Severing the Pipeline: Platform Bans and Misinformation in Closed Messaging Networks

Luis Meloni* Alexandros Cavgias† Felipe Bailez‡ Tiziano Rotesi§

May 2026

Abstract

We estimate how a regulator-imposed platform ban propagates into the content of an adjacent, closed messaging network, a setting whose downstream response to upstream shocks is largely unmeasured. We study Brazil's 39-day suspension of X/Twitter in 2024 using 145 million messages from 25,814 public WhatsApp groups and a daily difference-in-differences design on the 319 high-X-intensity groups for which treatment dose is well-defined. High-X-intensity groups lost 90.6% of their X-link traffic, 26% of their forwarded messages, and 31% of their potential misinformation during the ban, while only 14.7% of the lost X-link supply was recovered through alternatives such as Bluesky and Threads. The aggregate decline is driven almost entirely by the top 5% of pre-ban senders. The patterns are consistent with X functioning as a complement, not a substitute, for alternative upstream platforms: removing one platform quiets the downstream network rather than reshuffling content across alternatives.

JEL Codes: D83, L86, O33, P16

Keywords: platform regulation, deplatforming, closed messaging networks, WhatsApp, misinformation, Brazil

*Universidade de São Paulo (USP). Email: luis.meloni@usp.br.

†Ghent University.

‡Palver.

§University of Turin, Department of Economics.

1 Introduction

Closed messaging networks have become primary carriers of political information in much of the world. In Brazil, roughly 80% of the population uses WhatsApp, and 60% of Brazilian internet users report accessing information daily through messaging applications, a share that already exceeds traditional radio and television (Cetic.br, 2026; Kemp, 2024). Similar patterns hold across India, Indonesia, Nigeria, and much of the Global South, where WhatsApp and comparable platforms have become the main channel through which political news, campaign material, and misinformation circulate.[1] A growing literature documents how this content shapes electoral behavior, protest mobilization, and beliefs about public institutions.[2]

Yet what actually circulates inside these networks is largely invisible to outside observers: messages are encrypted by default, organized in private and group-level channels, and only imperfectly connected to the open social-media platforms whose data researchers and regulators can observe directly. This creates a gap between the political weight of closed messaging networks and our understanding of how they respond to policy interventions elsewhere in the information ecosystem. When policymakers intervene on the open platforms that feed content into these closed networks, as Brazil's Supreme Court did in August 2024 by suspending X/Twitter for 39 days, it is unclear what happens downstream. We investigate three questions. First, does a ban on an upstream platform reduce content flow in the adjacent closed network, or do users substitute toward alternatives? Second, if a reduction occurs, does it propagate to the composition of content—specifically to forwarded messages, potential misinformation, and aggression—beyond the mechanical loss of platform-specific links? Third, which senders drive any aggregate response, and does the pattern recover once the

[1] Examples include Resende et al. (2019) on Brazil, Garimella and Tyson (2018) on India, and Machado et al. (2019) on cross-platform flows.

[2] Caetano, Almeida, and Marques-Neto (2023), Recuero, Soares, and Volcan (2023), and Ventura et al. (2025), all on Brazil.

ban ends?

This paper estimates how a regulator-imposed platform ban propagates into the content composition of an adjacent, closed messaging network, a setting whose ecosystem-level response to upstream shocks is rarely observed systematically, even by regulators, platforms, or researchers with privileged access. We study Brazil's 39-day suspension of X/Twitter (August 30–October 8, 2024) as a supply shock to the upstream social-media input of Brazilian public WhatsApp groups. Brazil is a well-suited setting on three counts. WhatsApp is the dominant messaging platform in the country and a primary carrier of political content; cross-platform flows of news and misinformation from X/Twitter to WhatsApp are well documented in both the academic literature and the popular press; and the ban itself was abrupt, nationwide and enforced via DNS-level blocking, producing a clean removal of the upstream source rather than a gradual moderation adjustment. The ban also spanned the first round of the 2024 municipal elections (October 6), anchoring the shock in a politically salient window in which treated and control groups face a symmetric electoral environment.

We use a dataset of 145 million messages from 25,814 public WhatsApp groups covering the full year of 2024, combined with an external gold standard of 3,796 claims verified as false by four major Brazilian fact-check agencies. The paper makes three contributions: we causally estimate how an upstream platform ban propagates into the content composition of a closed downstream network; we quantify the extent of cross-platform substitution at the ecosystem level, showing that roughly 85% of the lost upstream supply is not recovered through alternatives; and we identify the sender population that drives the aggregate response.

Our empirical strategy is a daily difference-in-differences design at the group-day level that exploits cross-group variation in pre-ban reliance on X/Twitter. Starting from a corpus of 145 million messages from 25,814 public WhatsApp groups collected over 2024, we restrict attention to groups that actively shared X-links before the ban, so that the treatment dose is well-defined, and we compare groups above and below the median of pre-ban X-link in-

tensity.[3] This design contrasts X-heavy and X-light groups that are otherwise similar, and identifies how much of each group's activity and content depends on the upstream X pipeline. The specification absorbs group and calendar-day fixed effects, and identification rests on a standard parallel-trends assumption: absent the ban, X-heavy and X-light groups would have followed similar trajectories. We test this assumption with an event-study specification and find no systematic pre-trends in the main outcomes. We confirm the X-specific interpretation with a placebo exercise that substitutes eight alternative pre-ban dimensions into the treatment indicator; only X-intensity and, to a smaller extent, Telegram-intensity reproduce the ban-period X-link decline.

We document three main findings. First, the ban produced a large reduction in X-link traffic in treated groups (90.6%, or 2.80 links per group-day), reflecting both the mechanical loss of access to X-hosted content and a behavioral drop in cross-platform sharing that persists where the mechanical channel alone would not predict it. Substitution toward alternative platforms was limited: Bluesky and Threads together recovered only 14.7% of the lost X-link supply. Other candidates often discussed in U.S. and European deplatforming debates (Mastodon, Parler, Gab, Truth Social) showed no measurable response, as expected given their minimal footprint in the Brazilian information ecosystem.

Second, the disruption propagated to the content composition of WhatsApp messaging. Forwarded messages fell by 26%, aggression by 30%, and potential misinformation, measured using validated dictionary classifiers, by 31%. We corroborate the dictionary-based result with an externally anchored measure: messages matching claims that Brazilian fact-check agencies verified as false also fall during the ban, tracking the dictionary classifier and confirming that the content effects do not hinge on our own rhetorical patterns.

Third, the aggregate effect is concentrated in a small minority of heavy users. A sender-level decomposition shows that the top 5% of senders by pre-ban volume account for the

[3] We discuss the construction of the estimating sample in detail in Section 3, including thresholds and outlier treatment.

majority of X-link sharing, and their activity drives essentially all of the aggregate decline. Platform bans, at least in this episode, do not induce large migrations; they quiet a small set of high-intensity brokers whose activity fed the adjacent network. Beyond the ban period itself, we also find that the partial substitution toward Bluesky and Threads reverts once X access is restored and that mainstream news-link sharing remains depressed; Section 5.6 decomposes these post-ban patterns against the municipal election cycle that concluded for most municipalities in late October.

Taken together, the patterns are consistent with X functioning as a *complement* to WhatsApp group activity rather than a substitute that users can replace by multi-homing to alternative upstream platforms: removing X reduces the downstream content stream by far more than it reallocates it, and the decline concentrates in the same brokers who were previously importing X-sourced content into the network.

This paper contributes to three strands of the literature on media, platform regulation, and political behavior.

First, we contribute to a growing body of causal work on the effects of platform-level interventions on information consumption and political outcomes. This strand uses variation in platform access, penetration, or intensity to identify downstream effects on political behavior, from voter choice (Fujiwara, Müller, and Schwarz, 2024) to collective action (Enikolopov, Makarin, and Petrova, 2020; Qin, Strömberg, and Wu, 2024), news consumption and polarization (Allcott, Braghieri, et al., 2020; Levy, 2021; Allcott, Gentzkow, et al., 2024), and offline hate-directed behavior (Müller and Schwarz, 2021; Müller and Schwarz, 2023). We instead estimate the content response of a closed messaging network when an upstream open platform is suddenly removed. Relative to McCabe et al. (2024), who study reductions in misinformation reach within Twitter after the removal of 70,000 accounts, we shift the analysis from within-platform user-level removals to the downstream response of an adjacent closed network to the removal of an entire platform.

Second, we contribute to the literature on cross-platform information flows and the con-

sequences of platform access for the quality of online political communication. This strand shows that users multi-home across platforms and that information diffuses asymmetrically between them (Allcott, Braghieri, et al., 2020; Garimella and Tyson, 2018), and uses penetration or intensity measures to identify downstream political behavior (Enikolopov, Makarin, and Petrova, 2020; Bursztyn et al., 2024; Qin, Strömberg, and Wu, 2024). We quantify how much of a suddenly interrupted upstream content stream is recovered through alternatives: roughly 85% of the lost X-link supply is not reconstituted elsewhere, leaving the downstream network measurably quieter during the ban.

Third, we contribute to an emerging literature on platform regulation and political communication in large democracies outside the United States, and on public WhatsApp groups in Brazil in particular. This strand combines descriptive, experimental, and observational evidence on how messaging apps carry political content in the Global South.[4] We provide ecosystem-level causal evidence on how public WhatsApp groups respond to an upstream platform ban, complementing the individual-level deactivation experiment in Ventura et al. (2025): where they estimate user-level effects of limiting an individual's exposure to WhatsApp content, we estimate population-level effects of removing an entire upstream platform. Relative to earlier work on open-platform deplatforming of users or subreddits,[5] we study a regulator-imposed shutdown of an entire platform and measure its spillover into a closed downstream network.

The remainder of the paper is organized as follows. Section 2 describes the institutional setting. Section 3 presents the data, sample construction, the two complementary content measures (dictionary-based classifiers and fact-check agency-verified claims), and descriptive evidence on pre-ban activity patterns, with an exploratory LLM analysis documented in Appendix D. Section 4 details the empirical strategy. Section 5 presents the main results,

[4]Examples include Resende et al. (2019) and Machado et al. (2019) on misinformation flows, Caetano, Almeida, and Marques-Neto (2023) and Recuero, Soares, and Volcan (2023) on political behavior, Ozawa et al. (2024) on fact-checking, and Ventura et al. (2025) on deactivation effects.

[5]Examples include Chandrasekharan et al. (2017) on Reddit, Jhaver et al. (2021) on cross-platform migration, Horta Ribeiro, Jhaver, et al. (2021) on fringe communities, and Ali et al. (2023) on Twitter.

including activity, content, platform substitution, heterogeneity, sender-level decomposition, and the post-ban analysis. Section 6 reports threshold and specification robustness checks, placebo exercises that probe the X-specific interpretation of the ban-period coefficient, and classifier validation. Section 7 discusses policy implications; Section 8 concludes.

2 Institutional Background

2.1 The Brazilian X/Twitter Suspension

On August 30, 2024, Brazil's Supreme Court ordered the national telecommunications regulator to block access to X/Twitter throughout the country.[6] The block was implemented within hours through DNS- and IP-level restrictions, with VPN circumvention also prohibited. The ban remained in force for 39 days, until October 8, 2024, when service was restored.

Two features of the ban's timing matter for identification. First, the suspension began abruptly: while the STF had publicly threatened suspension if X failed to reinstate its legal representative in Brazil—issuing two consecutive 24-hour ultimatums on August 28 and 29—the specific shutdown date materialized within a 48-hour window rather than being set in advance, limiting the scope for behavioral anticipation; no grace period was provided. Consistent with this, the event study in Section 4 shows an upward drift in X-link sharing in the two weeks before the ban—the opposite of the pre-emptive reduction that behavioral anticipation would predict, and traceable to the public escalation of the STF–X conflict itself. Second, the first round of the 2024 municipal elections fell inside the ban window (October 6), so treated and control groups experienced the electoral shock symmetrically while X was suspended; this leaves identification of the ban-period effect unaffected by the election cycle. The post-ban window, by contrast, overlaps with the closing of the electoral cycle in most

[6]The order followed a prolonged conflict between the platform and the Brazilian judiciary over removal requests of accounts accused of spreading disinformation, and was preceded by X's decision to close its Brazilian office and withdraw its legal representative. See Recuero, Soares, and Volcan (2023) for background on the underlying judicial investigation.

municipalities, and we read the post-ban estimates with that caveat in mind (Section 4).

2.2 WhatsApp in Brazilian Political Communication

WhatsApp is the dominant messaging platform in Brazil, with over 170 million active users, roughly 80% of the population (Kemp, 2024). Unlike in the United States and Europe, where WhatsApp is primarily used for personal communication, in Brazil the platform serves as infrastructure for news consumption, political organizing, commercial transactions, and community coordination (Garimella and Tyson, 2018; Resende et al., 2019).

A relevant subset of WhatsApp groups are public, meaning their invite links are shared openly on websites, social media, or printed materials. Public groups function as broadcast channels: a small number of active senders produce and forward content to a large audience. The role of WhatsApp in Brazilian elections has been extensively documented, particularly during the 2018 and 2022 presidential cycles, when groups were identified as a primary channel for political misinformation (Machado et al., 2019; Resende et al., 2019; Ventura et al., 2025). Recent survey evidence shows that messaging apps have become one of the main channels through which Brazilians access political information, alongside or ahead of traditional broadcast media (Cetic.br, 2026). Because the platform is end-to-end encrypted, content moderation within groups is limited, though WhatsApp has restricted message forwarding to curb viral spread.

2.3 The X-to-WhatsApp Content Pipeline

The connection between X/Twitter and WhatsApp is directional: content produced on X is shared into WhatsApp groups via links, screenshots, and forwarded posts. X served as an upstream content-production platform where political commentators, news outlets, and activist accounts generated material that was then distributed downstream through WhatsApp's group infrastructure. This pipeline is asymmetric, since WhatsApp content rarely flows back to X, and concentrated among a small number of heavy users. In our data,

the top 5% of senders by pre-ban activity account for the majority of X-link sharing.

The ban severed this pipeline. When X became inaccessible, the upstream source of shareable content disappeared. The question we study is whether the downstream effects were confined to the mechanical loss of X links, or whether the disruption propagated to broader messaging behavior, content characteristics, and the activity patterns of different types of senders.

3 Data and Sample

3.1 Corpus

Our data consist of 145,262,238 messages from 25,814 public Brazilian WhatsApp groups, spanning the full calendar year 2024 (January 1 through December 31). For each message we observe a group identifier, a sender identifier (hashed at collection), message text, a timestamp, a content type label (text, image, video, audio, sticker, document), a forwarding indicator, and derived text fields from OCR on images and transcription of audio messages.[7]

We classify groups into nine thematic categories (politics, sales, sports, religion, education, condominium management, health, adult content, and other) using a keyword-based algorithm applied to message text within each group. A group receives a primary type label when the share of messages containing keywords for that category exceeds 25%. Groups that do not clear the 25% threshold are labeled "other." Within political groups, we further classify political orientation as left, right, mixed, or neutral using a scoring rule based on partisan keywords and domain links.[8] A group is classified as left or right when at least 60% of its political-signal messages lean in one direction; groups below this threshold are

[7]The corpus is stored as 122 compressed partitions with record-level metadata. Sender identifiers are SHA-256 hashed at collection. The forwarding indicator takes values 0 for original messages, 1–4 for the number of forwarding hops, and 127 as a capped indicator for highly forwarded content.

[8]Examples of partisan domains include *brasil247.com* and *cartacapital.com.br* on the left, *oantagonista.com* and *jovempan.com.br* on the right; the full list of partisan keywords and domains appears in Appendix C.

classified as mixed.

3.2 Analysis Sample

The analysis sample comprises groups with meaningful pre-ban exposure to X/Twitter content. We apply four sequential filters. First, we compute each group's average number of X links shared per week, denoted *x_per_week*, over the 13-week pre-treatment window from June 1 to August 29, 2024, and retain groups with *x_per_week* at least 1. This filter selects groups where X content was a regular feature of the information environment, not a one-off occurrence. Second, we impose a balanced-tail requirement: a group must have been active (at least one message) in at least 2 of the 4 weeks immediately preceding the ban (weeks starting July 29, August 5, August 12, and August 19). This ensures that groups in the sample were genuinely operating shortly before treatment onset, ruling out long-dormant groups that happen to have accumulated X links earlier in the window. Third, we impose *no* post-ban activity requirement. Groups that fell completely silent after the ban are retained, because dropping them would condition on the outcome: the groups most affected by the ban are those most likely to cease activity. Fourth, we drop six hub outliers with *x_per_week* above 100, which behave as amplification nodes with idiosyncratic dynamics: their pre-ban trajectories diverge sharply from the rest of the sample and account for roughly 85% of the residual differential pre-trend. Capping at 100 retains 98% of qualifying groups while substantially improving the parallel-trends diagnostic. Section 6 reports estimates without this cap.

These filters yield a canonical sample of 319 groups: 159 treated and 160 control (Appendix G.1 reports a pre-ban balance table). The treatment variable treated_g equals 1 if a group's *x_per_week* exceeds the sample median of 5.03 links per week, and 0 otherwise. Groups above the median shared, on average, substantially more X-sourced content and are therefore more exposed to the supply shock created by the ban. We construct a balanced *daily* panel covering the 228 calendar days from January 1 to December 29, 2024 that fall

inside the active observation window, yielding 72,672 group-day observations. Days in which a group sent zero messages are filled with zeros for all outcome variables rather than treated as missing, ensuring that group silence is measured as an outcome.

Table 1 compares the analytical sample to the full corpus. The sample is highly selected on activity and political engagement. Sample groups send roughly twelve times more messages per year than the average corpus group (38,283 vs. 3,124), are active four times more days (228 vs. 56), and sustain larger participation (median 39 unique senders vs. 2). They share X-link content at rates two orders of magnitude higher than the typical corpus group (median 5.0 X links per week pre-ban vs. 0.0). Composition is also strongly skewed: 76% of sample groups are politically themed compared to 7% in the full corpus, and within political groups, right-leaning groups dominate the sample (41% of all sample groups vs. 2.4% of corpus). The sample therefore describes the intensive margin of politically engaged, X-dependent WhatsApp groups, not the average Brazilian WhatsApp group.

Table 1: Sample characterization: full corpus vs. analytical sample

	Full corpus		Analytical sample	
	Mean	Median	Mean	Median
Groups (N)	25,814		319	
Panel A: Group activity				
Messages (2024 total)	3,123.57	135.00	38,282.89	18,865.00
Messages per active day	29.31	5.67	164.44	85.55
Days active	56.26	22.00	227.81	243.00
Distinct senders (daily max)	9.73	2.00	61.46	39.00
Forwarded messages per week	8.86	0.02	140.45	57.68
Share of messages forwarded (%)	7.69	0.41	22.43	21.85
Panel B: Content shared				
Total links per week	17.86	0.34	215.62	111.52
X/Twitter links per week	0.16	0.00	7.22	3.63
News links per week	0.12	0.00	6.50	1.64
Panel C: Dominant thematic category (% of groups)				
Politics	6.71%		76.18%	
Sales	46.86%		13.17%	
Religion	10.68%		5.33%	
Sports	4.81%		4.08%	
Education	3.06%		0.31%	
Health	1.94%		0.63%	
Condominium	1.90%		0.00%	
Other	24.05%		0.31%	
Panel D: Political orientation (% of groups)				
Left-wing	1.89%		13.17%	
Right-wing	2.37%		41.07%	
Mixed	2.44%		21.94%	
Neutral / non-political	93.29%		23.82%	

Notes: This table presents a comparison of group-level summary statistics between the full 25,814-group public Brazilian WhatsApp corpus collected from January 1 to December 29, 2024, and the 319-group analytical sample used in the daily difference-in-differences estimation. The analytical sample is defined by four filters: (i) at least one X/Twitter link shared per week over the pre-treatment window (June 1 to August 29, 2024); (ii) the group was active in at least two of the four weeks immediately before the ban; (iii) no post-ban activity requirement; and (iv) a cap at 100 X-links per week that excludes six high-intensity hub groups. Panels A and B report cross-group means and medians. Panels C and D report the share of groups in each category. "Messages per active day" is the 2024 total divided by the number of days with at least one message. "Days active" is the number of days with at least one message. "Distinct senders (daily max)" is the largest number of distinct sender_id values observed in a single day in 2024. "Forwarded messages" counts messages with a non-zero WhatsApp forwarding score. "X-links per week" is the weekly average over the 13-week pre-ban window (June 1 to August 29, 2024) used to define treatment. Group type is assigned when the share of messages containing keywords for a category exceeds 25%, and "Other" covers groups below the threshold for any single theme. Political orientation is computed within politically classified groups using a 60% partisan-keyword and partisan-domain threshold, with non-political groups coded as neutral.

3.3 Measuring Potential Misinformation

A central challenge for our analysis is measuring the content of WhatsApp messages. No single classifier fully captures what "misinformation" is, so we use two complementary approaches that differ in precision, recall, and what they actually measure. Our primary approach is a dictionary-based classifier applied to all 145 million messages. We complement it with an externally-validated gold standard built from claims that Brazilian fact-check agencies have verified as false. We also explored a large-language-model (LLM) classifier on a subsample but excluded it from our content analysis after validation revealed severe permissiveness against the dictionary; we document this in Appendix D.

Dictionary-based classifiers. Our first and broadest approach uses regular-expression classifiers applied to all messages. We focus on four categories. *Potential misinformation* captures rhetorical markers characteristic of misinformation, including urgency-and-share directives, censorship narratives, conspiracy labels, and wake-up rhetoric; the classifier captures the *form* of misinformation rather than verifying the truth of specific claims. *Aggression* captures hostile, insulting, or threatening language such as political epithets used as insults, personal insults, and violent threats; it is broader than hate speech and does not require that the target be a protected group. *Hate speech (broad)* captures hostile group-targeted language including direct misogynist slurs, homophobic epithets, and anti-LGBT rhetoric. *Hate speech (strict)* is a narrower version that captures only severe slurs directed at protected groups defined by race, gender identity, sexual orientation, religion, or nationality. The classifiers incorporate negative-lookahead patterns to exclude defensive statements, exclusion filters for idiomatic vulgarity, and a news-article filter that raises the misinformation threshold when the message contains markers of legitimate journalism.[9] To assess the precision of the dictionary classifiers, we construct a hand-labeled audit: a set of 199 messages drawn

[9]Examples of excluded patterns: defensive constructions mentioning slurs in condemning contexts; idiomatic vulgarity such as "puta que pariu" used as generic frustration; and messages with URLs to major news domains plus markers like "leia mais" or "fonte:" which signal that the content is a news article rather than a bare claim.

from the pre-ban window (June 11–13, 2024), each read and hand-coded by a human coder for the four content categories.[10] Precision ranges from 0.86 (aggression) to 1.00 (hate strict), recall from 0.71 (potential misinformation) to 1.00, and F1 from 0.80 to 1.00. Appendix A reports full audit details.

LLM-based classification (exploratory only). We also explored GPT-4o-mini (zero-shot, deterministic; no fine-tuning or task-specific examples) as a classifier for the same content categories, but validation revealed severe permissiveness relative to the dictionary: positive rates are 4–50× higher across categories, with Cohen's κ near zero for hate speech and potential misinformation (Appendix D). We therefore exclude LLM-classified outcomes from our content analysis and rely on the dictionary classifier and the externally anchored *Fake news* measure.

Fact-check agency gold standard. Our third measure anchors content measurement to an external benchmark: four IFCN-certified Brazilian fact-checking agencies (*Aos Fatos*, *Agência Lupa*, *Boatos.org*, and *e-Farsas*), yielding 3,796 claims verified as false in 2024. Using an inverted index over the 319-group message panel, we flag each message that matches a claim on at least three proper-noun tokens and a minimum composite keyword score, provided the message date precedes the debunking publication date. This externally-anchored outcome corroborates the dictionary classifier and enters the *Fake news* row of Table 2 and Table 4; full matching details and precision audit are in Appendix B.

Domain-level measures. Finally, we construct domain-level outcomes by extracting URLs from all messages and classifying host domains into 30 pre-registered categories, including left-wing Brazilian media, right-wing Brazilian media, mainstream news, fact-checking organizations, and each major social-media service. The full list of hosts per category appears in Appendix C.

[10] The sample is stratified to over-represent machine-flagged positives (because the population positive rate is below 1% for every category, a random draw would produce too few positives to estimate precision). 199 is the final size after dropping empty and duplicate messages from an initial draw of 200. Hand-coding takes several minutes per message; we ran the exercise in-house with one coder on the authors' team and the sample size reflects what could be coded with consistent quality.

3.4 Data Access and Ethics

The data were collected by the Brazilian technology company *Palver* using the public-group monitoring methodology described in Bailez (2022). Palver joins public WhatsApp groups (groups whose invite links are shared openly on websites, social media posts, or printed materials) and records the messages exchanged.[11] Because the invite links are public, membership is open to anyone who follows the link and there is no expectation of privacy equivalent to that of private, invitation-only groups. Palver does not circumvent or decrypt WhatsApp's end-to-end encryption; the messages we observe are identical to those visible to any group member. The collection complies with the General Data Protection Law (Lei Geral de Proteção de Dados, LGPD, Law 13.709/2018): all personal identifiers are masked or hashed at the point of collection, and the raw dataset that we access contains no names, phone numbers, or location fields.

The analysis operates at the group-day level, and for sender heterogeneity at the sender-group-day level using SHA-256 hashed sender identifiers, making re-identification of individuals infeasible.

3.5 Descriptive Evidence

Before turning to the formal identification strategy, we present descriptive evidence on three patterns visible in the raw data: the collapse in X-link sharing during the ban, the broader decline in messaging activity, and the partial substitution toward alternative platforms.

X-link sharing. Figure 1, Panel A plots the daily mean of X/Twitter links per group across the 159 treated groups, with a 7-day moving average overlaid. The ban (shaded region) produces a sharp, near-immediate drop in X-link sharing that persists for the full 39-day suspension. The decline is not gradual: link counts fall to near zero within the first two

[11] The same collection infrastructure has been used in prior academic work on Brazilian political communication (Peron, Pereda, and Bailez, 2025).

days, consistent with the abruptness of the DNS and IP-level blocks described in Section 2. When the ban is lifted on October 8, X-link sharing rebounds, though not to pre-ban levels. The post-ban average remains below the June to August baseline, suggesting that the ban disrupted sharing habits beyond the mechanical period of inaccessibility.

Platform substitution. Figure 1, Panel B plots daily link counts for Bluesky, Threads, and the Bluesky+Threads aggregate across the 159 treated groups. During the ban, X links collapse while Bluesky and Threads links spike sharply. Bluesky activity rises from a near-zero baseline (the platform was barely used in the canonical groups before the ban) to roughly 0.4 links per group-day at the peak; Threads spikes to about 0.5. The combined surge is concentrated in the first weeks of the ban and begins to fade before the ban ends. After X is restored, both Bluesky and Threads revert toward their pre-ban baseline. Section 5.2 quantifies these substitution patterns formally and Section 5.6 discusses their post-ban dynamics.

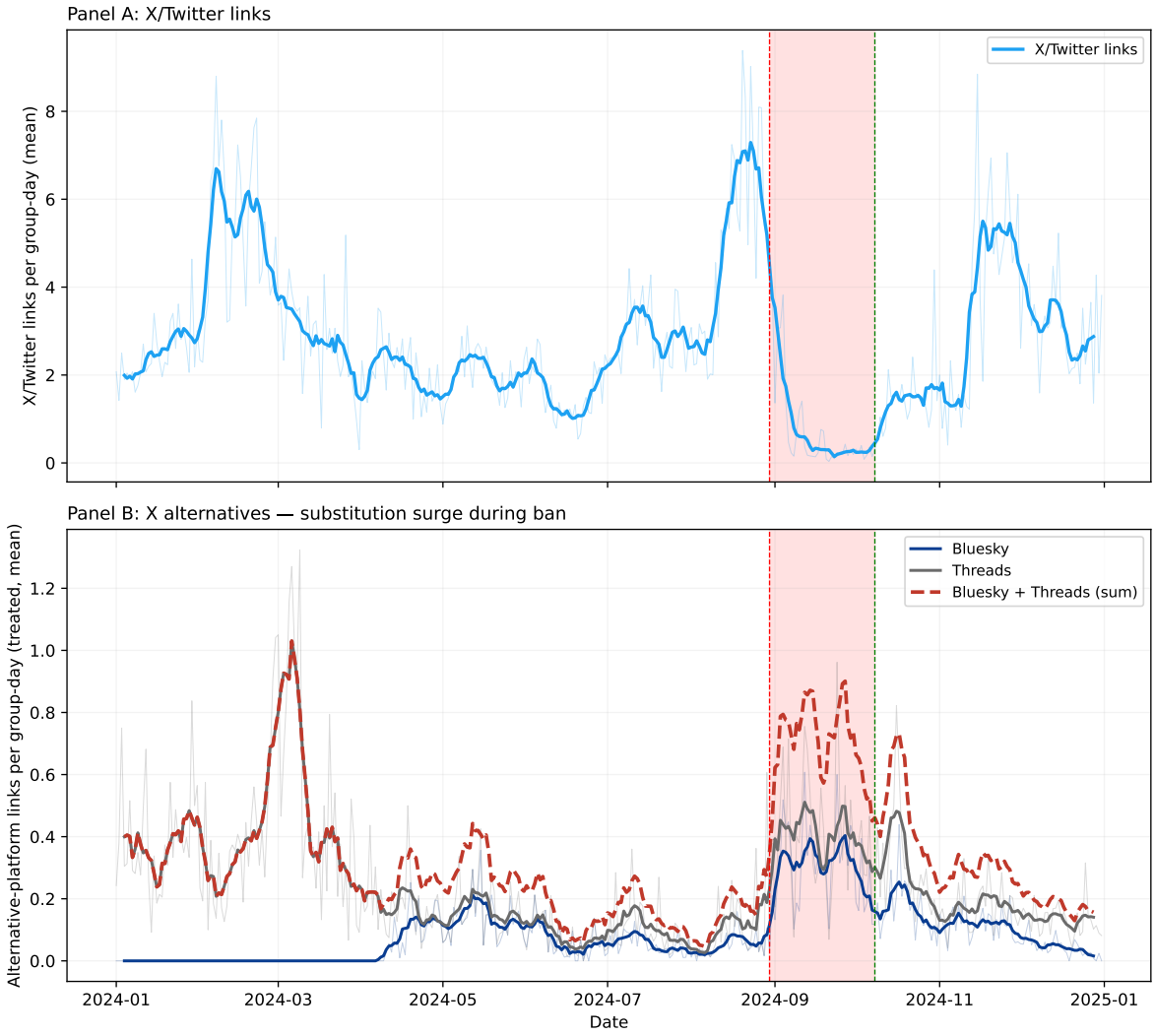


Figure 1: Daily X/Twitter and alternative-platform link sharing in WhatsApp groups

Notes: Daily link counts per group, mean across the 159 treated groups (above-median pre-ban X-intensity). Faint lines show raw daily means; bold lines show 7-day centered moving averages. Panel A plots X/Twitter links (blue line). Panel B plots Bluesky (dark blue), Threads (grey), and the Bluesky+Threads aggregate (red). The dashed red line marks the start of the X suspension (August 30, 2024); the dashed green line marks its end (October 8, 2024). The shaded area indicates the ban period.[a] Panel A shows that X-link sharing collapses during the ban and rebounds incompletely after restoration. Panel B shows that Bluesky (near-zero baseline) and Threads spike during the ban window and revert to pre-ban levels once X access returns.

---

[a]Treatment intensity *x_per_week* is constructed as follows: for each group, count the number of URLs whose host domain is `twitter.com`, `x.com`, `t.co`, or `mobile.twitter.com` sent between June 1 and August 29, 2024, then divide by 13 weeks. The analytical sample retains groups with *x_per_week* ≥ 1 and ≤ 100. The full list of X/Twitter host domains is reported in Appendix C.

4 Empirical Strategy

4.1 Design

We estimate the effect of the X ban on WhatsApp group outcomes using a difference-in-differences design at the group-day level that exploits variation in pre-ban X-link intensity across groups. The treatment variable treated_g equals 1 for groups whose pre-ban X-link sharing rate (*x_per_week*, computed over June 1 to August 29) exceeds the sample median of 5.03 links per week, and 0 otherwise. Groups above the median depend more heavily on X-sourced content and are therefore more exposed to the supply disruption. Because the ban was national, both treated and control groups experienced some reduction in X-link traffic; the DiD estimates the *differential* effect of higher pre-ban X-intensity, and the calendar-day fixed effects absorb the common supply shock shared by all groups on each day.

Our estimating equation is:

$$Y_{gd} = \alpha_g + \delta_d + \beta\,(\text{treated}_g \times \text{Ban}_d) + \theta\,(\text{treated}_g \times \text{Post}_d) + \varepsilon_{gd} \tag{1}$$

where Y_{gd} is the outcome for group g on day d, α_g are group fixed effects that absorb time-invariant differences in group activity levels, δ_d are calendar-day fixed effects (one dummy per day) that absorb common time shocks shared across groups, Ban_d is an indicator for the ban period (August 30 through October 8, 2024), and Post_d is an indicator for days from October 9 onward. The interactions $\text{treated}_g \times \text{Ban}_d$ and $\text{treated}_g \times \text{Post}_d$ are the difference-in-differences terms. Standard errors are clustered at the group level ($G = 319$), since treatment is assigned at the group level and shocks to group-level activity are likely correlated within groups over time.

The coefficient β is the parameter of primary interest: the average treatment effect on the treated during the ban period, expressed in outcome units per group-day. The coefficient θ captures what happens after the ban ends and the mechanical access channel is

restored. Section 5.6 analyses θ separately: the post-ban window overlaps with the October 27 municipal-election runoff and the December holiday season, and we disentangle those from genuine post-ban behavioral persistence by estimating a re-specified model on the ban + post-ban sample with the ban as the reference period.

The ban is a single-date shock: all 319 analytical groups (and all 25,814 corpus groups) are exposed to the same treatment start on August 30, 2024, so Equation 1 has a single treatment cohort rather than staggered adoption. The recent literature on heterogeneity-robust estimators for staggered designs (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; Chaisemartin and D'Haultfoeuille, 2020; Borusyak, Jaravel, and Spiess, 2024) therefore does not bind here: the two-way fixed-effects estimator is not contaminated by bad-comparison-group weights when there is only one cohort and one adoption date. Pre-ban variation in the date of group creation does not constitute staggered treatment timing because treatment is defined by the ban, not by group entry.[12]

4.2 Parallel Trends and Election Timing

The identifying assumption is that, absent the X ban, high- and low-intensity groups would have followed parallel trends in messaging outcomes. We assess this assumption using an event-study specification that replaces the ban and post-ban indicators in Equation 1 with a full set of daily leads and lags:

$$Y_{gd} = \alpha_g + \delta_d + \sum_{k \neq -1} \mu_k \left(\text{treated}_g \times \mathbb{1}[d - d_{\text{ban}} = k] \right) + \varepsilon_{gd} \tag{2}$$

[12] As a direct test of the stable unit treatment value assumption, Appendix G.7, Table 25 reports mean outcomes for the 160 control groups during the pre-ban and ban periods. X-link sharing fell 58% in control groups during the ban, consistent with a platform-wide supply shock: the ban suspended X access for all Brazilian users regardless of group treatment status. Crucially, potential misinformation is essentially unchanged in control groups (-1.4% relative to the pre-ban mean), and the decline in aggression (-17%) tracks the overall message-volume contraction (-23%) rather than any cross-group content diffusion. These patterns indicate that content spillovers from treated to control groups are negligible; the DiD identifies the intensive-margin effect of X-dependency.

where k indexes event-time days relative to the ban start (d_{ban} = August 30, 2024), and $k = -1$ is the omitted reference day. We estimate μ_k for $k \in [-90, +90]$. Under the parallel trends assumption, lead coefficients in the pre-ban window should fluctuate around zero with no systematic drift. Figure 2 displays the X-link event-study coefficients, with a flat pre-period followed by a sharp, sustained drop at $k = 0$; Appendix Figure 11 and Figure 12 present the remaining outcomes. The seven-day moving average across pre-ban leads stays within ± 1 X-link per group-day for the main X-link outcome, less than 40% of the absolute magnitude of the ban-period effect.

Election timing. Brazil's 2024 elections were *regional*: voters elected municipal mayors and city councilors, not federal offices. The first round took place on October 6, 2024 (inside the ban window), and the runoff on October 27, nineteen days after the ban ended. Brazilian electoral law reserves runoffs for municipalities with more than 200,000 registered voters; in 2024 this produced 51 runoff cities out of roughly 5,570 total municipalities, fewer than 1% of municipalities and covering about 35% of the population (Tribunal Superior Eleitoral, 2024). The practical implication is that for the bulk of municipalities the electoral campaign concluded on October 6 and political messaging naturally deflated thereafter, regardless of any platform shock.

Within this timing, the ban period itself is the cleanest estimand. The first round sits inside the ban window and affects treated and control groups symmetrically; identification relies on X-intensity, not political identity. Non-political groups (sales, religion, sports) show a first-stage X-link decline comparable to political groups (Section 5.4, Table 4), which would not be expected if electoral mobilisation were the driver. Threshold sensitivity produces a monotonic dose-response in X-intensity (Table 6), also inconsistent with an election-driven alternative.

The more serious timing concern is not with the ban but with the *post-ban* window (Post_d in Equation 1), which spans the runoff on October 27 and the December holiday season.

We treat θ as describing the full post-ban experience rather than a clean post-restoration behavioral effect, and Section 5.6 decomposes it into an immediate post-first-round window and a post-runoff window that doubles as the holiday season.[13]

We adjust for multiple comparisons within outcome families using Benjamini-Hochberg q-values and report them alongside cluster-robust p-values.

5 Results

5.1 Main Activity Effects

The ban reduced X/Twitter link sharing by 2.80 links per group-day (SE = 0.35, $p < 0.01$), a 90.6% decline from the treated-group pre-ban mean of 3.09 links per day (Table 2, Panel A). High-X-intensity groups lost essentially all of their X-sourced link traffic during the suspension, for the mechanical reason that the platform was inaccessible. This first stage establishes that the supply shock was real and large. Whether it propagated beyond X-specific links to the substitution margin and to the broader content composition of WhatsApp groups is the paper's central question, addressed in Sections 5.2 and 5.3.

The effects extend beyond X-specific content. Forwarded messages fell by 13.08 per group-day during the ban (SE = 3.44, $p < 0.01$), a 26.0% decline from the treated-group pre-ban mean of 50.26. Total messages declined by 41.49 per group-day (SE = 23.06, $p < 0.10$), a 16.5% drop from the pre-ban mean of 251.97. Unique senders fell by 5.12 (SE = 1.75, $p < 0.01$), a 13.8% decline from 37.03. Granular media outcomes (videos, images, messages with links, mainstream news links) show comparable declines and are reported in Appendix Table 21. The pattern is consistent across link-sharing, visual content, and participation: the ban reduced the full range of media that circulates through WhatsApp groups, not just

[13]Standard errors clustered at the group level with $G = 319$ clusters sit well above the 50-cluster threshold that generates cluster-size concerns in the wild-bootstrap literature; we report wild-cluster-bootstrap diagnostics for the sender-level bucket analysis in Appendix F. See MacKinnon and Webb (2017) and Cameron, Gelbach, and Miller (2008).

the narrow category of X/Twitter URLs. Audio messages are flat (Appendix Table 21); voice messages are produced within WhatsApp and do not depend on upstream platform content. Their stability confirms that the ban affected content that originates on or flows through X while leaving indigenous WhatsApp activity undisturbed.

What happens after X access is restored on October 8 is a separate question, analysed in Section 5.6 with the election timing made explicit.

Figure 2 presents the event-study coefficients for the X-link outcome. Pre-ban coefficients fluctuate around zero, with the seven-day moving average remaining within ± 1 X-link per day for roughly eight weeks before treatment. At event-day $d = 0$ the coefficient drops sharply, and the effect persists through the full 39-day ban window. A localized upward movement in the final five pre-ban days reflects the public escalation of the Brazilian Supreme Court's conflict with X (see Section 2); excluding that final week from the lead window does not affect the main coefficients. Event-study estimates for the five other main activity outcomes and for the content outcomes appear in Appendix Figure 11 and Figure 12.

Pre-trend diagnostics. Joint F-tests on the leads $k \in \{-8, -4, -2, -1\}$ pass at conventional thresholds for total messages, forwarded messages, aggression, and potential misinformation in the pooled sample. The X-link outcome shows a transient upward drift in the final two pre-ban weeks, traceable to the public escalation of the STF–X conflict in late August 2024 (Section 2). The direction of this drift is opposite to the ban-period effect, so any pre-trend contamination biases the ban coefficient toward zero rather than inflating it. Heterogeneity pre-trends by group type and political orientation are reported in Appendix Table 23; sender-bucket pre-trends are in Appendix Table 24. The heterogeneity diagnostic reveals three buckets that reject the null of flat leads on the X-link outcome (politics, sports, and right-wing political groups), all traceable to the same pre-ban political escalation; the rejections are outcome- and bucket-specific and opposite in sign to the ban-period effect.

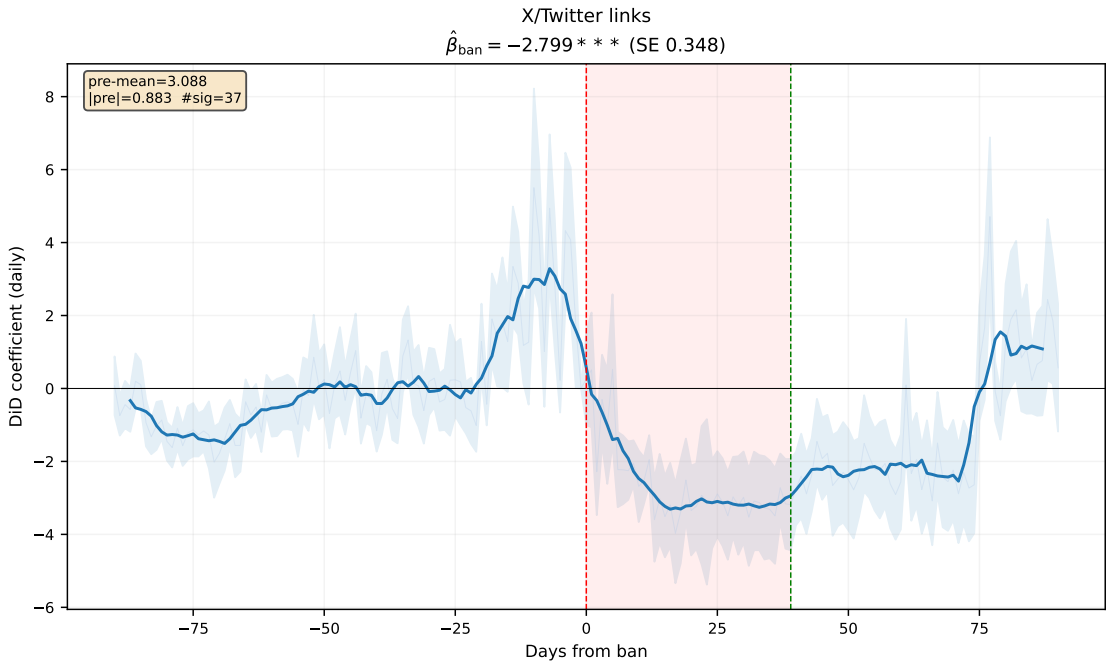


Figure 2: Event study: X/Twitter links per group-day

Notes: This figure presents daily event-study coefficients μ_k from Equation 2 for event-time $k \in [-90, +90]$ with reference day $k = -1$, estimated on the 319-group analytical sample. The faint line plots raw daily coefficients, the bold line plots the 7-day centred moving average, and the shaded band is the 95% confidence interval from group-clustered standard errors ($G = 319$). The dashed red vertical line marks ban onset at $k = 0$ (August 30, 2024) and the dashed green vertical line marks ban end at $k = 39$ (October 8, 2024).

Table 2: Effect of X ban on WhatsApp group activity and content

Outcome	Pre-mean (treated)	Ban effect Coef.	SE	$\%\Delta_{\text{ban}}$ (% of pre-mean)
Panel A: Activity				
X/Twitter links	3.09	-2.80^{***}	(0.35)	-90.6%
Forwarded messages	50.26	-13.08^{***}	(3.44)	-26.0%
Total messages	251.97	-41.49^{*}	(23.06)	-16.5%
Unique senders	37.03	-5.12^{***}	(1.75)	-13.8%
Panel B: Content				
Fake news	0.4636	-0.1439^{***}	(0.0475)	-31.0%
Fake news (broad)	0.6631	-0.1511^{**}	(0.0665)	-22.8%
Aggression	6.4156	-1.9055^{**}	(0.7925)	-29.7%
Hate speech (broad)	0.1468	-0.0228	(0.0342)	-15.5%
Fact-check links	0.0622	-0.0330^{**}	(0.0141)	-53.0%

Notes: This table presents daily DiD estimates from Equation 1 using the 319-group analytical sample (159 treated, 160 control; 72,672 group-day observations from January 1 to December 29, 2024). Treated is defined as above-median pre-ban *x_per_week*, with a median of 5.03 X-links per week. Each row reports a separate regression with the listed outcome as the dependent variable, per group-day. Panel A covers activity outcomes (raw counts per group-day). Panel B covers content outcomes constructed from three measurement strategies: regular-expression dictionary classifiers for *Potential misinformation*, *Aggression*, and *Hate speech (broad)* (Section 3.3, Appendix A; F1 scores 0.80, 0.82, and 0.85); domain-host counts for *Fact-check links* (URLs to Brazilian fact-checking organizations such as *Aos Fatos* and *Agência Lupa*); and external claim-matching against fact-check agency gold standard for *Fake news* (messages that share proper-noun tokens with a claim subsequently debunked by an IFCN-certified Brazilian agency, before the debunking publication date; Appendix B). "X/Twitter links" counts URLs to x.com, twitter.com, t.co, or mobile.twitter.com; "Forwarded messages" counts messages with a non-zero forwarding score; "Total messages" counts all messages; "Unique senders" counts distinct sender identifiers active in the group that day. The "Ban (vs pre)" column reports the treated-control differential during the ban period (August 30 to October 8, 2024) relative to the omitted pre-ban baseline (January 1 to August 29, 2024). The "$\%\Delta_{\text{ban}}$ vs pre-mean" column normalises the ban coefficient by the pre-ban mean of the outcome in the treated subsample. Group and calendar-day fixed effects; standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2 Platform Substitution

The substitution ratio (aggregate X-alternative substitution divided by the X decline) is $0.412/2.799 = 0.147$, or 14.7%. Roughly 85% of the X content supply that entered WhatsApp groups was not replaced by any alternative source. These patterns support a specific interpretation of how X content functioned within WhatsApp groups. X was an upstream content source, not a substitute for other platforms. Political commentators and activists produced content on X that was then forwarded into WhatsApp groups. When the upstream

source disappeared, the downstream forwarding pipeline collapsed, because producing equivalent original content on WhatsApp directly is costly relative to forwarding an X link. Users briefly explored Bluesky and Threads as alternative upstream sources, but the content supply on those platforms was thin and interest faded within weeks. Section 5.6 examines the recovery of each platform once X access is restored.

A central question for platform regulation is whether banning one platform simply redirects content to alternatives. We test this by estimating Equation 1 separately for each pre-registered platform category, with daily link-sharing on each platform as the outcome. Table 3 reports results, restricted to platforms with measurable baseline presence in our sample (pre-ban mean ≥ 0.01 links per group-day). The "alt-tech" platforms commonly invoked in North American and European deplatforming debates (Mastodon, Parler, Gab, Truth Social, Minds, Rumble, Odysee, BitChute) all fall below this threshold and are omitted; the ban produced no measurable increase in any of them.

The first stage confirms the main result: X/Twitter links fell by 2.799 per group-day (SE = 0.348, $p < 0.01$) from a pre-ban mean of 3.088, a 90.6% decline. Against this decline, only two platforms show positive substitution. Bluesky links increased by 0.231 per group-day (SE = 0.047, $p < 0.01$), a fourfold rise from a pre-ban mean of 0.054. Threads links increased by 0.185 per group-day (SE = 0.069, $p < 0.01$), an 84.1% rise from a pre-ban mean of 0.220. Both effects are large in proportional terms but small in absolute terms: the combined Bluesky + Threads aggregate added 0.412 links per group-day (SE = 0.105, $p < 0.01$).

Every other major platform either showed no change or declined. YouTube links fell by 3.149 per group-day (SE = 0.772, $p < 0.01$) from a pre-ban mean of 12.93, a 24.4% drop. Telegram fell by 1.133 (SE = 0.305, $p < 0.01$), an 87.4% decline from a thin baseline. The Telegram result is noteworthy: the platform most often flagged in popular discussion as the natural refuge for displaced political content showed the opposite pattern. TikTok and Kwai showed no significant ban-period effect.

The simultaneous decline of most other platforms is consistent with the aggregate activity drop documented in Table 2: overall messaging volume fell by about 16% in treated groups during the ban, so link-sharing across most categories should mechanically fall with it. The substitution that we can detect has to be read against this shrinking base, and the fact that only Bluesky and Threads grow (while the rest of the link portfolio contracts along with overall activity) reinforces the interpretation that users did not broadly reallocate their sharing to other upstream platforms.

Table 3: Platform Substitution: DiD Estimates by Platform

	Pre-mean	Ban effect		$\%\Delta_{\text{ban}}$
Platform	(treated)	Coef.	SE	(% of pre-mean)
Panel A: X/Twitter and alternatives				
X / Twitter	3.088	-2.799^{***}	(0.348)	-90.6%
Bluesky	0.054	$+0.231^{***}$	(0.047)	+424.2%
Threads	0.220	$+0.185^{***}$	(0.069)	+84.1%
All X-alternatives (aggregate)	0.279	$+0.412^{***}$	(0.105)	+147.7%
Panel B: Other major platforms				
YouTube	12.926	-3.149^{***}	(0.772)	-24.4%
Instagram	6.656	-0.747^{*}	(0.444)	-11.2%
Facebook	3.206	-0.380	(0.260)	-11.9%
TikTok	2.575	-0.206	(0.425)	-8.0%
Telegram	1.297	-1.133^{***}	(0.305)	-87.4%
Kwai	1.870	-0.281	(0.192)	-15.0%
Panel C: Ideologically aligned and aggregate				
Right-wing BR domains	5.617	-1.203^{**}	(0.531)	-21.4%
Left-wing BR domains	0.935	-0.249^{***}	(0.074)	-26.6%
Mainstream news	1.252	-0.322^{***}	(0.124)	-25.7%
Fact-check	0.070	-0.033^{**}	(0.016)	-47.3%
WhatsApp	7.022	-1.706	(1.420)	-24.3%
Total links (any)	66.527	-11.800^{**}	(4.622)	-17.7%

Notes: This table presents DiD estimates from the daily specification in Equation 1 using the 319-group analytical sample (72,672 group-day observations). Treated is defined as above-median pre-ban *x_per_week* (the median is 5.03 X-links per week). Each row reports a separate regression with a different platform-specific daily link count as the outcome. "Pre-mean (treated)" is the average daily count in treated groups before the ban; "$\%\Delta_{\text{ban}}$" is the ban coefficient as a share of the pre-mean. The "All X-alternatives (aggregate)" row sums Bluesky, Threads, Mastodon, Parler, Gab, Truth Social, and Minds; the latter five had near-zero baseline presence. Platforms with pre-ban mean below 0.01 links per group-day (Mastodon, Parler, Gab, Truth Social, Minds, Rumble, Odysee, BitChute) are omitted from the table and all produced null ban effects. The recovery ratio $0.412/2.799 = 14.7\%$ quantifies the aggregate X-alternative substitution as a share of the X-link decline. Platform categories are pre-registered (Appendix C). Group and calendar-day fixed effects; standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.3 Content and Behavioral Effects

Potential-misinformation patterns fell by 0.144 messages per group-day (SE = 0.048, $p < 0.01$), a 31.0% decrease from the treated-group pre-ban mean of 0.464 (Table 2, Panel B). This classifier captures the rhetorical packaging of misinformation (urgency-and-share directives, censorship narratives, conspiracy labels, and wake-up rhetoric) and achieves F1 = 0.80 in the pre-ban audit and F1 = 0.89 ($\kappa = 0.83$, two coders) in a ban-period revalidation on 200 messages from August 30 to October 8, 2024 (Appendix A.3, Table 10).[14] Aggression declined by 1.91 messages per group-day during the ban (SE = 0.79, $p < 0.05$), a 29.7% decrease from the pre-ban mean of 6.42.[15]

Hate speech (broad) is a precise null (−0.023, SE = 0.034, $p = 0.50$). This null rules out the interpretation that the ban suppressed identity-targeted hostility against protected groups: once the hate-speech dictionary is tightened to exclude political insults misclassified as identity-targeted hostility, the ban's apparent hate-speech reduction disappears. What the ban affected was the aggression category (political epithets, violent insults), not identity-based hostility. The strict hate-speech classifier (severe slurs only, $N = 5$ gold positives with F1 = 1.00) is also null. Appendix A documents the classifier validation in full.

Domain-level measures reveal effects across the political spectrum. Right-wing Brazilian media links fell by 1.203 per group-day (SE = 0.531, $p < 0.05$) during the ban, a 21.4% de-

[14]We extend the Rambachan and Roth (2023) relative-magnitude sensitivity analysis to content outcomes. For potential misinformation the breakdown $\bar{M} = 0.5$: the negative effect is significant at $\bar{M} = 0$ (original PT CI $[-0.33, -0.06]$) but loses significance when post-ban violations reach half the maximum pre-period deviation. For aggression the weekly event-study average does not robustly exclude zero even at $\bar{M} = 0$ (original PT CI $[-0.40, +1.15]$); the daily DiD estimate of −1.91 should therefore not be treated as robust to event-time aggregation. Appendix Figure 13 presents the sensitivity plots.

[15]The ban-period content estimates are reduced-form effects that cannot fully separate the platform supply shock from concurrent electoral dynamics. The first round of the 2024 municipal elections (October 6) falls inside the ban window; common electoral forces are differenced out because both treated and control groups face identical electoral conditions. However, if political engagement amplifies misinformation and aggression rhetoric above the platform-supply baseline, the DiD estimates overstate the platform-specific component. Two features bound this concern: the DiD contrasts high- versus low-X-intensity groups rather than political versus non-political groups, and Table 4 shows that non-political groups (sales, religion, sports; $N = 76$) exhibit no significant content decline despite a clean X-link first stage, consistent with the effects operating primarily through the X-supply channel.

cline from the pre-ban mean of 5.62. Left-wing domains fell by 0.249 (SE = 0.074, $p < 0.01$), a 26.6% decline from 0.935. Mainstream news links also declined (-0.73 per group-day, SE = 0.22, $p < 0.01$; pre-mean 2.70; Appendix Table 21). Table 3 reports a more restrictive mainstream-news subaggregation (pre-mean 1.25, coefficient -0.32, -25.7%) that omits domains classified elsewhere in the platform-substitution categorisation; the proportional decline is similar. The ban did not selectively suppress one side of the political spectrum; it reduced ideological media sharing across the board.

Fact-checking links declined by 0.033 per group-day (SE = 0.016, $p < 0.05$), a 47.3% drop from the small pre-mean of 0.070. The ban reduced not just misinformation but also the corrective discourse that responds to it: when conspiracy rhetoric fell, so did fact-checks. Misinformation flows dropped more than four times as much as fact-checking flows in absolute terms (0.144 vs. 0.033 per group-day), while the proportional decline in fact-checking is larger. Section 7 returns to the welfare implications. The main findings in Table 2 survive Benjamini-Hochberg correction within outcome families ($q = 0.001$ for X-links, $q = 0.005$ for forwarded messages, potential-misinformation patterns retains significance after adjustment).

5.4 Heterogeneity by Group Type and Orientation

The ban's first stage (the decline in X-link sharing) could reflect dynamics specific to political groups, which dominate our sample and are the most likely to share X content. Or it could reflect a broader disruption that affects any group that relied on X as an upstream content source. Table 4 tests this by estimating the ban-period coefficient separately by group type and, within political groups, by political orientation.

The X-link first stage appears in every group type that has sufficient sample size. Political groups ($N = 243$) show a decline of 3.01 links per group-day (SE = 0.44, $p < 0.01$). Sales groups ($N = 42$) show a smaller decline of 1.40 links (SE = 0.33, $p < 0.01$). The merged Religion + Sports bucket ($N = 30$) shows an average decline around three links per group-day, the largest in absolute terms. The presence of a first stage in non-political groups (sales,

religion, sports) confirms that the ban operated through X content dependency, not through political mobilization. An election confound cannot explain why religious or sports groups with high pre-ban X-link sharing reduced their X content during the ban.

The downstream effects differ by group type. Political groups show significant declines in forwarded messages, aggression, and potential misinformation patterns. Non-political groups show a clean first stage on X-links but muted downstream effects. The asymmetry reinforces the pipeline interpretation from Section 5.2: X content entered WhatsApp primarily through political channels, and the behavioral consequences of the ban concentrated where the X-to-WhatsApp pipeline was densest. Political groups used X as a source of politically charged content—misinformation framing, aggressive commentary, ideological media—that lacks a close substitute within WhatsApp itself; non-political groups drew on X more incidentally, so losing it reduced link counts without disrupting the rhetorical tone of the group.

Within political groups, we further disaggregate by orientation. Left-wing groups ($N = 42$) show the largest X-link decline in absolute terms (-10.00 links per group-day, SE $= 2.90$, $p < 0.01$), reflecting much higher baseline X-link sharing. Right-wing groups ($N = 131$) show a smaller X-link first stage (-2.07, SE $= 0.36$) but a broader downstream collapse across forwarded messages, aggression, and potential misinformation. Mixed-orientation groups ($N = 70$) sit between the two. The right-wing asymmetry—smaller X-link first stage but broader downstream collapse—is consistent with X functioning as a centralized rhetorical hub in the Brazilian right-wing ecosystem: right-wing groups relied on X not only for link volume but for the misinformation and aggression cues that arrived packaged within those links.

Appendix Table 23 reports joint F-tests of pre-ban leads (weeks -8, -4, -2, -1) for each cell in Table 4. Most buckets show clean pre-trends at conventional thresholds, but three cells for the X-link outcome reject the null of flat leads: politics ($F = 6.15$, $p < 0.01$), sports ($F = 5.86$, $p < 0.01$), and right-wing political groups ($F = 8.44$, $p < 0.01$). The rejection traces to the public escalation of the STF–X conflict in late August 2024, which

produced a temporary pre-ban rise in X-link sharing concentrated in politically engaged groups. Two points mitigate the concern. First, the pre-ban rise is upward while the ban-period effect is downward and much larger, so any pre-trend bias works against finding an effect rather than toward it. Second, the rejection is outcome- and bucket-specific; forwarded messages, potential misinformation, aggression, and fact-check links pass the pre-trend test in all political and non-political buckets with one marginal exception.

Table 4: Ban Effect by Group Type and Orientation (outcomes in rows, subsets in columns)

Group type	N	X-links	Forwarded	Fake news	Aggression	Fact-check
Politics	243	-3.01^{***}	-17.68^{***}	-0.18^{***}	-2.49^{**}	-0.05^{**}
		(0.44)	(4.35)	(0.06)	(1.03)	(0.02)
Sales	42	-1.40^{***}	$+6.52$	-0.02	$+0.11$	$+0.02$
		(0.33)	(5.51)	(0.02)	(0.22)	(0.02)
Religion	17	-3.79^{**}	-7.60	-0.10	-0.60	-0.04
		(1.43)	(5.37)	(0.10)	(0.66)	(0.04)
Sports	13	-2.03^{***}	-1.27	-0.02	$+0.59$	$+0.00$
		(0.64)	(1.87)	(0.01)	(1.31)	(0.00)

Notes: This table presents ban-period DiD estimates on the 319-group analytical sample split by group type and, within political groups, by orientation. Outcomes appear in rows; group subsets in columns. Each cell reports the coefficient on Treated × Ban (per group-day) from a separate regression of Equation 1 estimated on the corresponding subsample, with the group-clustered standard error in parentheses below. Group sizes are shown under each column header. The first five rows (*X-links*, *Forwarded*, *Potential misinformation*, *Aggression*, *Fact-check links*) use the full-year panel (Jan 1–Dec 29, 2024) and are constructed from activity and dictionary-classifier outcomes and from domain classification as in Table 2. The *Fake news* row counts messages matched, before the debunking publication date, to claims verified as false by Brazilian fact-check agencies, using the matching procedure described in Section 3.3 and Appendix B. The Religion + Sports column reports a weighted average of the two group types. Treated is defined as above-median pre-ban X-link intensity within the canonical 1–100 sample. Group and calendar-day fixed effects; standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.5 Who Responds? Sender Heterogeneity

The aggregate effects are consistent with two stories: many senders each reducing slightly, or a small set of brokers losing their platform. The decomposition shows the latter. The aggregate results pool across all senders within a group. But messaging activity in WhatsApp groups is highly concentrated: the top 1% of senders by pre-ban message volume account for a disproportionate share of all messages, and the top 0.1% are even more concentrated.

If the ban's effect is driven by a small number of heavy senders, the aggregate coefficient masks important variation in who responds to the supply shock and how.

We decompose the effect by sender activity level. Each sender active during the clean January to May 2024 window (a window that precedes the June to August pre-treatment measurement period and is therefore not contaminated by treatment-period dynamics) is assigned to one of eight buckets based on total January to May message count. Expressed as messages per week over the 22-week window, the thresholds are approximately 1.2 msgs/wk (Q1 cutoff), 2.8 (Q2), 8.5 (Q3), 28 (Q4), 59 (top 10%), 197 (top 5%), 630 (top 1%), and more than 633 msgs/wk in the top 0.1%. We then estimate Equation 1 at the sender-group-day level, replacing the group fixed effect with a sender-by-group fixed effect, and cluster standard errors at the group level.

We study four outcomes: (i) total messages per sender-group-day, (ii) posted X-link (binary, equal to one if the sender shared at least one X/Twitter link on the day), (iii) forwarded messages per sender-group-day, and (iv) posted potential misinformation (binary, equal to one if the sender posted a message flagged by the potential misinformation classifier). For X-links we report the binary rather than the daily count because individual senders rarely post more than one X-link per day, which makes the binary the cleaner cross-sender comparison. The X-link count is available at the group-day level from Table 2; the sender-level binary in Figure 3 and Figure 9 is the natural sender analog.

Figure 3 and Figure 9 report the ban-period coefficients by bucket with 95% confidence intervals. The X-link first stage rises monotonically across the activity distribution. Q1 senders show no detectable change. Q2 through Q4 show small but precise declines. The top tail drives the aggregate effect: top 5% senders reduced the probability of posting an X-link by 0.48 percentage points per sender-day (SE = 0.12, $p < 0.01$), top 1% by 1.19 percentage points (SE = 0.33, $p < 0.01$), and top 0.1% by 3.00 percentage points (SE = 1.57, $p < 0.10$). The gradient is the visual confirmation of the pipeline story: a small population of high-volume senders drove the aggregate X-link decline.

Forwarded-message effects concentrate even more narrowly. The top 0.1% bucket shows a decline of 1.28 forwarded messages per sender-day (pre-ban treated mean: 1.43; $p < 0.01$), a 90% reduction: a small population of hyperforwarders whose behaviour drives the aggregate forwarded-message collapse. Other buckets show null or small coefficients. Posted potential misinformation effects show the same top-heavy pattern. Total messages per sender-day decline modestly across buckets, with the strongest effects in the top 1% and top 5%.

The profile of the top-tail senders is distinctive. They are active in several groups simultaneously, produce more original than forwarded content, and derive over half their content from X. This signature matches the "information brokers" described in the literature on WhatsApp's role in Brazilian elections: individuals or organisations that maintain a presence across multiple WhatsApp groups and distribute content sourced from X. We do not label these senders as automated accounts (the data cannot distinguish coordinated campaigns from highly engaged individuals), but their behavioural response to the ban is consistent with a small, highly concentrated set of content producers who lost their upstream platform and reduced their cross-group activity accordingly.

Appendix Table 24 reports joint F-tests of pre-ban leads for each bucket-outcome pair. Most cells (27 of 32) pass the pre-trend test at the 5% level. The one persistent rejection is the top-5% bucket on the posted-X-link outcome ($F = 3.70$, $p < 0.01$), which traces to the same pre-ban political escalation that drives the group-level rejections discussed in Section 5.4: politically active heavy senders increased their X-link posting in the final weeks of August. The rejection works against finding a ban effect, not toward it. Additional robustness checks on the sender analysis (within-treatment bucketing, dominance cutoff sensitivity, continuous dose specifications, log-bin bucket definitions, and a selection check on dropped senders) appear in Appendix F and confirm the concentration of first-stage effects in the top tail.

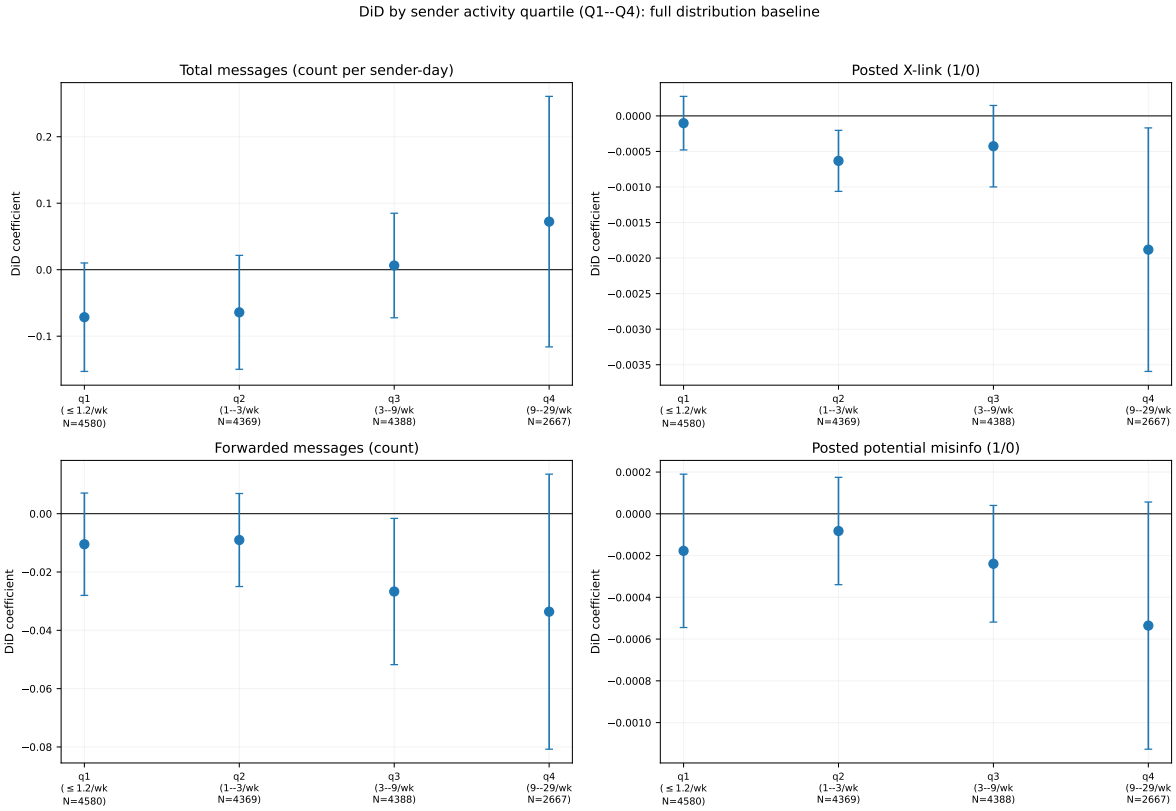


Figure 3: Ban effect by sender activity quartile (Q1–Q4): full distribution baseline

Notes: This figure presents ban-period DiD coefficients for four outcomes, one panel per outcome, across the four sender activity quartiles (Q1–Q4) defined on total messages sent during the clean pre-ban window (January 1 to May 31, 2024). Bucket labels show the messages-per-week range and the number of distinct senders in each quartile. Whiskers are 95% confidence intervals from group-clustered standard errors. "Messages" is the total-message count per sender-day. "Posted X-link" is an indicator equal to one if the sender shared at least one X/Twitter link on the day. "Forwarded" is the forwarded-message count per sender-day. "Posted potential misinformation" is an indicator equal to one if the sender posted a message flagged by the potential misinformation classifier. The heavy-tail zoom on the most active senders (top 10%, 5%, 1%, and 0.1%) is reported in Figure 9.

5.6 What Happens After the Ban?

The post-ban analysis is informative but mixed; the election-cycle decomposition in Table 5 distinguishes platform-level recovery from electoral-cycle deflation.

The ban ends on October 8. For the two and a half months that follow, two forces operate in parallel: the reduction in X-link sharing identified during the ban attenuates as access is restored, and the 2024 municipal election cycle completes itself (first round October 6, runoff October 27 for the 51 cities eligible to hold one). We want to know whether the

ban-period reductions persist, recover, or mix with these post-election forces. To separate them, we re-estimate Equation 1 on the ban and post-ban sample only (dropping the pre-ban period) so that the ban period becomes the reference, using a *nested* specification so that the runoff-window coefficient reads as an incremental shift on top of the post-first-round level:

$$Y_{gd} = \alpha_g + \delta_d + \gamma_1 \left(\text{treated}_g \times \text{Post1st}_d\right) + \gamma_2 \left(\text{treated}_g \times \text{PostRunoff}_d\right) + \varepsilon_{gd}, \tag{3}$$

where Post1st_d equals one for all days from October 9 onward (the full post-ban window) and PostRunoff_d equals one for days from October 28 onward (nested inside Post1st_d). Under this nested coding, γ_1 is the treated-control shift in the October 9 to October 27 window versus the ban-period baseline, and γ_2 is the *additional* shift once the runoff + holiday window begins; the total treated-control differential for the post-runoff window relative to the ban baseline is $\gamma_1 + \gamma_2$.

Table 5 reports the results in three panels: activity, content, and platform substitution. Panel A examines the activity outcomes from Table 2. Forwarded messages decline in the post-first-round window in both political ($\gamma_1 = -5.13$, SE $= 2.29$, $p < 0.05$) and non-political ($\gamma_1 = -8.83$, SE $= 4.66$, $p < 0.10$) groups, with the larger point estimate appearing in non-political groups. This pattern is consistent with a broader election-cycle deflation rather than a politics-specific channel: both group types continue to contract activity beyond the ban baseline in the window immediately after October 8. Total messages contract further in the post-first-round window within political groups ($\gamma_1 = -33.09$, SE $= 13.30$, $p < 0.05$), with no additional contraction by the post-runoff period ($\gamma_2 = +27.77$, SE $= 16.58$, $p < 0.10$, nearly offsetting the Post1st decline). Non-political groups show no significant additional shift in total messages at either horizon. The activity deflation is concentrated in the October 9 to October 27 window, not in the post-runoff + holiday window.

Panel B examines content outcomes. Potential misinformation rhetoric falls further beyond the ban-period level in the post-first-round window ($\gamma_1 = -0.175$, SE $= 0.035$, $p < 0.01$

for political groups) and does not move further in the post-runoff window ($\gamma_2 = +0.034$, insignificant). The broad potential misinformation measure follows the same pattern. Even as X-link sharing recovers (Panel C), the rhetorical templates the X-to-WhatsApp pipeline carried do not rebound. Potential misinformation declines are sharper in political groups ($\gamma_1 = -0.175$) than in non-political groups ($\gamma_1 = -0.069$, insignificant), consistent with the interpretation that the election cycle partially masks the post-ban content effect: political groups that would otherwise have continued producing election rhetoric deflate more because both the platform-level reduction and the post-election cooldown push in the same direction. Hate speech and fact-check links are null in both post-ban windows.

Panel C reports the platform substitution dynamics. X/Twitter links are flat in the post-first-round window within political groups ($\gamma_1 = +0.40$, SE = 0.36, insignificant) and jump incrementally in the post-runoff + holiday window ($\gamma_2 = +2.40$, SE = 0.77, $p < 0.01$), reaching a total post-runoff rebound of +2.80 relative to the ban-period baseline and closing most of the ban-period gap. Non-political groups show no X-link recovery at either horizon ($\gamma_2 = +0.03$, insignificant), an asymmetry consistent with political accounts re-entering X in the weeks after the municipal campaigns had closed. Bluesky and Threads link-sharing drops back toward pre-ban baseline in both post-ban windows, confirming that the substitution documented in Section 5.2 was temporary.

Table 5: After the ban: ban-referenced re-estimation, by political/non-political split

Outcome	All groups		Political only		Non-political only	
	Post 1st rd. (vs ban)	Post runoff (incr.)	Post 1st rd. (vs ban)	Post runoff (incr.)	Post 1st rd. (vs ban)	Post runoff (incr.)
Panel A: Activity						
X/Twitter links	+0.386	+1.800***	+0.403	+2.399***	+0.056	+0.030
	(0.291)	(0.591)	(0.364)	(0.774)	(0.410)	(0.282)
Forwarded	−5.95***	+3.45	−5.13**	+2.21	−8.83*	+7.51
	(2.07)	(2.68)	(2.29)	(3.11)	(4.66)	(5.59)
Total messages	−25.04	+12.81	−33.09**	+27.77*	−6.57	−25.75
	(17.75)	(21.31)	(13.30)	(16.58)	(58.35)	(67.23)
Unique senders	−2.05*	+0.53	−1.33	−0.70	−4.59	+4.96
	(1.15)	(1.40)	(1.22)	(1.55)	(2.88)	(3.19)
Panel B: Content						
Potential misinformation	−0.1476***	+0.0349	−0.1752***	+0.0344	−0.0688	+0.0417
	(0.0335)	(0.0229)	(0.0354)	(0.0246)	(0.0807)	(0.0587)
Potential misinformation (broad)	−0.1835***	+0.0879***	−0.1704***	+0.0534	−0.2427**	+0.1927**
	(0.0392)	(0.0336)	(0.0423)	(0.0357)	(0.1012)	(0.0807)
Aggression	−0.4951*	+0.1155	−0.6266**	+0.3831	−0.2038	−0.6719
	(0.3000)	(0.3473)	(0.2908)	(0.3356)	(0.8075)	(0.9340)
Hate speech (broad)	+0.0094	−0.0005	+0.0077	+0.0036	+0.0108	−0.0113
	(0.0229)	(0.0204)	(0.0280)	(0.0250)	(0.0293)	(0.0283)
Fact-check links	−0.0019	−0.0032	+0.0026	−0.0046	−0.0158	+0.0017
	(0.0149)	(0.0167)	(0.0189)	(0.0219)	(0.0195)	(0.0057)
Panel C: Platforms						
X / Twitter links	+0.405	+1.812***	+0.421	+2.417***	+0.079	+0.024
	(0.293)	(0.593)	(0.366)	(0.776)	(0.412)	(0.283)
Bluesky links	−0.1186***	−0.0835***	−0.0880*	−0.1091***	−0.2160**	−0.0078
	(0.0443)	(0.0253)	(0.0498)	(0.0328)	(0.0949)	(0.0123)
Threads links	−0.0670	−0.1787***	−0.0585	−0.2374***	−0.0845*	+0.0048
	(0.0572)	(0.0558)	(0.0723)	(0.0728)	(0.0494)	(0.0140)
Bluesky + Threads	−0.186**	−0.262***	−0.147	−0.347***	−0.301**	−0.003
	(0.093)	(0.078)	(0.114)	(0.101)	(0.126)	(0.020)

Notes: This table presents DiD estimates from Equation 3 on the ban + post-ban sample of the 319-group analytical sample (pre-ban period dropped). Treated is defined as above-median pre-ban X-link intensity. The ban period (August 30 to October 8, 2024) is the omitted reference. The coefficients are reported under a *nested* coding: *Post 1st rd. (vs ban)* is γ_1, the treated-control shift in the October 9 to October 27 window relative to the ban baseline, and *Post runoff (incr.)* is γ_2, the *additional* shift on top of γ_1 during the October 28 onward window. The total treated-control shift relative to the ban baseline in the post-runoff window is $\gamma_1 + \gamma_2$. Each row is a separate regression with the listed outcome as the dependent variable, per group-day. Columns 1–2 pool all 319 groups; columns 3–4 restrict to the 243 political groups; columns 5–6 restrict to the 76 non-political groups (the 30 religion + sports groups and the 42 sales groups displayed in Table 4, plus 4 small-N non-political groups in health, education, condominium, and adult-content buckets that are aggregated here but omitted from the by-type display). Group and calendar-day fixed effects; standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6 Robustness and Placebos

We present robustness checks organized by the type of concern each addresses—sample construction, specification choice, and inference method—and close with a treatment-definition placebo that probes whether the X-link decline reflects X-specific intensity or a broader

dimension of group activity.

Threshold sensitivity. Because the analytical sample requires groups to average at least one X link per week and at most 100 X links per week, the treatment indicator is a binary function of the same count used to filter the sample. Table 6 reports the X-link ban-period coefficients across alternative left-tail thresholds (≥ 0.5, ≥ 1.0, ≥ 2.0) and right-tail caps (none, 200, 100, 50, 30). The first stage grows monotonically as the right-tail cap loosens—from -2.799 at the cap of 100 to -4.223 with no cap—reflecting the contribution of the six groups with pre-ban X-links per week above 100 ("hub outliers," whose pre-ban trajectories diverge from the rest of the sample) that we exclude from the primary specification. Across all combinations the X-link coefficient remains negative, large, and significant at the 1% level: the main result is not an artifact of the sample-filter choice.

Table 6: Robustness: X-link ban-period coefficients across alternative sample definitions

Truncation (X-links/week)	N groups	Ban effect on X-links	
		Coef.	SE
≥ 1, no cap	325	-4.223^{***}	(0.743)
≥ 1, ≤ 200	322	-3.174^{***}	(0.427)
≥ 1, ≤ 100 (primary)	319	-2.799^{***}	(0.348)
≥ 1, ≤ 50	312	-2.078^{***}	(0.209)
≥ 1, ≤ 30	297	-1.524^{***}	(0.160)
≥ 0.5, ≤ 100	394	-2.330^{***}	(0.292)
≥ 2, ≤ 100	247	-3.108^{***}	(0.444)

Notes: This table presents ban-period DiD coefficients for X-link sharing under alternative sample-filter combinations on the 319-group analytical sample. Each row reports the ban coefficient for a different combination of the left-tail threshold (minimum pre-ban X-links per week to qualify for the sample) and the right-tail cap (maximum pre-ban X-links per week). The primary specification uses ≥ 1, ≤ 100 and yields $N = 319$ groups. Group and calendar-day fixed effects; standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Continuous-dose specification on the full corpus. To rule out that the sample-construction step drives the headline X-link result, we re-estimate on the full 25,814-group corpus using three continuous dose definitions: $\log(1 + x\_per\_week)$, raw x_per_week, and

percentile rank. Groups with zero pre-ban X-links receive dose zero and act as null-dose controls. Table 7 reports results for both the full corpus (Panel A) and the 319-group canonical sample (Panel B). The X-link coefficient is negative and significant at $p < 0.01$ in every row. The log-dose coefficient attenuates from −2.583 (canonical sample) to −1.930 (full corpus), consistent with the large fraction of zero-dose groups diluting the slope; all five headline outcomes remain negative and significant in Panel A. The sample selection is therefore not what produces the X-link result.

Table 7: Robustness: continuous-dose ban-period coefficients, full corpus vs. 319-group canonical sample

Dose definition	X-links	Forwarded	Messages	With links	Misinform.
Panel A: Full corpus (N groups = 25,814, N group-days = 1,452,195)					
Pre-ban mean (top-half *x_per_week*)	0.964	24.31	140.51	32.62	0.196
$\log(1 + x\_per\_week)$	-1.930^{***}	-4.676^{***}	-21.879^{***}	-4.440^{***}	-0.411^{***}
(SE)	(0.336)	(0.954)	(5.744)	(0.862)	(0.143)
x_per_week	-0.166^{***}	-0.136^{**}	-0.545^{*}	-0.128^{**}	-0.071^{***}
(SE)	(0.015)	(0.065)	(0.291)	(0.061)	(0.008)
Percentile rank of *x_per_week*	-2.234^{***}	-10.843^{***}	-51.489^{***}	-9.472^{**}	-0.341^{***}
(SE)	(0.347)	(2.071)	(12.327)	(4.112)	(0.115)
Panel B: 319-group canonical sample (N groups = 319, N group-days = 72,672)					
Pre-ban mean (top-half *x_per_week*)	3.088	50.26	251.97	56.59	0.464
$\log(1 + x\_per\_week)$	-2.583^{***}	-7.833^{***}	-25.264^{*}	-5.730^{**}	-0.128^{***}
(SE)	(0.393)	(2.552)	(15.244)	(2.735)	(0.048)
x_per_week	-0.219^{***}	-0.535^{**}	-1.773	-0.492^{*}	-0.011^{*}
(SE)	(0.039)	(0.256)	(1.211)	(0.250)	(0.006)
Percentile rank of *x_per_week*	-6.222^{***}	-21.702^{***}	-68.797	-13.666^{*}	-0.315^{***}
(SE)	(0.883)	(6.746)	(42.314)	(7.972)	(0.101)

Notes: Ban-period DiD coefficients from $y_{g,d} = \alpha_g + \delta_d + \beta\,(\text{dose}_g \times \text{Ban}_d) + \theta\,(\text{dose}_g \times \text{Post}_d) + \varepsilon_{g,d}$, estimated separately for each dose definition and outcome. Ban is the indicator for August 30 to October 8, 2024. *x_per_week* is computed from *n_twitter_links* summed over June 1 to August 29, 2024, divided by 13 weeks. Group and calendar-day fixed effects; standard errors clustered at the group level. Outcomes: X-links (*n_twitter_links*), Forwarded (*n_forwarded*), Messages (*n_messages*), With links (*n_with_links*), Misinform. (*fakenews_pattern_v2*). Panel A uses all corpus groups (including the 24,904 with zero pre-ban X-links, which receive dose = 0 under all three definitions). Panel B uses the 319-group canonical sample described in Section 3.2. Pre-ban means reported for the subset of within-panel groups with *x_per_week* above the within-panel median. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Event-study on the full corpus. Figure 4 extends the daily event study to the full 25,814-group corpus using a continuous log-dose. Pre-period coefficients cluster near zero in the 25 days before ban onset; the coefficient drops sharply at $k = 0$ and recovers partially after

$k = 39$, replicating the canonical-sample signature with attenuated magnitudes consistent with the large zero-dose mass in the full corpus.

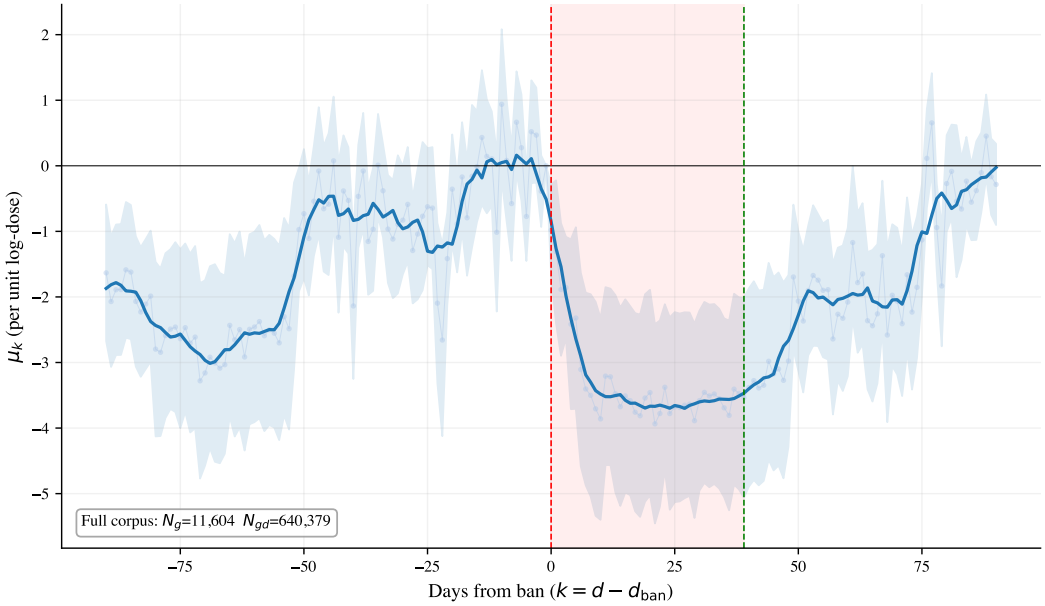


Figure 4: Event-study on the full 25,814-group corpus with continuous log-dose

Notes: Daily event-study coefficients μ_k from $Y_{gd} = \alpha_g + \delta_d + \sum_{k \neq -1} \mu_k(\text{dose}_g \times \mathbb{1}[d - d_{\text{ban}} = k]) + \varepsilon_{gd}$ on the full 25,814-group corpus (unbalanced panel; $N_g = 11{,}604$ groups with at least one observed day in the $k \in [-90, +90]$ window; $N_{gd} = 640{,}379$ group-day observations), with $\text{dose}_g = \log(1 + x\_per\_week_g)$. Faint markers show raw daily μ_k; the bold blue line is the 7-day centred moving average; the shaded band is the 95% confidence interval from group-clustered standard errors. The reference day is $k = -1$. The dashed red vertical line marks ban onset ($k = 0$, August 30, 2024); the dashed green vertical line marks ban end ($k = 39$, October 8). The shaded red region is the ban window. Coefficients are expressed per unit of log-dose: a group with $x\_per\_week = e - 1 \approx 1.72$ (one log-dose unit above a zero-dose group) experiences μ_k additional X-links per group-day at event-time k relative to the reference day.

Specification sensitivity. Because treated groups contain a higher share of politically oriented groups, political-cycle trends could contaminate the ban-period coefficient if not absorbed by the fixed-effects structure. The X-link ban-period coefficient is nonetheless stable across alternative fixed-effects and control choices. All four specifications agree within 1.3% of the primary estimate of -2.799 (SE $= 0.348$); details are in Appendix G.3 (Table 22). The calendar-day-FE specification is preferred for parsimony and for retaining the maximum amount of time-fixed variation. Appendix G.1 (Table 20) additionally shows that the main estimates are stable under propensity-score matching and inverse probability weighting, addressing the pre-existing activity-level differences between treated and control groups; the

one exception is total messages under IPW, whose point estimate falls to -28.0 and loses significance, consistent with that outcome's sensitivity to group-size variation.

Serial correlation. We assess robustness of the headline X-link coefficient to alternative treatments of serial correlation, following Bertrand, Duflo, and Mullainathan (2004), who show that clustering alone can understate sampling variance in long daily panels. Table 8 re-estimates $\widehat{\beta}_{\text{ban}}$ across five outcomes under four variance estimators: group-clustered standard errors (baseline), weekly aggregation to ISO-weeks with group and week fixed effects, Driscoll–Kraay HAC (Driscoll and Kraay, 1998) with a Bartlett kernel and bandwidth of 14 days (two calendar weeks, covering the dominant autocorrelation horizon in daily panels of social media activity), and a block cluster bootstrap that resamples whole groups with replacement over 1,000 replications. For X/Twitter links, the baseline estimate is $\widehat{\beta}_{\text{ban}} = -2.799$ ($p < 0.001$); weekly aggregation yields -17.972 ($p < 0.001$, larger in absolute value because weekly outcomes sum over 4–7 days); Driscoll–Kraay HAC yields the same point estimate with SE $= 0.475$ ($p < 0.001$); and the block bootstrap 95% percentile interval is $[-3.522, -2.188]$. Across every outcome in the table the $\widehat{\beta}_{\text{ban}}$ point estimate retains its sign and order of magnitude, and the headline X-link effect remains significant at the one-percent level under all four alternatives. Among the other outcomes, total messages is the only one whose block-bootstrap interval crosses zero ($[-89.5, +2.2]$); the remaining four outcomes stay significant under every variance estimator.

Table 8: Serial-correlation robustness: ban-period DiD coefficients under four variance estimators

	X-links	Forwarded	Messages	With links	Misinform.
Panel A: Group-clustered SE (baseline)					
$\widehat{\beta}_{\text{ban}}$	-2.799^{***}	-13.078^{***}	-41.488^{*}	-7.712^{**}	-0.144^{***}
(SE)	(0.348)	(3.437)	(23.059)	(3.646)	(0.047)
Obs.	72,672	72,672	72,672	72,672	72,672
Panel B: Weekly aggregation (group-clustered)					
$\widehat{\beta}_{\text{ban}}$	-17.972^{***}	-104.661^{***}	-384.631^{**}	-74.264^{***}	-1.221^{***}
(SE)	(2.222)	(27.701)	(180.776)	(25.984)	(0.298)
Obs.	12,531	12,531	12,531	12,531	12,531
Panel C: Driscoll-Kraay HAC (Bartlett, bw=14 days)					
$\widehat{\beta}_{\text{ban}}$	-2.799^{***}	-13.078^{***}	-41.488^{***}	-7.712^{***}	-0.144^{***}
(SE)	(0.475)	(2.388)	(8.956)	(1.459)	(0.038)
Obs.	72,672	72,672	72,672	72,672	72,672
Panel D: Block cluster bootstrap (1000 reps)					
$\widehat{\beta}_{\text{ban}}$	-2.799^{***}	-13.078^{***}	-41.488^{*}	-7.712^{**}	-0.144^{***}
[95% CI]	[−3.522, −2.188]	[−19.994, −6.523]	[−89.516, 2.240]	[−14.770, −0.504]	[−0.237, −0.050]
Obs.	72,672	72,672	72,672	72,672	72,672

Notes: Each entry is the ban-period Treated × Ban coefficient $\widehat{\beta}_{\text{ban}}$ from Equation 1 estimated on the 319-group canonical sample, one regression per (outcome, panel) cell. Outcomes are X/Twitter links (*n_twitter_links*), Forwarded messages (*n_forwarded*), Total messages (*n_messages*), Messages with links (*n_with_links*), and Potential misinformation patterns (*fakenews_pattern_v2*). Panel A reports the headline CRV1 standard errors clustered at the group level; Panel B collapses the panel to group-ISO-week totals (outcomes summed within week, group and week fixed effects, SE clustered by group); Panel C reports Driscoll-Kraay HAC standard errors using a Bartlett kernel and bandwidth of 14 days, constructed as the Newey-West covariance of cross-sectionally aggregated scores from fixed-effect-demeaned regressors; Panel D reports the percentile 95% confidence interval from a block cluster bootstrap with 1,000 replications in which whole groups are resampled with replacement. Panel B point estimates are not directly comparable to Panels A/C/D in magnitude because outcomes are summed over 4–7 days per ISO-week; sign, significance, and order of magnitude are the relevant comparison. Significance stars in Panels A–C: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Panel D uses CI-based inference and omits stars.

Randomization inference. As a distribution-free complement to the asymptotic clustered inference in Table 8, we implement a Fisher-style randomization test that permutes the binary treatment assignment under the sharp null of no effect and recomputes $\widehat{\beta}_{\text{ban}}$ for each of $R =$ 1,000 draws. In the primary design, we randomly reassign 159 of the 319 groups to the treated arm without regard for pre-ban X-activity, which yields a one-sided randomization p-value for the X-link outcome of $p_{\text{RI}} < 0.001$ (observed $\widehat{\beta} = -2.799$; Figure 5). A supplementary test that reshuffles treatment within deciles of pre-ban total-message volume — a proxy for overall group activity that is only weakly correlated with X-link intensity in our sample (Spearman $\rho \approx 0.24$) — yields $p_{\text{RI}} < 0.001$, confirming that the effect is tied specifically

to X-intensity rather than to a correlated measure of group size.[16] The observed X-link coefficient lies outside the entire range of the null distribution. Of the five main outcomes, X-links, forwarded messages, messages with links, and potential misinformation patterns survive randomization inference at the 0.05 level; total messages survive only at the 0.10 level under the primary design.

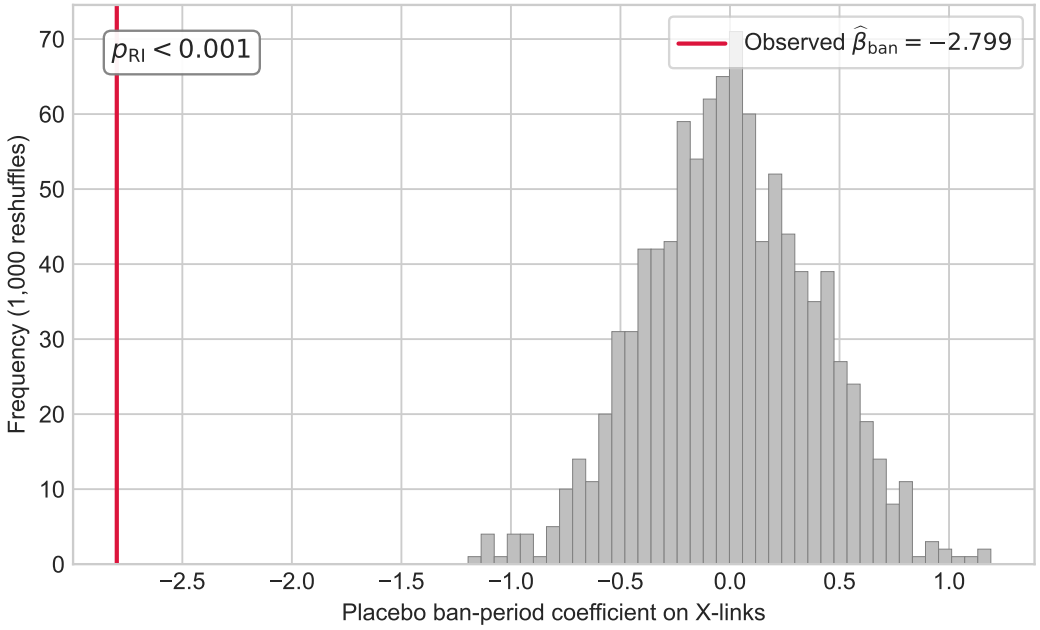


Figure 5: Randomization inference: null distribution of the ban-period X-link coefficient

Notes: Histogram of placebo ban-period coefficients $\widehat{\beta}^{(r)}_{\text{ban}}$ on the *X-links* outcome from $R = 1{,}000$ random reshuffles of the binary treatment indicator over the 319-group canonical sample (simple Fisher reshuffle: 159 treated, 160 control, drawn uniformly without regard for pre-ban X-activity). Each reshuffle re-estimates Equation 1 by the Frisch-Waugh-Lovell trick, residualising the outcome and interaction terms against chat and calendar-day fixed effects once. The red vertical line marks the observed coefficient $\widehat{\beta}_{\text{ban}} = -2.799$. The observed value falls outside the entire range of the 1,000 null draws, yielding $p_{\text{RI}} < 0.001$ (the Monte Carlo floor at $1/R$).

Robust inference under weakened parallel trends. The X-link event study in Figure 2 shows a small upward drift in the final two pre-ban weeks attributed to the public escalation of the STF–X conflict (Sections 3 and 5.1). Because that drift moves *against* the post-ban direction, it biases our DiD estimate toward zero; we nonetheless follow Rambachan and Roth (2023) and construct sensitivity bounds that explicitly allow post-ban deviations

[16] We stratify on message volume rather than on $x_{\text{per week}}$ itself because the treatment is defined as $\mathbb{1}\{x_{\text{per week}} > \text{median}\}$, which makes every $x_{\text{per week}}$ decile strictly above or strictly below the cutoff; a within-x-decile reshuffle is then deterministic.

from parallel trends. We work in their relative-magnitude (RM) class, which restricts the maximum post-treatment violation of parallel trends to at most $\bar{M}$ times the maximum pre-period violation, and report 95% robust confidence intervals for the average post-ban X-link effect at $\bar{M} \in \{0, 0.5, 1, 1.5, 2\}$ (weekly aggregation; parameter of interest: average treatment effect over post-ban weeks $k \in [0, 5]$).

The breakdown value at which the robust CI first contains zero is $\bar{M} = 1$. The negative post-ban effect remains statistically significant ($\text{CI}_{0.5} = [-10.90, -0.62]$) as long as post-ban violations are no larger than half the worst pre-period violation, and loses significance only when post-ban deviations match the most extreme pre-ban deviation ($\text{CI}_1 = [-15.49, 3.82]$). For comparison, the standard 95% CI is $[-7.07, -3.68]$ (width 3.39); the RM CI at $\bar{M} = 1$ is roughly 5.7× wider and at $\bar{M} = 2$ roughly 8.9× wider. The largest pre-ban violation falls in the two weeks of the STF–X escalation, when X-link traffic was elevated for reasons unrelated to the ban; this sets a conservative benchmark for plausible post-ban deviations.

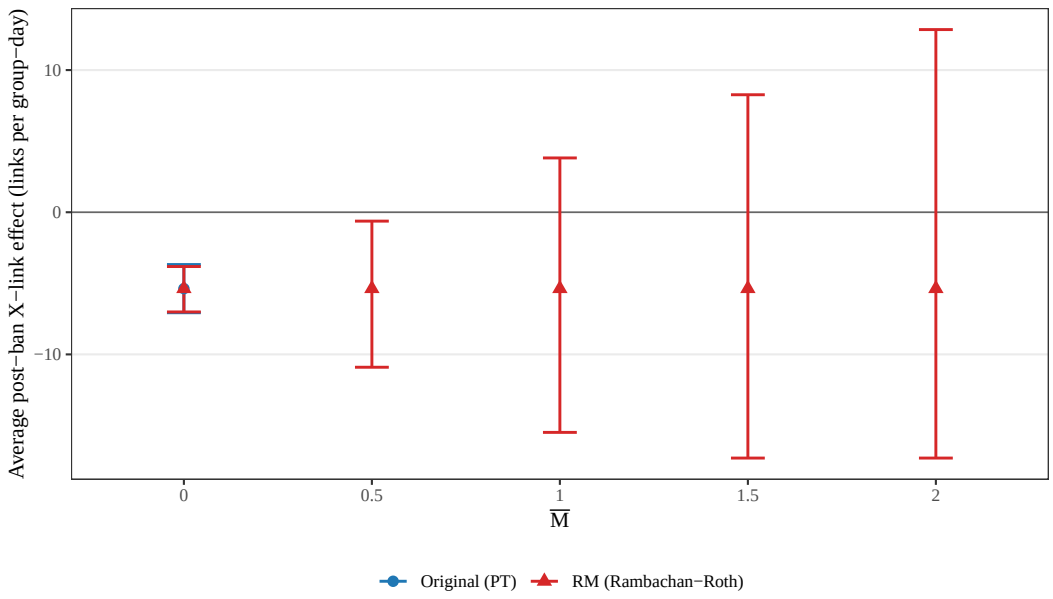


Figure 6: Rambachan-Roth sensitivity bounds on the average post-ban X-link effect

Notes: Sensitivity plot from Rambachan and Roth (2023) applied to a weekly aggregation of the canonical 319-group event study on the X-link outcome (pre-window $k \in [-13, -1]$ weeks, post-window $k \in [0, 5]$ weeks; the ban spans weeks 0 through 5). The parameter of interest is the average treatment effect over post-ban weeks 0–5. The leftmost bar labelled $\bar{M} = 0$ reports the standard parallel-trends 95% confidence interval (blue); the remaining bars (red) report relative-magnitude robust confidence intervals that allow the worst post-treatment parallel-trends violation to be at most $\bar{M}$ times the worst pre-treatment violation. The breakdown value at which the robust CI first contains zero is $\bar{M} = 1$.

Placebo across alternative treatment definitions. The ban-period X-link decline could in principle reflect a broader dimension of the group (generic platform activity, group size, or political engagement) rather than X-specific intensity. We test this by substituting alternative dimensions into the treatment definition and re-estimating the ban-period coefficient on the X-link outcome. For each candidate dimension z we compute the group-level pre-ban mean of z, take an above-median dummy within the 319-group canonical sample, and re-estimate Equation 1 with the substituted treatment indicator. Eight alternative dimensions are used: pre-ban YouTube, Facebook, Instagram, TikTok, and Telegram link counts per group-day; total messages per group-day; distinct senders per group-day; and a random 50/50 assignment as a pure-noise baseline. The real X-intensity split is reported in the same figure for comparison. If the X-link result reflects X-specific intensity, only the X-intensity split should produce a significantly negative coefficient; all other splits should hug zero.

Figure 7 reports the nine ban-period coefficients. The X-intensity split (in red) produces -2.80 per group-day ($p < 0.01$), reproducing the headline result in Table 2. The six non-Telegram platform and activity placebos hug zero: YouTube $+0.14$, Facebook $+0.27$, Instagram -0.32, TikTok -0.07, total messages -0.01, distinct senders -0.11, all with $p > 0.10$. The random baseline gives -0.49 (also insignificant). The one exception is the Telegram-intensity split, which produces -1.27 ($p < 0.01$). This is consistent with Telegram operating as a complement to X inside the Brazilian right-wing ecosystem rather than as a substitute, so high-Telegram groups are disproportionately high-X groups in a way the other platform splits are not.

The falsification confirms that the X-link result in Table 2 reflects X-specific intensity, not generic social-media intensity, overall activity, or group size. The exercise speaks to the X-link outcome, which is mechanically tied to X access. It does not extend cleanly to the other outcomes in Table 2: the ban window contains the first round of Brazil's 2024 municipal elections (October 6), so any dimension correlated with political engagement, including X-intensity and every platform-activity placebo, will absorb both the ban and the concurrent

electoral cycle. The design does not separate those two contributions for outcomes other than X-links. Section 5.6 treats the election-cycle mixing explicitly for the post-ban recovery.

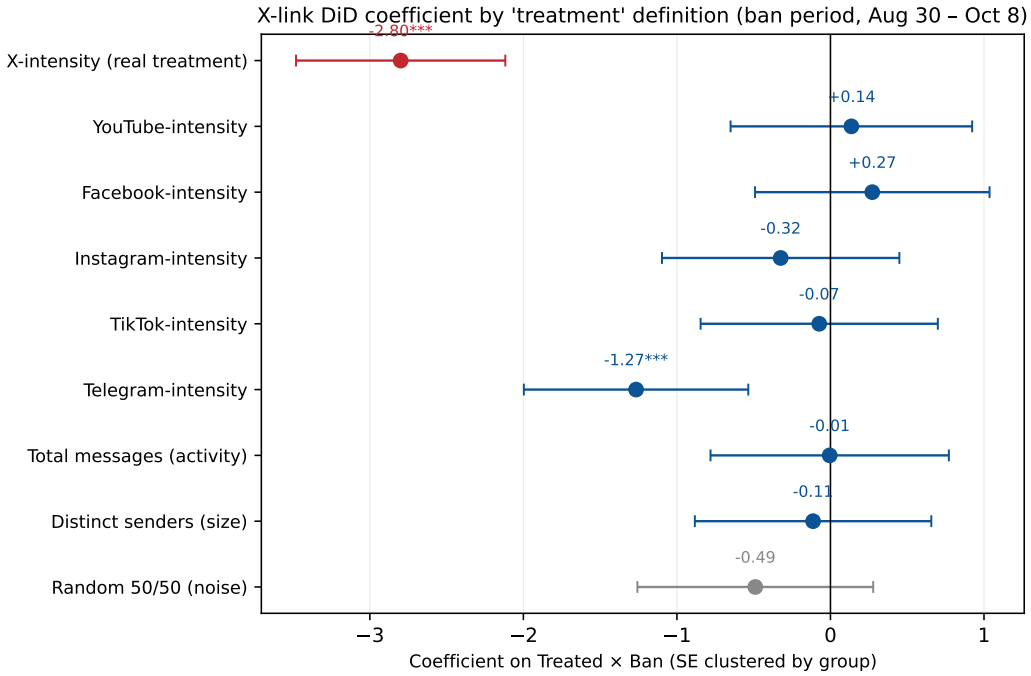


Figure 7: Placebo exercise: X-link DiD coefficient under alternative treatment definitions

Notes: This figure reports the ban-period DiD coefficient on X/Twitter links per group-day when the treated indicator is redefined on alternative group-level dimensions. Each point is a separate regression of the form $n\_twitter\_links_{g,d} = \alpha_g + \gamma_d + \beta \cdot T_g \cdot \text{Ban}_d + \varepsilon_{g,d}$ estimated on the 319-group canonical sample, where T_g is one above the within-sample median of pre-ban z and Ban is the indicator for August 30 to October 8, 2024. The dimensions are pre-ban X-link intensity (red, real treatment); YouTube, Facebook, Instagram, TikTok, and Telegram link counts per group-day (blue); total messages per group-day (blue); distinct senders per group-day (blue); and a random 50/50 assignment (grey). Whiskers are 95% confidence intervals from standard errors clustered at the group level. Only the X-intensity split produces the headline result; the Telegram coefficient is non-zero because Telegram and X intensity co-vary within the Brazilian right-wing media ecosystem.

7 Discussion

Our results support a complementarity interpretation of how upstream platforms shape closed messaging networks. When X was removed as a source of shareable content, the downstream network responded primarily by quieting rather than by reallocating activity to alternative upstream sources. About 85% of the lost X-link traffic was not recovered through any other platform, and the aggregate decline concentrates in a small set of high-intensity senders who had previously been importing X-sourced content. Combined with

the ban-period drops in forwarded messages (26%), potential misinformation (31%), and aggression (30%, though the latter is sensitive to event-time aggregation; see Appendix G.8), the patterns point to X functioning as a *complement* to WhatsApp group activity rather than as one of several substitutable upstream sources that users can swap at low cost. Removing the upstream platform reduces the downstream content stream by far more than it reallocates it.

This reading connects to a growing causal literature on platform-level interventions. Allcott, Braghieri, et al. (2020) and Allcott, Gentzkow, et al. (2024) show that voluntary individual deactivation of Facebook and Instagram reduces news consumption and political knowledge, consistent with complementarity between social media and other information sources rather than substitution across them. Fujiwara, Müller, and Schwarz (2024) find that exogenous variation in Twitter adoption affects U.S. election outcomes, implying that platform access shifts the balance of political information available to voters. McCabe et al. (2024) show that within-platform deplatforming of 70,000 misinformation accounts on Twitter after January 6 reduced misinformation reach on the platform, and Ventura et al. (2025) show that reducing an individual user's exposure to WhatsApp multimedia during the 2022 Brazilian election lowered recall of prominent false claims. Our paper complements this work by measuring the content response of an entire closed messaging network when the upstream open platform is removed at the population level: where the existing literature identifies effects on individual receivers or on within-platform content, we identify effects on the downstream ecosystem as a whole. The placebo exercise (Section 6) reinforces this reading. Only X-intensity and, to a smaller extent, Telegram-intensity reproduce the ban-period X-link decline; no other platform, activity, or group-size dimension replicates the pattern. The Telegram result is consistent with Telegram's correlation with X inside the Brazilian right-wing ecosystem rather than with generic cross-platform substitution.

Policy implications. Several implications follow. First, platform bans produce durable reductions in content flow on adjacent networks, with limited cross-platform substitution

in the short run. Policymakers considering platform bans should expect the intervention to quiet downstream messaging networks rather than to displace content to less-regulated alternatives. This contrasts with evidence from the within-fringe deplatforming literature: Horta Ribeiro, Hosseinmardi, et al. (2023) show that the deplatforming of Parler in early 2021 was followed by a large increase in user activity on other fringe social-media platforms, an outcome consistent with substitutable alternative platforms. In our setting the alternatives are either absent (Mastodon, Parler, Gab, Truth Social, all essentially non-existent in Brazil) or quickly saturated (Bluesky and Threads), and the aggregate downstream content stream contracts rather than migrating. Second, platform bans are blunt instruments. The ban reduced potential misinformation by 31% but also reduced fact-check links by nearly half, and left-wing domain sharing at similar proportional rates. A welfare assessment must weigh these competing effects rather than report the reductions in isolation. Third, the aggregate ban effect is driven by a small set of high-intensity senders, active across multiple groups and sourcing much of their material from X. This suggests an intermediate-policy design: targeted interventions aimed at high-intensity information brokers, in the spirit of the within-platform deplatforming studied by McCabe et al. (2024), would deliver a large share of the content reduction with smaller collateral effects on the rest of the network.

Limitations. Our estimates have several limitations. The analysis sample represents the intensive margin of X-dependent public WhatsApp groups rather than the average Brazilian group, and the shock is a single 39-day episode rather than a steady-state regulatory regime. Our content classifiers capture rhetorical patterns rather than verifying factual accuracy, so the estimated content effects are conservative under classical measurement error. The data cover public groups only, so if the ban shifted some content from public to private channels our estimates would understate the corresponding private-channel activity. Finally, while the ban-period coefficient on X-link sharing is clean (the first round of the 2024 municipal elections falls inside the ban window and affects treated and control groups symmetrically, and the placebo exercise in Section 6 confirms that only X-intensity repro-

duces the X-link decline), the post-ban window overlaps with the closing of the electoral cycle in most municipalities; the post-ban estimates are secondary and we interpret them against the electoral-cycle decomposition in Section 5.6.

External validity. The results apply most directly to settings in which an upstream platform is a content source for downstream messaging networks, the messaging platform has public-group infrastructure that enables observation and broad content distribution, and the platform ban is sudden, comprehensive, and enforced. These conditions were met in Brazil in 2024 and plausibly apply to other large democracies where WhatsApp or similar messaging platforms are secondary distribution channels for social media content, including India, Indonesia, Nigeria, and Turkey. India is a particularly comparable case: WhatsApp is the primary channel for political information sharing and documented misinformation campaigns (Garimella and Tyson, 2018), and the Indian digital ecosystem lacks large-scale domestic alternatives to major international platforms, so that our complementarity mechanism—upstream removal quieting downstream flow rather than redirecting it—would apply in that setting as well. The results do not speak to symmetric platform competition, gradual content-moderation policies (algorithmic downranking rather than outright bans), or private messaging channels, which differ in relevant ways from the public-group, sudden-ban setting studied here.

8 Conclusion

We study Brazil's 39-day ban of X/Twitter (August 30 to October 8, 2024) as a natural experiment on how upstream platform interventions propagate into an adjacent, closed messaging network. Using a daily difference-in-differences design that exploits cross-group variation in pre-ban X-link intensity among public WhatsApp groups, we document three main facts. First, the ban produced a large reduction in X-link sharing in treated groups (90.6%, or 2.80 links per group-day), only partially offset by limited substitution to alternative platforms:

Bluesky and Threads together recovered only 14.7% of the lost X-link supply, and other candidates often discussed in U.S. and European deplatforming debates (Mastodon, Parler, Gab, Truth Social) showed no measurable response. Second, the disruption propagated to the broader content composition of WhatsApp messaging: forwarded messages fell by 26% and potential misinformation by 31% (corroborated by an externally anchored fact-check-matched measure), with aggression also declining by 30% in the daily specification (though this estimate is sensitive to event-time aggregation; Appendix G.8). Third, the aggregate effect is concentrated in a small minority of heavy users: the top 5% of senders by pre-ban volume account for the majority of X-link sharing and drive essentially all of the aggregate decline.

A placebo exercise that redefines treatment on eight non-X pre-ban dimensions (platform intensities, total activity, group size, random assignment) confirms that only X-intensity reproduces these patterns, ruling out generic platform activity or political engagement as alternative explanations.

The three facts together suggest that X functioned as a complement to WhatsApp group activity rather than as a substitute that users can replace by multi-homing to alternative upstream platforms. The ban did not induce large cross-platform migrations: it quieted a small set of high-intensity brokers whose activity had fed the adjacent network. In ecosystems where WhatsApp and similar messaging platforms are secondary distribution channels for social-media content, interventions on the upstream platform can produce durable downstream reductions in content flow without triggering substantial substitution to alternatives.

Several directions for future work follow. Fine-tuned transformer classifiers would enable more precise content categorisation than our dictionary-based approach. Longer-run follow-up data would reveal whether the behavioural changes we document persist beyond the 39-day window or dissipate as content-sharing habits re-form. Cross-country comparisons with other platform ban episodes, in settings with different platform ecosystems and different regulatory regimes, would test whether the upstream-downstream complementarity

we identify generalises beyond the Brazilian context.

References

Ali, S., M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini (2023). "Understanding the Effect of Deplatforming on Social Networks". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). "The Welfare Effects of Social Media". In: *American Economic Review* 110.3, pp. 629–676.

Allcott, H., M. Gentzkow, W. Mason, A. Wilkins, P. Barberá, T. Brown, et al. (2024). "The Effects of Facebook and Instagram on the 2020 Election: A Deactivation Experiment". In: *Proceedings of the National Academy of Sciences* 121.21, e2321584121. DOI: 10.1073/pnas.2321584121.

Bailez, F. (2022). "Monitoring public WhatsApp groups: methodology and measurement". Working paper / methodology note for the Palver platform.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" In: *Quarterly Journal of Economics* 119.1, pp. 249–275. DOI: 10.1162/003355304772839588.

Borusyak, K., X. Jaravel, and J. Spiess (2024). "Revisiting Event-Study Designs: Robust and Efficient Estimation". In: *Review of Economic Studies* 91.6, pp. 3253–3285. DOI: 10.1093/restud/rdae007.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2024). *Social Media and Xenophobia: Evidence from Russia*. NBER Working Paper 26567. Revised February 2024. National Bureau of Economic Research. URL: https://www.nber.org/papers/w26567.

Caetano, J. A., J. M. Almeida, and H. T. Marques-Neto (2023). "Public WhatsApp Groups in Brazilian Politics: Audience and Content Patterns". In: *Online Social Networks and Media*.

Callaway, B. and P. H. Sant'Anna (2021). "Difference-in-Differences with Multiple Time Periods". In: *Journal of Econometrics* 225.2, pp. 200–230.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *Review of Economics and Statistics* 90.3, pp. 414–427.

Cetic.br (2026). *Painel TIC Integridade da Informação*. https://cetic.br/. Survey on information integrity in Brazil collected August–September 2025; 60% of Brazilian internet users access information daily through messaging apps.

Chaisemartin, C. de and X. D'Haultfoeuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110.9, pp. 2964–2996. DOI: 10.1257/aer.20181169.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017). "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. Vol. 1, pp. 1–22.

Driscoll, J. C. and A. C. Kraay (1998). "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data". In: *Review of Economics and Statistics* 80.4, pp. 549–560. DOI: 10.1162/003465398557825.

Enikolopov, R., A. Makarin, and M. Petrova (2020). "Social Media and Protest Participation: Evidence from Russia". In: *Econometrica* 88.4, pp. 1479–1514.

Fujiwara, T., K. Müller, and C. Schwarz (2024). "The Effect of Social Media on Elections: Evidence from the United States". In: *Journal of the European Economic Association* 22.3, pp. 1495–1539. DOI: 10.1093/jeea/jvad058.

Garimella, K. and G. Tyson (2018). "WhatApp, Doc? A First Look at WhatsApp Public Group Data". In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Gilardi, F., M. Alizadeh, and M. Kubli (2023). "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks". In: *Proceedings of the National Academy of Sciences* 120.30, e2305016120. DOI: 10.1073/pnas.2305016120.

Goodman-Bacon, A. (2021). "Difference-in-Differences with Variation in Treatment Timing". In: *Journal of Econometrics* 225.2, pp. 254–277. DOI: 10.1016/j.jeconom.2021.03.014.

Horta Ribeiro, M., H. Hosseinmardi, R. West, and D. J. Watts (2023). "Deplatforming Did Not Decrease Parler Users' Activity on Fringe Social Media". In: *PNAS Nexus* 2.3, pgad035. DOI: 10.1093/pnasnexus/pgad035.

Horta Ribeiro, M., S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West (2021). "Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

Jhaver, S., C. Boylston, D. Yang, and A. Bruckman (2021). "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. Vol. 5, pp. 1–30.

Kemp, S. (2024). *Digital 2024: Brazil*. DataReportal Global Digital Insights report. URL: https://datareportal.com/reports/digital-2024-brazil.

Levy, R. (2021). "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment". In: *American Economic Review* 111.3, pp. 831–870.

Machado, C., B. Kira, V. Narayanan, B. Kollanyi, and P. N. Howard (2019). "A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections". In: *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 1013–1019.

MacKinnon, J. G. and M. D. Webb (2017). "Wild Bootstrap Inference for Wildly Different Cluster Sizes". In: *Journal of Applied Econometrics* 32.2, pp. 233–254.

McCabe, S. D., D. Ferrari, J. Green, D. M. J. Lazer, and K. M. Esterling (2024). "Post-January 6th Deplatforming Reduced the Reach of Misinformation on Twitter". In: *Nature*. DOI: 10.1038/s41586-024-07524-8.

Müller, K. and C. Schwarz (2021). "Fanning the Flames of Hate: Social Media and Hate Crime". In: *Journal of the European Economic Association* 19.4, pp. 2131–2167.

Müller, K. and C. Schwarz (2023). "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". In: *American Economic Journal: Applied Economics* 15.3, pp. 270–312.

Ozawa, J. V. S., J. Lukito, F. Bailez, and L. M. Fakhouri (2024). "Brazilian Capitol Attack: The Interaction between Bolsonaro's Supporters' Content, WhatsApp, Twitter, and News Media". In: *Harvard Kennedy School Misinformation Review* 5.2. DOI: 10.37016/mr-2020-139.

Pangakis, N., S. Wolken, and N. Fasching (2023). "Automated Annotation with Generative AI Requires Validation". arXiv working paper 2306.00176. URL: https://arxiv.org/abs/2306.00176.

Peron, F., P. Pereda, and F. Bailez (2025). "Gender stereotypes in politics: Insights from social media data". In: *Economics Letters* 254, p. 112403. DOI: 10.1016/j.econlet.2025.112403.

Qin, B., D. Strömberg, and Y. Wu (2024). "Social Media and Collective Action in China". In: *Econometrica* 92.6, pp. 1993–2026.

Rambachan, A. and J. Roth (2023). "A More Credible Approach to Parallel Trends". In: *Review of Economic Studies* 90.5, pp. 2555–2591. DOI: 10.1093/restud/rdad018.

Recuero, R., F. B. Soares, and T. Volcan (2023). "Disinformation and Coordinated Behavior on Brazilian Twitter during the 2022 Election". In: *Information, Communication & Society*.

Resende, G., P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto (2019). "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures". In: *Proceedings of the World Wide Web Conference (WWW)*, pp. 818–828.

Tribunal Superior Eleitoral (2024). *Divulgação de Candidaturas e Contas Eleitorais – Eleições 2024*. https://divulgacandcontas.tse.jus.br. Official results of the Brazilian 2024 municipal elections (first round October 6, runoff October 27); 51 municipalities with over 200,000 voters proceeded to a runoff.

Ventura, T., R. Majumdar, J. Nagler, and J. A. Tucker (2025). "Misinformation Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil". In: *Journal of Politics*. Online first.

A Classifier Audit

The content-based outcomes in this paper (hate speech, aggression, potential misinformation patterns) rely on regex dictionary classifiers applied to Portuguese-language WhatsApp messages. Dictionary methods are brittle to context, sarcasm, and polysemy, and their performance must be quantified before results are interpreted. This appendix describes the audit methodology, reports precision and recall on a hand-labeled gold standard, and documents the design choices that produce the classifier specification used in all main results.

A.1 Audit Sample Construction

We drew a stratified random sample from a three-day pre-ban extract (June 11–13, 2024; 701,000 messages). Filtering to the canonical analysis groups, Brazilian country code, and non-empty text fields yielded 34,300 messages. From these we sampled up to 100 positive and 100 negative examples per category (hate strict, hate broad, aggression, potential-misinformation pattern), producing 941 unique messages after deduplication across categories.

One researcher hand-labeled 199 of these messages against the following gold-standard definitions:

- **Hate speech (strict):** severe hostility targeting a protected group (race, gender, sexuality, religion, nationality). Political insults and personal attacks not targeting a protected identity do not qualify.
- **Hate speech (broad):** broader hostile language targeting a group. Includes direct misogynist, transphobic, and homophobic slurs directed at a person. Excludes defensive or condemning statements about slurs.
- **Aggression:** any hostile, insulting, or threatening language, whether political or personal. Includes political epithets used as insults.
- **Potential misinformation patterns:** rhetorical markers characteristic of misinfor-

mation, including urgency cues, conspiracy framing, and conspiratorial labels. Excludes news articles about potential misinformation and casual affirmations.

Health and electoral misinformation categories had too few positive instances in the three-day window (28 and 4, respectively) for meaningful validation and are treated as exploratory throughout the paper.

A.2 Results

Table 9 reports precision, recall, and F1 for each classifier on the 199-message gold-standard sample. The strict hate-speech classifier reaches F1 = 1.00 on $N = 5$ gold positives (wide Wilson CI given small sample). Broad hate speech reaches F1 = 0.85. Aggression reaches F1 = 0.82 and potential-misinformation patterns reach F1 = 0.80. Precision on all four classifiers exceeds 0.86, and recall exceeds 0.71. The key classifier design choices that produce these results are: (i) negative-lookahead patterns that exclude defensive statements about slurs; (ii) exclusion filters that remove idiomatic vulgarity not aimed at a target; (iii) a news-article filter that raises the potential misinformation threshold when the message carries markers of legitimate journalism (URLs to major news domains plus template markers such as "leia mais" or "fonte:"); and (iv) demonstrative- or possessive-context requirements for misogynist slurs, so that isolated keywords do not produce false positives.

Table 9: Classifier performance on 199 hand-labeled messages

Category	Precision	Recall	F1	Gold positives
Hate speech (strict)	1.000	1.000	1.000	5
Hate speech (broad)	0.917	0.786	0.846	14
Aggression	0.863	0.784	0.821	88
Potential misinformation patterns	0.923	0.706	0.800	17

Notes: This table presents classifier performance metrics on 199 hand-labeled messages drawn from a stratified random sample of the pre-ban period (June 11 to 13, 2024). Gold-standard labels were produced by one researcher. Wilson 95% confidence intervals on precision are: hate speech (strict) [48%, 100%], hate speech (broad) [65%, 99%], aggression [77%, 93%], and potential misinformation [67%, 99%]; confidence intervals are wide for rare categories with few flagged observations. Precision is the share of flagged messages that are true positives against the gold label; recall is the share of gold positives that are flagged by the classifier; and F1 is the harmonic mean of precision and recall.

A.3 Ban-Period Revalidation: Potential Misinformation

To address the three principal limitations of the pre-ban audit for the potential-misinformation classifier—single coder, small positive sample, and temporal gap from the treatment period—we conducted a second validation directly on ban-period messages.

Sample. We drew a stratified random sample of 200 messages from the canonical 319-group panel restricted to August 30 to October 8, 2024 (the ban window): 80 messages classified as positive by the dictionary and 120 classified as negative. Messages were selected uniformly at random within each stratum (seed = 42).

Annotation. A second coder (Claude Sonnet 4.5, Anthropic; briefed with the full codebook from Section E and seven gold-standard few-shot examples from the original June audit) independently annotated all 200 messages. The first author then reviewed all 18 cases where the two coders disagreed (9% of the sample) and adjudicated a final label. We note that using an LLM as a second coder differs from the zero-shot LLM classification evaluated in Appendix D: the task here is binary classification against a specific codebook with few-shot examples and a narrow rhetorical target (urgency-and-share directives, censorship narratives), not zero-shot content categorisation. The LLM permissiveness documented elsewhere—where positive rates are 4–50× higher than the dictionary—arises from the absence of a codebook constraint and from labelling broad content types rather than specific rhetorical markers; the codebook-constrained few-shot protocol substantially reduces that permissiveness, as confirmed by the low false-positive rate (18.8%) in the current exercise.

Results. The two-coder agreement rate is 92.0% and Cohen's $\kappa = 0.828$, indicating substantial to almost-perfect agreement (Landis and Koch, 1977). Against the human-adjudicated ground truth, the dictionary achieves precision of 0.812 (Wilson 95% CI: [0.71, 0.88]), recall of 0.985, and F1 of 0.89. These results confirm that the classifier maintains its

performance during the ban period despite the different rhetorical context of the municipal election campaign. The dominant source of false positives (15 of 80 flagged messages) is news articles from partisan outlets accompanied by WhatsApp group-recruitment links (*"Compartilhe — Grupo do WhatsApp"*), which trigger the urgency-and-share pattern; all are factual news shares rather than misinformation templates. The single remaining false negative is a conspiracy narrative about deliberately set fires written in non-standard vocabulary absent from the dictionary.

Table 10: Ban-period revalidation: potential misinformation classifier

Metric	Value
Sample period	Aug 30 – Oct 8, 2024
Sample size	200 (80 dict+, 120 dict−)
Coders	2 (Claude + first author adjudication)
Cohen's κ	0.828
Agreement rate	92.0%
Precision (vs. ground truth)	0.812
Wilson 95% CI on precision	[0.71, 0.88]
Recall	0.985
F1	0.890
False positives (FP)	15/80 flagged
False negatives (FN)	1/120 unflagged

Notes: Ground truth labels produced by two coders (Claude annotated all 200 messages using the codebook in Appendix E with seven few-shot examples; the first author adjudicated all 18 disagreements). The dominant false-positive pattern is news-article shares with group-recruitment links triggering the urgency-and-share rule. Wilson CI uses the exact (Clopper-Pearson) method.

A.4 Limitations of the Audit

Three limitations remain. The ban-period revalidation covers only the potential-misinformation classifier; the aggression, hate speech (strict), and hate speech (broad) classifiers retain single-coder June labels.

The gold-standard sample is small for hate speech (strict, $N = 5$) and hate speech (broad, $N = 14$), so Wilson 95% confidence intervals on precision remain wide for those categories. Dictionary classifiers face a performance ceiling of roughly 60–80% F1 for hate speech in

Portuguese; fine-tuned transformer models such as BERTimbau trained on HateBR could likely reach 85–90% F1 but are deferred to future work to preserve reproducibility in the current paper.

B Fact-Check Matching Procedure

This appendix documents the procedure that produces the *Fake news* row of Table 2 and Table 4. The outcome is a per-group-day count of WhatsApp messages whose content matches a claim that a Brazilian fact-checking agency later verified as false, measured *before* the debunking publication date. The procedure has three stages: claim ingestion, message–claim matching, and aggregation to the group-day panel. Our design draws on the Poynter International Fact-Checking Network (IFCN) certification standard and follows Ozawa et al. (2024), who use a similar keyword-based matching architecture to link WhatsApp messages from the Palver panel to Brazilian fact-check claims.

Claim corpus. We scrape all 2024 articles from four IFCN-certified Brazilian fact-checking agencies: *Aos Fatos* (2,985 URLs), *Boatos.org* (1,121), *Agência Lupa* (390), and *e-Farsas* (213), yielding 4,709 articles. For each agency we parse the public sitemap (schema `urlset`), restrict to entries with `lastmod` in 2024, and dereference each URL to extract the `ClaimReview` and `NewsArticle` JSON-LD blocks. From these we retain the headline, the `claimReviewed` string, the verdict, the `datePublished` timestamp, and a body snippet. *Aos Fatos*, *Boatos.org*, and *e-Farsas* publish schema.org `ClaimReview` payloads that expose the verdict directly in `reviewRating.alternateName`. *Agência Lupa* does not emit a `ClaimReview` block, but its `NewsArticle` JSON-LD surfaces the verdict as one of the entries in the `keywords` array (e.g. `"Falso"`, `"Enganoso"`, `"Verdadeiro"`); we match each Lupa article against the same verdict lexicon used for the other agencies. After verdict normalisation (mapping local terms such as `falso`, `enganoso`, `boato` to the `false` bucket and discarding `verdadeiro`, `parcialmente verdadeiro`), 3,796 of the 4,709 articles carry a

`false` verdict and a parseable publish date. This is the claim corpus used for matching; all four agencies contribute.

Matching algorithm. The matcher scans the 11.8 million messages in the 319-group analytical sample against the 3,796 false-verdict claims. Text normalisation is identical on both sides: Unicode NFKD decomposition with diacritic stripping, lowercasing, URL removal, a token filter of minimum length 4, and a stop-list that combines Portuguese function words with fact-checking vocabulary (`falso`, `verdade`, `checamos`, `desmentido`, etc.). For each claim we extract three keyword sets from the original-case text:

1. **Proper nouns** — tokens matching `\b[A-ZÀ-Ü][A-Za-z0-9À-ü.\-]{2,}\b` in the raw headline (following Brazilian Portuguese capitalisation conventions for names, places, and institutions).
2. **Numbers with units** — currency (`R$`), percentages, and quantities with suffixes `mil`, `milhões`, `bi`.
3. **Content tokens** — all normalised non-stopword tokens of length ≥ 4 from the headline and `description`.

An inverted index maps each keyword to the set of claims containing it. Tokens that appear in more than $\max(3, 5\%)$ of claims are pruned from the index to remove non-discriminative vocabulary (e.g. *Brasil*). For each message we look up its tokens against the index, collect candidate claims, and score each (message, claim) pair as

$$\text{score} = 3 \cdot \text{proper\_hits} + \text{content\_hits} + 4 \cdot \text{number\_hits},$$

where the hit counts are set intersections between the message's tokens/proper nouns and the claim's. A candidate qualifies as a *match* if (i) proper_hits $\geq K$, and (ii) score $\geq S$. The paper uses $K = 3$ and $S = 7$; we report sensitivity to K in Table 11.

Category assignment. A qualifying match is then classified by the message's temporal position relative to the fact-check and by whether the message itself contains debunking markers. If the message contains a fact-check-agency URL (`aosfatos.org`, `lupa.uol.com.br`, `boatos.org`, `e-farsas`, `g1.globo.com/fato-ou-fake`, etc.), it is labelled `fc_url`. Otherwise, if the message contains any of twenty-one debunking phrases (*é falso que*, *não é verdade que*, *desmentido*, *checamos*, etc.), it is labelled `fc_share`. Otherwise, if the message date strictly precedes the claim's `datePublished`, it is labelled `original_claim`; this is the category that feeds the *Fake news* row. Remaining matches where the message date is on or after the claim date without debunking markers are labelled `post_date_reshare`. When a message matches multiple claims, we assign a single label using the priority `fc_url` $\succ$ `fc_share` $\succ$ `original_claim` $\succ$ `post_date_reshare`, which is conservative: a message containing a fact-check URL or a debunking phrase is classified as a debunking share even if it also triggers a (spurious) `original_claim` match against another claim.

Outcome construction. For each 319-group $\times$ day cell we count the number of messages labelled `original_claim`; this count enters the *Fake news* row in Table 2 and Table 4. Zero-count cells are assigned 0 (no imputation). Of the 319 analytical groups, all receive a non-zero count at some point in the 2024 panel.

Threshold sensitivity. Table 11 reports the number of unique messages classified into each category as a function of the proper-noun threshold K, holding the composite-score floor fixed at $S = 7$. Tightening K from 2 to 3 reduces the pool of matched messages by roughly a factor of four and materially improves precision (see audit below).

Precision audit. We draw a seeded random sample of 100 messages from the $K = 3$ `original_claim` pool (stratified by agency and claim-publication month) and produce a hand-audit CSV at `quality_reports/factcheck_fix/factcheck_audit_k3_sample.csv`. Two machine-auditable similarity proxies — BM25 scored over the false-claims corpus (pa-

Table 11: Fact-check matching: threshold sensitivity.

Threshold	Original-claim msgs	Any-match msgs
$K = 2$	487,195	858,128
$K = 3$ (paper)	60,396	130,993

Notes: Unique-message counts after collapsing the (message, claim) match rows to a per-message label with the priority rule described in the text. The composite-score floor is $S = 7$ throughout; all four IFCN-certified agencies contribute (*Aos Fatos*, *Boatos.org*, *e-Farsas*, and *Agência Lupa*). *Agência Lupa* verdicts are extracted from the JSON-LD `keywords` array (Lupa does not emit a `ClaimReview` block).

rameters $k_1 = 1.5$, $b = 0.75$) and word-token Jaccard on content tokens — are computed for each sampled pair. Against a null benchmark of 200 random (message, claim) pairs drawn from the same corpus, the $K = 3$ sampled matches exhibit a dramatically higher mean BM25 (25.65 vs. 0.61), and 100 of 100 sampled pairs clear the BM25 ≥ 3.0 high-signal threshold (versus only 4 of 100 at $K = 2$). BM25 is an *upper bound* on precision when used as a filter, since semantic matches with low lexical overlap will not be recovered; ground-truth precision requires human labels, for which the CSV carries empty `is_true_match` and `reviewer_notes` columns.

Caveats and interpretation. Two caveats bear on the *Fake news* row. First, the matcher's precision rests on lexical co-occurrence of proper nouns and content tokens; a claim paraphrased without reusing proper nouns will not be recovered. The role of the *Fake news* row in the paper is therefore to *corroborate* the direction of the effect identified by the dictionary-based misinformation classifier (Section 3.3), not to provide a precision-certified point estimate of fact-checked misinformation on its own. Second, the claim corpus is restricted to 2024 articles from four Brazilian agencies and inherits their editorial priors (Brazilian political news, health misinformation, viral-media claims); misinformation about local elections or niche topics that these agencies do not cover is invisible to the matcher. A full human-labelled precision study, using the audit CSV scaffolding, is left for a companion note; this paper uses the measure as one of two complementary content signals and relies on

their agreement.

C Platform Categorization

Table 12 lists all platform categories and their constituent hosts used in the URL extraction and platform substitution analysis (Section 5.2). The categorization was pre-registered before running the substitution analysis. Any host not appearing in this list is classified as "other."

Table 12: Platform Categories and Constituent Hosts

Category	Hosts
Major social platforms	
twitter_x	twitter.com, x.com, t.co, mobile.twitter.com
instagram	instagram.com
facebook	facebook.com, fb.com, fb.watch, fb.me, m.facebook.com
youtube	youtube.com, youtu.be, m.youtube.com
tiktok	tiktok.com, vm.tiktok.com
telegram	t.me, telegram.me
whatsapp	chat.whatsapp.com, wa.me
kwai	kwai.com, kwai-video.com
Twitter/X alternatives	
bluesky	bsky.app, bsky.social
threads	threads.net
mastodon	mastodon.social, mastodon.online, mstdn.social, and other instances
parler	parler.com
gab	gab.com
gettr	gettr.com
truthsocial	truthsocial.com
koo	kooapp.com, koo.com
minds	minds.com
post_news	post.news
spoutible	spoutible.com
nostr	nostr.com
Brazilian media	
mainstream_news	g1.globo.com, uol.com.br, folha.uol.com.br, estadao.com.br, r7.com, cnnbrasil.com.br, bbc.com, and others
rightwing_br	oantagonista.com.br, gazetabrasil.com.br, jovempan.com.br, pensandodireita.com, tercelivre.com.br, and others
leftwing_br	brasil247.com, cartacapital.com.br, revistaforum.com.br, theintercept.com, and others
factcheck	aosfatos.org, lupa.uol.com.br, e-farsas.com, and others
Alternative video platforms	
rumble	rumble.com
odysee	odysee.com
bitchute	bitchute.com
vimeo	vimeo.com
twitch	twitch.tv
dailymotion	dailymotion.com
Other platforms	
discord	discord.com, discord.gg
signal	signal.org
snapchat	snapchat.com
linkedin	linkedin.com
pinterest	pinterest.com

Notes: This table presents the pre-registered platform categorisation used throughout the paper. The classification was fixed before running the platform substitution analysis. "And others" denotes additional minor domains within the same editorial group. Any URL host not listed here is classified as "other."

D LLM Validation: Why We Do Not Use LLM Classification as a Content Measure

This appendix documents two pieces of evidence that led us to exclude LLM-classified outcomes from our content analysis: a side-by-side validation against the dictionary classifier (severe permissiveness) and a weekly event-study diagnostic (non-flat pre-trends).

Permissiveness against the dictionary classifier. Table 13 compares per-category positive rates between GPT-4o-mini binary classification (zero-shot, temperature 0) and our dictionary classifier (Section 3.3, Appendix A) on the 844,413-message overlap window (August 16 to September 12, 2024) for which both classifiers have labels and the dictionary has been hand-validated against a 199-message gold standard with F1 ranging from 0.80 to 1.00 across categories. The LLM produces 34× more positives than the dictionary for hate speech (strict), 50× more for hate speech (broad), 4× more for aggression, and 34× more for potential-misinformation patterns. Cohen's κ is 0.025 (hate broad), 0.034 (hate strict), 0.035 (potential misinformation), and 0.286 (aggression), indicating agreement only marginally above chance for the hate and misinformation categories. Treating dictionary positives as the comparison standard, the LLM achieves recall of 60–83% (it catches most of what the dictionary catches) but precision of only 1–19% across categories: 81–99% of LLM-flagged messages are not corroborated by the validated dictionary. Because the dictionary is the validated classifier, this pattern indicates that the LLM is severely permissive — positive rates are 4–50 times the validated base rate — and that its noise floor swamps any signal at the per-day count level used by the DiD specification.

Table 13: LLM versus dictionary classifier on the 844,413-message overlap window

	Positive rate			Dictionary as gold		
Category	LLM	Dictionary	LLM/Dict	Precision	Recall	Cohen's κ
Hate speech (strict)	0.54%	0.016%	33.95×	0.018	0.600	0.034
Hate speech (broad)	1.69%	0.034%	49.90×	0.013	0.650	0.025
Aggression	8.46%	1.923%	4.40×	0.189	0.832	0.286
Potential misinformation	5.69%	0.167%	33.99×	0.020	0.675	0.035

Notes: Per-category comparison of GPT-4o-mini binary classification (zero-shot, temperature 0) against the project's dictionary classifier (Section 3.3) on the 844,413 messages in the LLM-classified subsample for which both labels exist (August 16 to September 12, 2024). The "LLM" column reports the share of messages flagged as positive by the LLM; the "Dictionary" column the same for the dictionary classifier. "LLM/Dict" is the ratio of the two positive rates. Precision, Recall, and Cohen's κ treat the dictionary classifier as the comparison standard for purposes of arithmetic; this does not assume the dictionary is the truth, but the dictionary is the only one of the two classifiers with hand-coded validation against a 199-message gold standard from the same corpus (F1 = 0.80–1.00 across categories, Appendix A). The pattern — LLM positive rates 4–50× the dictionary, κ values close to zero for the hate and misinformation categories — indicates the LLM is severely permissive in this corpus.

Substantive limits. Several features of the WhatsApp Brazilian Portuguese corpus make zero-shot LLM classification unreliable in ways that prompt-engineering alone cannot eliminate. *Out-of-distribution language*: GPT-4o-mini is trained predominantly on web English; our corpus is Brazilian Portuguese with regional slang, internet abbreviations specific to Brazil (*kkkk*, *mlk*, *blz*), and political memes whose meaning depends on local context. *Defensive versus propagating use*: a message such as *"é falso que Lula vendeu o Brasil"* can be either a fact-checking denial or a setup for the same false claim; our dictionary handles the distinction with negative-lookahead patterns (Section 3.3), zero-shot LLMs do not. *Base-rate miscalibration*: the LLM expects population prevalences in the 5–10% range; the validated prevalences in WhatsApp Brazil for the relevant categories are 0.02–2%, so the LLM posterior is mis-anchored. These limits are documented in the broader literature on automated annotation (Gilardi, Alizadeh, and Kubli, 2023; Pangakis, Wolken, and Fasching, 2023); in our setting their joint effect is the order-of-magnitude permissiveness reported in Table 13.

Pre-trend diagnostic at weekly aggregation. Even setting the permissiveness problem aside, the LLM-classified outcomes also fail the parallel-trends assumption in a static DiD.

Figure 8 reports a weekly event study on the canonical 319-group sample over the 48-day classification window. The X-link panel (Panel A) is a sanity check: it reproduces the headline result at the weekly level, with a clean monotone post-ban drop and no apparent pre-trend at $k_w = -1$. The two LLM panels tell a different story. Pre-treatment coefficients at $k_w = -4$ and $k_w = -3$ are large and statistically significant in the same direction as the post-ban estimates. The reference week $k_w = -1$ (August 19–25) is itself an anomalous peak in political and institutional rhetoric driven by the public escalation of the STF–X conflict in the days preceding the ban. The static DiD on these outcomes therefore mixes a peak-versus-trough calendar contrast with the genuine ban onset.

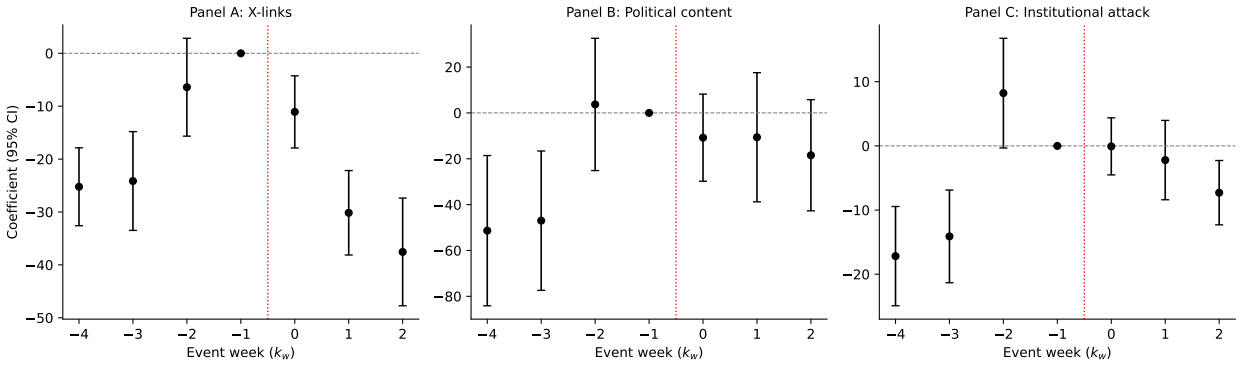


Figure 8: Weekly event study on LLM-classified outcomes (canonical 319-group sample, 48-day window)

Notes: Weekly event-study coefficients μ_{k_w} from $Y_{gw} = \alpha_g + \delta_w + \sum_{k_w \neq -1} \mu_{k_w}(\text{treated}_g \times \mathbb{1}[w - w_{\text{ban}} = k_w]) + \varepsilon_{gw}$ on the canonical 319-group sample over the 48-day LLM classification window. The reference week is $k_w = -1$ (the calendar week of August 19–25, 2024, immediately preceding ban onset on August 30). The dashed red vertical line marks ban onset (between $k_w = -1$ and $k_w = 0$). Whiskers are 95% confidence intervals from cluster-robust standard errors at the group level. Panel A shows the X-link outcome as a sanity check: a clean monotone post-ban drop. Panels B (Political content) and C (Institutional attack) show large, statistically significant pre-treatment coefficients at $k_w = -4$ and $k_w = -3$, indicating that the parallel-trends assumption does not hold over this window. The reference week $k_w = -1$ is itself an anomalous peak in political content driven by the public escalation of the STF–X conflict, which makes static DiD coefficients on these outcomes sensitive to the pre-window choice. Outcomes are summed within group-week.

Conclusion. The combination of severe permissiveness against a hand-validated dictionary classifier (Table 13) and non-flat pre-trends in a weekly event study (Figure 8) led us to exclude LLM-classified outcomes from our content analysis. Our content measures are the dictionary-based *Potential misinformation* row in Table 2 (full-year panel, flat pre-trends, hand-validated F1 = 0.80) and the externally anchored *Fake news* row in Table 2 and Table 4

(Appendix B). A longer LLM classification window with a quiet pre-period and case-specific calibration via prompt engineering, few-shot examples, or a labeled training set would be needed to make LLM classification a useful identification target on this corpus; we leave that to future work.

E Classifier Categories

This appendix describes the content categories used in each classifier at the level required to interpret the results. The full regular-expression patterns, exhaustive word and expression lists, and Portuguese-language domain hosts used by each classifier are available from the authors upon request. We intentionally do not reproduce the complete list of slurs, insults, and misinformation templates in the published appendix to avoid amplifying identity-targeted hostility in a publication intended for general academic circulation.

Hate speech (strict). The strict hate-speech classifier flags messages that contain severe identity-targeted hostility against a protected group, including racial slurs in direct-insult context, severe transphobic and homophobic insults, xenophobic slurs, group-targeted death threats, and anti-LGBT rhetoric that assigns pathologising or violent predicates to gender or sexual identity. Defensive statements that mention slurs in a condemning frame are excluded via negative-lookahead patterns, and idiomatic vulgarities that do not target an identity group are excluded via a separate filter.

Hate speech (broad). The broad classifier nests the strict classifier and additionally flags direct misogynist, transphobic, and homophobic slurs used in demonstrative or possessive context (i.e., directed at a specific person through determiners such as *seu/sua/esse/essa* or subject pronouns), slur-plus-hostile-modifier combinations, body-focused objectifications paired with slurs, misogynist prescriptive statements, and mental-health slurs used as group insults. The same exclusion set applies.

Aggression. The aggression classifier captures hostile, insulting, or threatening language in-

dependent of identity targeting. It covers core insults, rude imperatives, coarse sexual insults, political epithets used as insults, collective epithets applied to political groups, organised-crime group labels applied to political groups, violent rhetorical threats, and mental-health slurs used as personal insults. Idiomatic vulgarities trigger exclusion only in short messages (< 30 characters), on the logic that a long message containing vulgarity is still aggressive in character.

Potential misinformation patterns. The potential misinformation classifier captures rhetorical templates of misinformation rather than verifying the factual accuracy of specific claims. The templates include urgency-and-share directives, censorship narratives ("the media won't show you this"), appeal-to-authority templates ("leaked medical source"), miracle-cure claims, named conspiracy labels (including vaccine-chip, 5G-control, new-world-order, and deep-state tropes), wake-up rhetoric, grand-media-distrust templates, emoji-led share directives, and health-specific conspiracies common to Brazilian WhatsApp. Exclusion patterns cover news-article templates (where a URL to a major Brazilian news domain co-occurs with markers such as "leia mais" or "fonte:") and casual affirmations ("kkk," "rsrs"). When a message matches a news-article template and the potential misinformation match is weak, the classifier requires an additional strong conspiracy keyword before flagging.

Fact-check classifiers. Fact-check links are identified by URL host; the main patterns include *aosfatos.org*, *lupa.uol.com.br*, and *e-farsas.com*, with the full list available upon request. Left-wing, right-wing, and mainstream Brazilian media domains are identified against a pre-registered list described in Appendix C.

F Sender Heterogeneity Details

This appendix provides full details on the sender heterogeneity analysis summarized in Section 5.5. We document bucket construction, report effective cluster counts, present five robustness checks (within-treatment bucketing, dominance cutoff sensitivity, continuous dose-

response, log-bin specifications, and a selection check on dropped senders), and show a heavy-tail zoom on the top buckets.

F.1 Heavy-tail zoom

Figure 9 complements the quartile panel (Figure 3 in the main body) by zooming in on the four buckets with the most active senders. It is primarily diagnostic: the effective sample size declines sharply with the bucket percentile, so the heavy-tail estimates carry larger standard errors even when the points remain consistent with the quartile-level pattern.

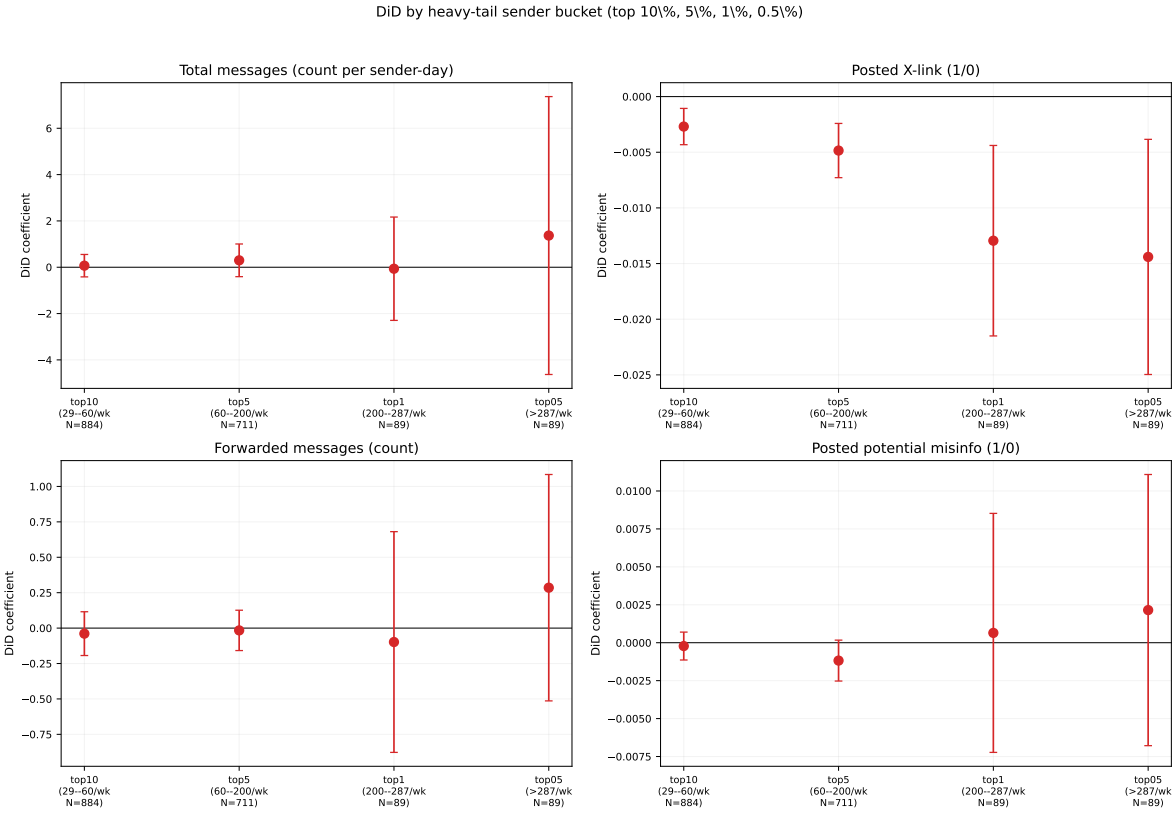


Figure 9: Ban effect by heavy-tail sender bucket (top 10%, 5%, 1%, 0.1%)

Notes: This figure presents the same four-outcome sender-level DiD coefficients as Figure 3 but on a heavy-tail zoom: the top 10%, top 5%, top 1%, and top 0.1% of senders by total messages sent during the clean pre-ban window (January 1 to May 31, 2024). The top 0.1% bucket contains 18 senders and is reported with the caveat that inference at the narrowest tail is limited by its small sample size. Bucket labels show the messages-per-week range and the number of distinct senders. Whiskers are 95% confidence intervals from group-clustered standard errors.

F.2 Bucket Definitions

We assign each sender to one of eight activity buckets based on total messages sent during the clean pre-ban window (January 1–May 31, 2024). This window precedes the June–August treatment-definition period, ensuring that bucket assignment is not mechanically related to the treatment variable. After applying the daily canonical sample (Section 3.2, 319 groups) and dropping senders with no activity in the Jan–May window, the bucketing population is $N = 17{,}777$. The thresholds correspond to percentiles of this population: 25th (= 27 messages), 50th (= 62), 75th (= 188), 90th (= 624), 95th (= 1,306), 99th (= 4,336), and 99.9th (= 13,926). The figures differ from the legacy weekly-companion specification ($N = 21{,}863$, lower thresholds) because the canonical sample has shifted from 325 to 319 groups and the bucketing now uses sender-day totals; conclusions are unaffected. Senders are assigned to their "dominant" group (the group where they sent the largest share of their messages) at the 70% dominance threshold. Section F.8 examines whether this exclusion introduces bias.

F.3 Effective Cluster Counts

Table 14 reports the number of senders, groups (clusters), and treated/control groups per bucket on the daily sender–group–day sample (319 groups, 17,777 senders). Cluster counts decline from 257 (Q1) to 66 (top 0.1%), reflecting the concentration of heavy senders across fewer groups. The top 0.1% bucket contains 18 senders spread across 66 groups (44 treated, 22 control), providing limited variation for reliable inference at the tail.

Table 14: Effective Cluster Counts by Sender Activity Bucket

Bucket	N senders	N groups	Treated groups	Control groups
Q1 (0–27 msgs)	4,580	257	129	128
Q2 (27–62)	4,369	250	124	126
Q3 (62–188)	4,388	267	138	129
Q4 (188–624)	2,667	289	147	142
Top 10% (624–1306)	884	247	131	116
Top 5% (1306–4335)	711	236	137	99
Top 1% (4335–13926)	160	190	109	81
Top 0.1% ($> 13,926$)	18	66	44	22

Notes: This table presents the number of senders, the number of distinct groups to which they belong (clusters for standard-error calculation), and the split of those groups between treated and control, by sender activity bucket, on the daily sender-group-day sample used in Figure 3 and Figure 9 (N = 17,777 senders across 319 groups). Bucket thresholds are percentiles of total messages sent in January to May 2024 among senders active in that window (25th, 50th, 75th, 90th, 95th, 99th, and 99.9th). Treated is defined as above-median pre-ban *x_per_week* in the group-level classification.

F.4 Within-Treatment Bucketing

To verify that the sender-level effects are not driven by treated-control compositional differences, we restrict the analysis to treated groups only and re-estimate bucket-specific ban-period coefficients on X-links. Table 15 reports the results: the negative coefficient grows monotonically in sender activity, from -0.000 in Q1 to -0.192 in the top 0.1%. The top 5% coefficient is -0.038 ($p < 0.001$) and the top 1% is -0.047 ($p < 0.001$), confirming that the effect concentrates among heavy senders even within treated groups.

Table 15: Within-Treatment Bucketing: X-Link Coefficients by Sender Bucket

Bucket	Coefficient	SE	p-value
Q1	-0.000^{**}	(0.000)	0.012
Q2	-0.002^{***}	(0.000)	< 0.001
Q3	-0.003^{***}	(0.000)	< 0.001
Q4	-0.012^{***}	(0.002)	< 0.001
Top 10%	-0.011^{***}	(0.002)	< 0.001
Top 5%	-0.038^{***}	(0.005)	< 0.001
Top 1%	-0.047^{***}	(0.011)	< 0.001
Top 0.1%	-0.192^{**}	(0.093)	0.046

Notes: This table presents ban-period coefficients on X/Twitter links by sender activity bucket using only the treated groups (above-median pre-ban *x_per_week*). Each row is a separate regression at the sender-group-day level on the bucket's sender-day observations. Specification includes sender-by-group fixed effects with a ban-period indicator; calendar-day fixed effects are dropped because they would absorb the ban indicator once treated-versus-control variation is removed. Standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

F.5 Dominance Cutoff Sensitivity

Our baseline assigns each sender to the group where they send at least 70% of their messages. We test sensitivity to this threshold by re-running the analysis at 60%, 80%, and 90% cutoffs. The top 5% and top 1% X-link coefficients remain significant at all four cutoffs, though the sample size shrinks at higher thresholds as more multi-group senders are excluded. The qualitative pattern of concentration of effects in the top 5% is stable across all specifications.

F.6 Continuous Dose-Response

We replace the discrete bucket indicators with a continuous dose variable (the standardised log of pre-ban total messages) interacted with the ban indicator, and re-estimate the DiD inside each bucket. Table 16 reports the results. The dose-response coefficient increases sharply in the upper tail, from -0.0002 in Q1 to -1.5262 in the top 0.1%. The monotonic

increase in dose-response magnitude across the upper tail confirms that the treatment effect rises smoothly with sender activity level rather than being an artifact of discrete bucket boundaries.

Table 16: Continuous Dose-Response: X-Link Coefficients by Sender Bucket

Bucket	Dose coefficient	SE	p-value
Q1	-0.0002	(0.0003)	0.594
Q2	-0.0023^{**}	(0.0010)	0.026
Q3	-0.0014	(0.0014)	0.329
Q4	-0.0099	(0.0071)	0.165
Top 10%	-0.0182^{*}	(0.0096)	0.059
Top 5%	-0.0768^{***}	(0.0167)	< 0.001
Top 1%	-0.0059	(0.0308)	0.848
Top 0.1%	-1.5262^{**}	(0.7152)	0.037

Notes: This table presents continuous dose-response coefficients on X/Twitter links by sender activity bucket. Dose is defined as the standardised log(1 + pre-ban total messages) interacted with the ban-period indicator, and is estimated at the sender-group-day level within each bucket. Each row is a separate regression on the bucket's sender-day observations. Specification includes sender-by-group and calendar-day fixed effects, with standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

F.7 Log-Bin Bucket Results

As an alternative to percentile-based buckets, we define bins on a logarithmic scale of total pre-ban messages. Table 17 reports the results. The effect onset occurs between the 300–1,000 and 1,000–3,000 ranges, and the strongest negative coefficients are in the 3,000–10,000 bin (-0.061, $p < 0.001$) and the 10,000–100,000 bin (-0.085, $p < 0.10$; $N = 34$, underpowered). This analysis confirms that the critical mass of X-dependent senders lies in the 300–10,000 message range, corresponding roughly to the top 5%–1% of the activity distribution.

Table 17: Log-Bin Bucket Results: X-Link Coefficients

Bin (msgs in Jan–May)	N senders	Coefficient	SE	p-value
0–10	1,125	+0.001	(0.001)	0.239
10–30	4,076	−0.000	(0.000)	0.842
30–100	5,842	−0.001*	(0.000)	0.061
100–300	3,521	−0.002*	(0.001)	0.072
300–1,000	2,047	−0.006**	(0.002)	0.013
1,000–3,000	852	−0.019***	(0.004)	< 0.001
3,000–10,000	280	−0.061***	(0.012)	< 0.001
10,000–100,000	34	−0.085*	(0.048)	0.081

Notes: This table presents ban-period DiD coefficients on X/Twitter links by log-scale message bin. Bins are defined on total messages sent in January to May 2024. Each row is a separate regression at the sender-group-day level on the bin's sender-day observations, with sender-by-group and calendar-day fixed effects and the ban and post indicators interacted with treated status. Standard errors clustered at the group level. The 10,000 to 100,000 bin is underpowered ($N = 34$). Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

F.8 Selection Check

Senders excluded from the sender-level analysis (those failing the 70% dominance threshold or the balanced-panel filter) differ systematically from retained senders. Dropped senders have 3–4 times higher mean pre-ban message counts than retained senders (Table 18). Dropped control-group senders are heavier than dropped treated-group senders (1,081 vs. 708 mean pre-ban messages). This asymmetry means that the dominance filter preferentially removes heavy senders from the control group, which, if anything, *attenuates* the treated-control contrast. The DiD estimates are therefore conservative: including these senders would strengthen the estimated treatment effect.

Table 18: Selection Check: Retained vs. Dropped Senders

Status	Treatment group	N senders	Mean pre-ban msgs
Retained	Control	10,524	287.6
Retained	Treated	7,867	292.3
Dropped	Control	2,411	1,080.7
Dropped	Treated	1,061	707.5

Notes: This table presents the distribution of senders between those retained in the sender-level analysis and those dropped by the 70% dominance threshold or the balanced-panel filter, with pre-ban message counts computed over January to May 2024. Dropped senders are systematically heavier than retained senders and disproportionately concentrated in the control group, which attenuates the treated-control contrast and makes the DiD estimates conservative.

F.9 Why Within-Group Sender-Level Identification Fails

The sender heterogeneity analysis uses between-group variation in pre-ban X-link intensity; a natural alternative would compare, within each group, senders who used X heavily against those who used it rarely. We pursued this design and encountered two obstacles. First, pre-trends fail: 7 of 12 pre-ban week indicators are significantly different from zero on the *posted_x_link* outcome under a sender-by-group / group-by-week FE specification, because X-link users are more politically responsive than non-users within the same group and the group-by-week FE cannot absorb that heterogeneity. Second, within-group variation is thin: only 103 of the 319 canonical groups (32%) contain two or more X-dependent senders (at least one X-link per week in the pre-ban window), and the median such group contains two. The group-level design in the main analysis is the appropriate unit for this identification problem.

G Additional Robustness and Pre-Trend Diagnostics

This appendix collects additional robustness material referenced in the main text.

G.1 Treated vs. Control Balance and Propensity-Score Diagnostics

Table 19 compares pre-ban characteristics of treated and control groups and documents how the imbalance evolves after propensity-score matching (PSM) and inverse probability weighting (IPW). Columns (1)–(2) report group means; column (3) reports the raw difference; column (5) reports the raw standardized difference. Columns (6)–(8) report standardized differences after PSM (1:1 nearest-neighbor matching, no replacement, $N = 246$) and after IPW (stabilized ATT weights, $N = 319$). By design, the groups differ on X-links per week—the treatment variable—and this gap is unchanged by matching or weighting. On group-composition characteristics (share political, share right-wing) the two sets are balanced before and after. On activity-level variables (messages, forwarded messages, distinct senders) treated groups are larger and more active; IPW largely restores balance on these dimensions, while PSM does so less because the scalar propensity score does not guarantee balance on individual covariates. The group fixed effects in the main specification absorb these time-invariant level differences regardless.

Figure 10 presents the same information as a love plot. The dashed lines mark the $|\text{std diff}| = 0.10$ threshold (conventional "well-balanced" criterion).

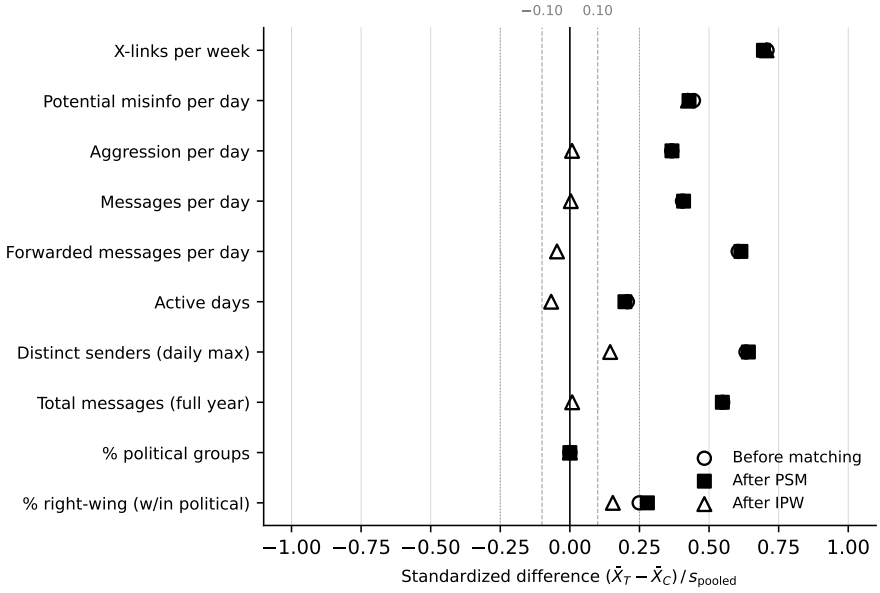
Table 20 repeats the main DiD estimates under PSM and IPW reweighting. Coefficients are broadly stable across columns, confirming that the baseline results are not sensitive to the pre-existing activity differences between treated and control groups.

Table 19: Pre-ban balance: treated vs. control groups

	Mean				Std. difference		
Variable	Treated	Control	Difference	p-value	Raw	PSM	IPW
X-links per week [a]	16.747	2.470	+14.277	0.000***	+1.211	+0.695	+0.706
Messages per day	145.7	81.9	+63.8	0.001***	+0.371	+0.408	+0.003
Forwarded messages per day	28.6	14.5	+14.1	0.000***	+0.466	+0.614	-0.046
Active days	62.5	59.5	+3.0	0.176	+0.152	+0.198	-0.067
Distinct senders (daily max)	57.7	36.8	+20.9	0.000***	+0.424	+0.641	+0.145
Total messages (full year)	47,224	27,036	+20,189	0.001***	+0.373	+0.548	+0.008
Share political groups	0.755	0.769	-0.014	0.770	-0.033	+0.000	+0.000
Share right-wing (within pol.)	0.617	0.463	+0.153	0.016**	+0.310	+0.279	+0.154
Potential misinfo per day	0.473	0.107	+0.367	0.000***	+0.762	+0.428	+0.424
Aggression per day	3.352	1.247	+2.105	0.000***	+0.499	+0.367	+0.008
N	159	160				123[b]	159

Notes: Pre-ban period: June 1–August 29, 2024. Treated groups have above-median X-links per week (> 5.03). Differences are treated minus control. p-values from two-sample Welch t-tests. Standardized differences $= (\bar{X}_T - \bar{X}_C)/s_{\text{pooled}}$. [a] Treatment variable; differs by design. [b] PSM column uses 1:1 nearest-neighbor matched sample ($N = 246$, 123 per arm). IPW uses stabilized ATT weights (full $N = 319$). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Figure 10: Covariate balance before and after propensity-score matching and weighting



Notes: Open circles = raw standardized differences; filled squares = after PSM; open triangles = after IPW. Vertical dashed lines mark |std diff| = 0.10. X-links per week is the treatment variable.

Table 20: DiD estimates: full sample, PSM-matched, and IPW-weighted

	(1) Full sample	(2) PSM matched	(3) IPW weighted
X/Twitter links per day			
Ban period	-2.799***	-2.797***	-3.036***
	(0.348)	(0.348)	(0.480)
Forwarded messages per day			
Ban period	-13.078***	-13.091***	-10.615***
	(3.437)	(3.442)	(3.452)
Total messages per day			
Ban period	-41.488*	-41.502*	-27.991
	(23.059)	(23.090)	(21.190)
Potential misinfo per day			
Ban period	-0.144***	-0.144***	-0.121***
	(0.047)	(0.047)	(0.045)
Aggression per day			
Ban period	-1.906**	-1.906**	-1.472**
	(0.792)	(0.793)	(0.653)
Fact-check links per day			
Ban period	-0.033**	-0.033**	-0.030**
	(0.014)	(0.014)	(0.013)
Observations	72,672	72,536	72,672
Groups	319	246	319
Group FE	✓	✓	✓
Day FE	✓	✓	✓
Cluster SE (group)	✓	✓	✓
Analytic weights			✓

Notes: Dependent variables in levels. Each panel reports the ban-period DiD coefficient $\hat{\beta}_{\text{ban}}$ from equation (1). Column (1): full estimation sample ($N = 319$ groups). Column (2): 1:1 nearest-neighbor PSM on propensity score, no replacement ($N = 246$ groups). Column (3): IPW-weighted DiD using stabilized ATT weights ($N = 319$). Standard errors clustered at the group level in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

G.2 Other Outcomes

Table 21 reports daily DiD estimates for the secondary outcomes that did not appear in the main results table (Table 2). Panel A covers granular activity measures (messages with links, videos, images, audio messages, mainstream news links). Panel B covers secondary

content outcomes (strict hate speech, health misinformation, electoral misinformation). The specification is identical to the main results.

Table 21: Other outcomes: granular activity and secondary content

Outcome	Pre-mean (treated)	Ban effect Coef.	SE	$\%\Delta_{\text{ban}}$ (% of pre-mean)
Panel A: Other activity outcomes				
Messages with links	56.59	-7.71^{**}	(3.65)	-13.6%
Videos	44.60	-13.22^{***}	(3.18)	-29.6%
Images	53.81	-10.74^{*}	(5.49)	-20.0%
Audio messages	13.02	-0.06	(3.79)	-0.5%
Mainstream news links	2.70	-0.73^{***}	(0.22)	-27.0%
Panel B: Other content outcomes				
Hate speech (strict)	0.0453	-0.0141	(0.0086)	-31.0%
Health misinformation	0.1178	$+0.0276$	(0.0280)	+23.4%
Electoral misinformation	0.0817	-0.0348^{***}	(0.0107)	-42.6%

Notes: This table presents secondary outcomes that do not appear in Table 2, using the same daily specification on the 319-group analytical sample (group and calendar-day fixed effects; standard errors clustered at the group level). Panel A covers granular activity measures (messages with links, videos, images, audio messages, and mainstream-news links). Panel B covers secondary content outcomes (strict hate speech, health misinformation, and electoral misinformation). Audio messages serve as a natural placebo because they are produced inside WhatsApp and have no upstream X dependency. Hate (strict) is the narrowest classifier, covering severe slurs directed at protected groups (F1 = 1.00 on five gold positives). Health and electoral misinformation are exploratory classifiers that were not validated against the audit sample because the three-day audit window contained too few positives. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

G.3 Specification Sensitivity (Full Table)

Table 22 reports the X-link ban-period coefficient across the four alternative specifications summarised in Section 6 ("Specification sensitivity" paragraph). The primary (group + calendar-day FE) specification yields -2.799 (SE = 0.348); adding group-level ISO-week FE in place of calendar-day FE yields -2.834 (SE = 0.355); adding day-of-week + holiday dummies yields -2.799 (SE = 0.348); and residualising against political-orientation pre-ban linear trends yields -2.835 (SE = 0.347). All four estimates agree within 1.3%.

Table 22: Robustness: X-link ban coefficient across specifications

	Ban effect on X-links	
Specification	Coef.	SE
Primary (group + day FE)	-2.799^{***}	(0.348)
Group + ISO-week FE	-2.834^{***}	(0.355)
Group + day FE + day-of-week + holiday	-2.799^{***}	(0.348)
Group + day FE + pol-orientation pre-ban trends	-2.835^{***}	(0.347)

Notes: This table presents the X-link ban-period DiD coefficient under four alternative specifications on the daily panel of the 319-group analytical sample (72,672 group-day observations). The primary specification uses group and calendar-day fixed effects. The second row replaces calendar-day fixed effects with ISO-week fixed effects; the third row adds day-of-week dummies and a Brazilian holiday dummy; the fourth row residualises the outcome against political-orientation-specific pre-ban linear trends, with the trends estimated on the pre-ban subsample only and projected forward to avoid the bad-control problem of fitting trends over the treatment period. Standard errors clustered at the group level. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

G.4 Full Event Study Figures

Figure 11 reports event-study estimates for the main activity outcomes (X-links, forwarded messages, total messages, messages with links, videos, unique senders). Figure 12 reports event-study estimates for the main content outcomes (potential misinformation primary, potential misinformation broad, aggression, fact-check links). Pre-ban coefficients cluster around zero across all outcomes, and the sharp break at $d = 0$ is visible for the significant outcomes identified in the main tables.

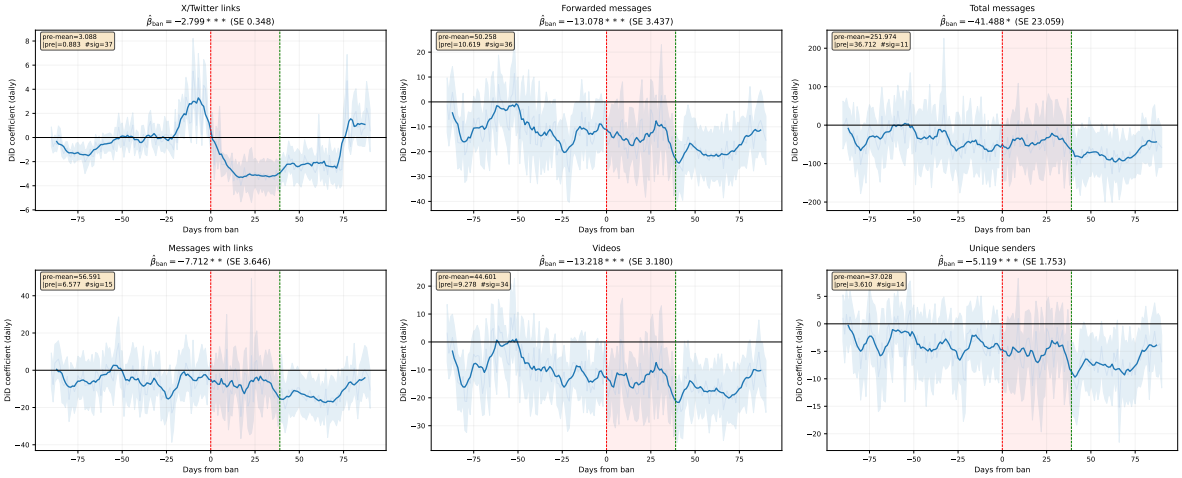


Figure 11: Event study estimates: main activity outcomes (Table 2, Panel A)

Notes: This figure presents daily event-study coefficients μ_k from Equation 2 for event-time $k \in [-90, +90]$ with reference day $k = -1$, estimated on the 319-group analytical sample. The faint blue line plots raw daily coefficients, the bold blue line plots the 7-day centred moving average, and the shaded band is the 95% confidence interval from group-clustered standard errors ($G = 319$). The dashed red vertical line marks ban onset at $d = 0$ and the dashed green vertical line marks ban end at $d = 39$. Outcomes shown include X-links, forwarded messages, total messages, messages with links, videos, and unique senders.

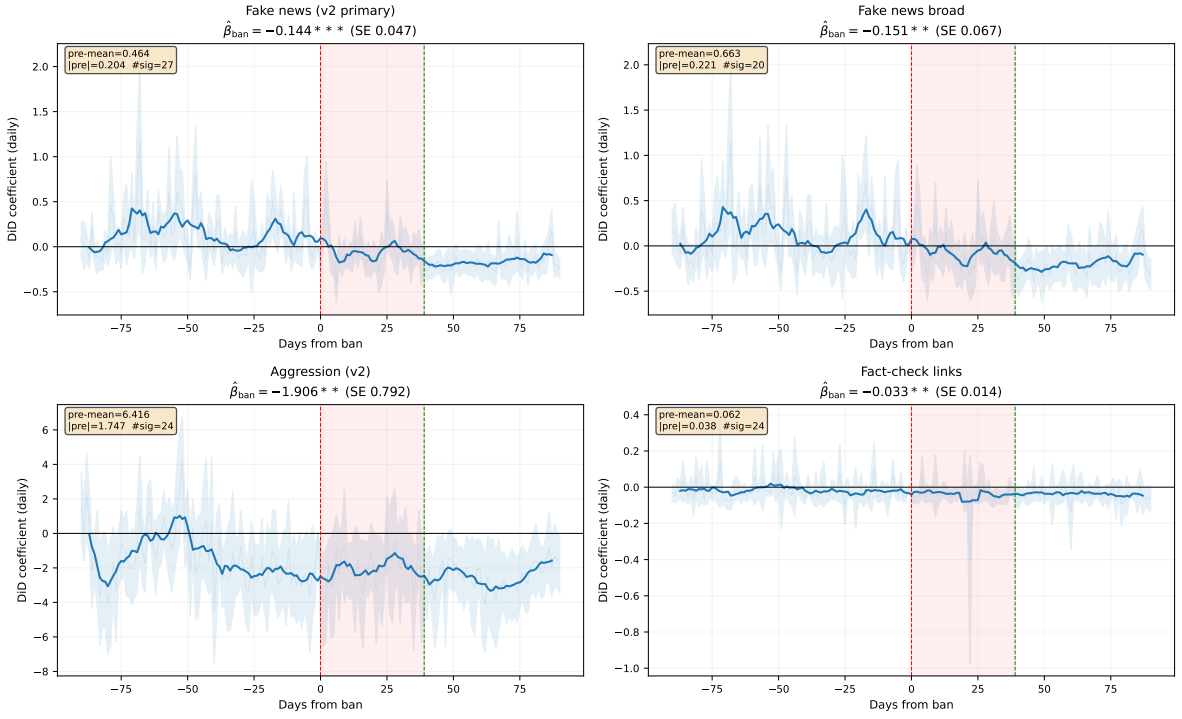


Figure 12: Event study estimates: content outcomes (Table 2, Panel B)

Notes: This figure presents daily event-study coefficients using the same specification as Figure 11. Outcomes shown are potential misinformation (primary), potential misinformation (broad), aggression, and fact-check links.

G.5 Pre-Trend F-Tests for Heterogeneity Buckets

Table 23 reports joint F-tests of pre-ban leads (weeks -8, -4, -2, -1) for each heterogeneity bucket in Table 4, across the five main outcomes. Most cells pass the pre-trend test at the 10% level. Exceptions concentrate on the X-link outcome for politically engaged buckets (politics, sports, right-wing groups), where the public escalation of the Brazilian Supreme Court's conflict with X in late August 2024 produced a temporary rise in X-link sharing. The direction of the pre-trend is opposite to the ban-period effect, so any bias works against finding an effect rather than toward it.

Table 23: Pre-trend F-tests for heterogeneity buckets (Table 4)

Bucket	N groups	X-links	Forwarded	Fake news	Aggression	Fact-check
Panel A: Group type (Table 4)						
Politics	243	6.15^{***}	5.23^{***}	3.42^{***}	3.26^{**}	1.38
		$[p = 0.000]$	$[p = 0.000]$	$[p = 0.008]$	$[p = 0.011]$	$[p = 0.240]$
Sales	42	5.05^{***}	0.55	2.04^{*}	0.28	0.56
		$[p = 0.000]$	$[p = 0.703]$	$[p = 0.086]$	$[p = 0.890]$	$[p = 0.695]$
Religion	17	0.82	2.90^{**}	0.99	1.82	0.53
		$[p = 0.515]$	$[p = 0.021]$	$[p = 0.414]$	$[p = 0.122]$	$[p = 0.714]$
Sports	13	5.86^{***}	1.92	2.59^{*}	2.30^{*}	–
		$[p = 0.001]$	$[p = 0.124]$	$[p = 0.051]$	$[p = 0.075]$	
Panel B: Political orientation (Table 5)						
Left	42	0.12	2.07^{*}	0.82	2.13^{*}	0.63
		$[p = 0.974]$	$[p = 0.082]$	$[p = 0.510]$	$[p = 0.074]$	$[p = 0.638]$
Right	131	8.44^{***}	4.17^{***}	1.73	2.92^{**}	1.53
		$[p = 0.000]$	$[p = 0.002]$	$[p = 0.140]$	$[p = 0.020]$	$[p = 0.189]$
Mixed	70	2.10^{*}	1.64	2.13^{*}	1.41	3.02^{**}
		$[p = 0.078]$	$[p = 0.162]$	$[p = 0.075]$	$[p = 0.226]$	$[p = 0.017]$

Notes: This table presents joint F-statistics from an event-study specification at the group-week level, testing the null that the lead coefficients at weeks -8, -4, -2, and -1 are jointly zero. Each cell corresponds to one of the heterogeneity buckets reported in Table 4. p-values are reported in brackets. Cells with insufficient within-bucket variation are marked with a dash. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

G.6 Pre-Trend F-Tests for Sender Buckets

Table 24 reports the equivalent diagnostic at the sender-bucket level used in Section 5.5 and Figure 3 and Figure 9. Of the 32 bucket-outcome pairs, 27 pass the pre-trend test at the 5% level. The persistent rejection is the top-5% bucket on the posted-X-link outcome ($F = 3.70$, $p < 0.01$), which traces to the same pre-ban political escalation.

Table 24: Pre-trend F-tests for sender activity buckets (Figure 3 and Figure 9)

Bucket	N senders	Messages	Posted X (1/0)	Forwarded	Posted FN (1/0)
Q1	4580	0.24	2.16*	0.49	0.99
		[$p = 0.918$]	[$p = 0.071$]	[$p = 0.744$]	[$p = 0.413$]
Q2	4369	1.37	1.39	2.29*	0.70
		[$p = 0.242$]	[$p = 0.233$]	[$p = 0.057$]	[$p = 0.595$]
Q3	4388	0.52	0.74	1.21	1.85
		[$p = 0.723$]	[$p = 0.564$]	[$p = 0.306$]	[$p = 0.116$]
Q4	2667	0.73	0.87	2.32*	0.82
		[$p = 0.571$]	[$p = 0.481$]	[$p = 0.054$]	[$p = 0.510$]
Top 10%	884	1.57	0.08	1.49	0.86
		[$p = 0.179$]	[$p = 0.988$]	[$p = 0.203$]	[$p = 0.486$]
Top 5%	711	1.09	3.70***	0.16	0.74
		[$p = 0.357$]	[$p = 0.005$]	[$p = 0.960$]	[$p = 0.565$]
Top 1%	160	1.97*	2.44**	0.32	1.75
		[$p = 0.096$]	[$p = 0.045$]	[$p = 0.863$]	[$p = 0.137$]
Top 0.1%	18	1.19	1.79	2.49**	1.03
		[$p = 0.311$]	[$p = 0.128$]	[$p = 0.041$]	[$p = 0.392$]

Notes: This table presents joint F-statistics from an event-study specification at the sender-week level, testing the null that the lead coefficients at weeks -8, -4, -2, and -1 are jointly zero. Each cell corresponds to one of the sender activity buckets reported in Figure 3 and Figure 9. p-values are reported in brackets. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

G.7 SUTVA: Control-Group Outcome Trends

Table 25 reports mean outcomes per group-day for the 160 control groups during the pre-ban period (January 1 to August 29, 2024) and the ban period (August 30 to October 8, 2024). The 58% decline in X-link sharing reflects the platform-wide supply shock; the near-zero change in potential misinformation (-1.4%) and the activity-proportional decline in aggression (-17% against -23% in total messages) confirm that content spillovers from

treated to control groups are negligible.

Table 25: Control-group outcome means: pre-ban vs ban period.

Outcome	Pre-ban mean	Ban mean	% change
X-links (Twitter/X)	0.5556	0.2311	-58.4%
Potential misinformation	0.1261	0.1244	-1.4%
Aggression	2.1644	1.7971	-17.0%
Total messages	143.7551	110.4461	-23.2%
Forwarded messages	23.8484	20.7138	-13.1%

Notes: Daily means computed across control groups (treated_g = 0) in the canonical 319-group sample. Pre-ban period: 2024-01-01 to 2024-08-29. Ban period: 2024-08-30 to 2024-10-08. Means are first averaged across groups within each day, then averaged across days.

G.8 Rambachan-Roth Sensitivity Bounds: Content Outcomes

Figure 13 applies the Rambachan and Roth (2023) relative-magnitude sensitivity analysis to the potential misinformation and aggression outcomes, using the same setup as Figure 6: weekly aggregation of the canonical 319-group event study, parameter of interest equal to the average treatment effect over post-ban weeks $k \in [0, 5]$, $\bar{M}$ grid $\{0, 0.5, 1, 1.5, 2\}$. For potential misinformation the breakdown $\bar{M}$ is 0.5: the negative effect retains significance when post-ban violations are no larger than half the observed pre-period drift, but loses significance beyond that threshold. For aggression the weekly-aggregated point estimate is positive and the original robust CI straddles zero; the daily DiD estimate is therefore not robust to event-time aggregation.

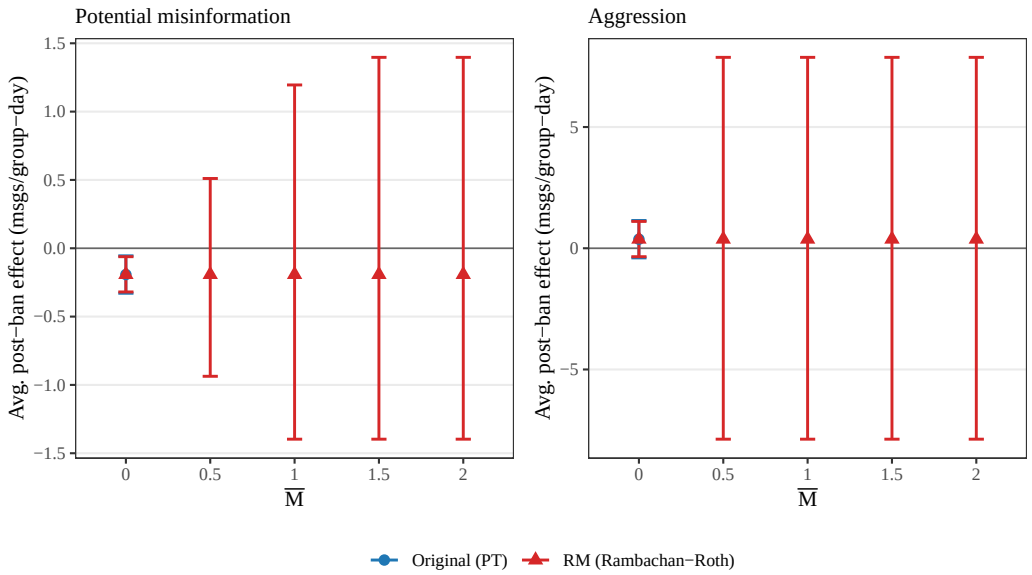


Figure 13: Rambachan-Roth sensitivity bounds: content outcomes

Notes: Sensitivity plots from Rambachan and Roth (2023) applied to the weekly canonical 319-group event study for potential misinformation (*fakenews_pattern_v2*, left panel) and aggression (*aggression_v2*, right panel). Specification and $\bar{M}$ grid identical to Figure 6. The breakdown $\bar{M}$ for potential misinformation is 0.5; for aggression the original robust CI straddles zero at $\bar{M} = 0$.