TACKLING HATE

From Shock to System

The Rise of Anti-Muslim Hate in Australia

Matteo Vergani, Susan Carland, Andrea Giovannetti, Stephanie Ng, Muhammad Sakib Khan Inan, Kewen Liao, Huu Phuc Hong, Yinsong Chen, Haily Tran, Amiee Taylor, Dan Goodhardt

Suggested citation: Vergani, M., Carland, S., Giovannetti, A., Ng, S., Inan, M.S.K., Liao, K., Hong, H.P., Chen, Y., Tran, H., Taylor, A., & Goodhardt, D. (2026). From Shock to System: The Rise of Anti-Muslim Hate in Australia. Tackling Hate Lab.

This report contains discussions of hateful and distressing language, including explicit examples, as well as themes of violence and discrimination. Some content may be confronting or upsetting for readers.

Please engage with this material at your own pace and seek support if needed.

The Tackling Hate Lab (www.tacklinghate.org) is a multi-institutional research initiative combining artificial intelligence, computational analytics, and social science to study online hate, extremism, and social cohesion through transparent, community-informed, and scientifically rigorous research.

EXECUTIVE SUMMARY

This report uses a big data approach to examine online anti-Muslim hate in Australia between January 2023 and January 2026, including more than 1 million online posts and hundreds of real-world, offline anti-Muslim incidents.

Drawing on multiple classification systems and computational methods, it analyses trends in anti-Muslim hate, anti-Palestinian hate, toxicity, and identity attacks targeting Muslims and Palestinians.

The report also examines how major trigger events reshaped online discourse, investigates the relationship between online anti-Muslim hate and verified offline incidents targeting Muslim communities, and analyses how hateful content spreads through online interaction networks.

By combining community-informed classification with large-scale computational analysis, the report provides a detailed and empirically grounded account of how anti-Muslim hate evolved and escalated to unprecedented levels during the Gaza war period and after the Bondi terrorist attack.

Key Findings

1. A structural shift.

Online anti-Muslim hate in Australia increased to a new baseline after 7 October 2023 and escalated further following the Bondi terrorist attack on 14 December 2025. Before October 2023, anti-Muslim hate remained relatively low and fluctuating, averaging 18.2 hateful posts per day. Following the Hamas terrorist attack and the start of the Gaza war, from October 7 2023 to 22 March 2025, that average increased to 121.3 posts per day. This represents a more than sixfold increase in daily levels. The escalation was not limited to overall hate volume: toxicity and identity attacks targeting Muslims also increased substantially, indicating a rise in explicitly hostile and identity-directed language. The Bondi attack produced a second and much larger rupture in online discourse. In the month following the attack, anti-Muslim hate increased more than ninefold relative to the previous month, rising from an average of 205.6 to approximately 1,917.9 posts per day. On 15 December 2025, anti-Muslim hate reached 7,786 posts in a single day, while toxicity targeting Muslims and identity attacks targeting Muslims also rose into the thousands. Unlike the post-October 2023 shift, which raised baseline discourse into the low hundreds, the post-Bondi period generated sustained daily counts in the high hundreds and thousands. The findings suggest that the Bondi attack activated an already elevated online hate ecosystem to an unprecedented scale.

Anti-Palestinian hate also increased after October 2023 and Bondi, but followed a different trajectory. While anti-Palestinian hate was highly volatile and event-driven, it did not exhibit the same sustained baseline growth observed in anti-Muslim hate.

The report finds that anti-Muslim hostility became increasingly embedded within dominant narratives linking Muslims to violence, terrorism, threat, and collective responsibility.

2. Demonstrable link between Australian offline and online anti-Muslim hate

The relationship between online anti-Muslim hate and offline incidents changed fundamentally after October 7 2023. Before the Gaza war, online and offline Anti-Muslim activity operated largely independently. Online hateful posts generated further posts in short bursts, but offline incidents did not measurably amplify online discourse.

After October 2023, this pattern shifted substantially. Previously, the two domains operated largely independently, but after October 2023, the offline and online domains formed a tightly coupled system in which real-world incidents often triggered sustained waves of online anti-Muslim hate. Each offline incident generated, on average, approximately 12 additional online hateful posts, while online self-excitation also intensified markedly. The online hate ecosystem became increasingly reactive and interconnected, with both online posts and offline incidents capable of generating prolonged waves of amplification lasting up to a full day.

One plausible explanation is that highly visible offline incidents attracted media coverage and public attention, increasing opportunities for hostile narratives to spread and gain social reinforcement online. In practical terms, this means that offline aAnti-Muslim incidents no longer remain isolated local events; they become catalysts for broader online mobilisation and identity-based hostility.

3. Online anti-Muslim hate morphed from fragmented users into organised, hierarchical webs

Network analysis shows that online anti-Muslim hate became more structurally organised after October 2023. While the overall volume of hateful interactions increased substantially, the online ecosystem also became more concentrated, cohesive, and hierarchical. A relatively small number of highly central users came to shape a growing share of hateful interactions, indicating that influence within the network became increasingly concentrated among key actors.

The report also finds that hateful networks underwent a process of crisis-driven expansion followed by reconsolidation. Immediately after October 2023, the influx of new users fragmented the network. Over time, however, these users became increasingly connected within large, cohesive clusters organised around central hubs. This indicates that anti-Muslim hate online is not simply diffuse or spontaneous, but circulates through identifiable pathways of amplification and influence.

Taken together, the findings suggest that online anti-Muslim hate in Australia became more reactive, more interconnected with offline events, more structurally concentrated, and more systemically embedded over time. The post-October 2023 period established a persistently elevated baseline of hostility, while the Bondi attack demonstrated the capacity for that environment to escalate rapidly into mass online mobilisation.

RECOMMENDATIONS

The findings of this report indicate that Australia is no longer managing a cyclical or episodic issue of anti-Muslim hate, but a new and elevated floor of hostility from which each successive trigger event launches a higher spike coordinated through structured networks.

Anti-Muslim online hate is now demonstrably much higher than pre-Oct 7 and operating in new ways. These findings should be taken as an early-warning alarm for foundational change in Australia's social cohesion, and are of prime relevance to the Royal Commission on Antisemitism and Social Cohesion.

R1. Fund a National Online Hate Observatory

We recommend that the Department of Home Affairs fund a stable and independent Observatory of Online Hate that uses a range of artificial intelligence models to monitor hate targeting diverse communities, including and especially the Muslim community. Such an observatory could provide ongoing threat assessment, early warning of escalating community tensions, and evidence-based insights to support timely prevention and policy responses.

Further discussion on R1.

- R1-A. This submission presents evidence that large-scale, longitudinal analysis of online discourse can identify structural escalation, network concentration, and narrative embedding before they manifest in visible offline harms. The Australian Government should fund the development of an independent, ongoing national online hate monitoring capability, modelled on the methodological approach demonstrated in this research, that tracks hate discourse across major platforms, provides real-time and quarterly reporting to relevant agencies, and feeds into the trigger-aware preparedness framework recommended above. This would represent Australia's first systematic early warning infrastructure for online hate and community leveraging a range of artificial intelligence tools.

- R1-B. The finding that anti-Muslim narratives associating Muslims with terrorism and collective threat have become embedded in mainstream online discourse and not confined to fringe platforms has direct national security implications. The normalisation of terrorism-linked framing and collective blame narratives creates a permissive discursive environment for radicalisation, harassment, and communally targeted violence. The Department of Home Affairs should integrate longitudinal online hate monitoring into existing threat assessment infrastructure and ensure that the escalation documented in this submission – and in future work of the proposed Online Hate Observatory – is formally registered in social cohesion risk frameworks and national threat assessments.

- R1-C. The data demonstrate a clear and consistent relationship between major national and international events (including domestic and international terror attacks and escalations in conflict in the Middle East) and surges in domestic anti-Muslim hate online. This pattern highlights the importance of establishing an Online Hate Observatory to monitor online hate trends in real time and to support the development and evaluation of interventions during periods of heightened social tension. Such an observatory would provide an evidence base to inform public statements and communications by Australian officials regarding international conflicts involving Muslim-majority countries, with greater awareness of how political and media discourse may unintentionally amplify or moderate domestic anti-Muslim sentiment. The Department of Home Affairs should work collaboratively with Department of Foreign Affairs and Trade and Department of the Prime Minister and Cabinet to develop communication guidelines that explicitly consider impacts on domestic social cohesion.

R2. Develop long-term interventions that move beyond short-term responses to trigger events

We recommend that all federal and state agencies involved in funding, planning and delivering interventions to mitigate online hate (e.g. education, awareness and literacy campaigns, inoculation interventions) address the structural conditions that sustain and escalate anti-Muslim hate over time, documented in this report. These interventions should be designed and implemented in partnership with community organisations to ensure they are culturally informed, locally grounded, and responsive to the lived experiences of affected communities.

Further discussion on R2.

- R2-A. Current intervention frameworks are predominantly reactive and incident-focused. The evidence presented in this submission demonstrates that structural escalation is analytically distinct from event-driven spikes.

- R2-B. The data demonstrate that three categories of trigger event are reliably associated with major surges in anti-Muslim online hate: local violent incidents, escalation in Middle East conflict, and symbolic anniversaries. This predictability allows for actionable response. The eSafety Commissioner should develop, in consultation with relevant government agencies and the Muslim community, a national high-risk calendar identifying foreseeable trigger windows (including anniversary dates of significant incidents). This calendar should serve as the basis for pre-emptive coordination for interventions with major platform operators, community organisations, and government communications teams.

R3. Support evidence-based content moderation practices

Current moderation approaches tend to focus on removing individual posts or accounts after harmful content has already spread. Our findings suggest that anti-Muslim hate operates through broader online ecosystems in which influential accounts, recurring narratives, and predictable trigger events shape the circulation and amplification of harmful content. This highlights the need for more evidence-based moderation practices that incorporate network analysis, narrative-based monitoring, and anticipatory responses during periods of heightened risk. Such approaches would enable platforms to intervene earlier and more strategically, rather than relying solely on reactive content removal.

Further discussion on R3.

- R3-A. Existing guidance to platform operators is predominantly content-focused: oriented toward identifying and removing specific posts or accounts on the basis of individual violations. The network evidence in this submission demonstrates that online hate ecosystems develop hierarchical structures in which a small number of structurally central actors disproportionately shape the circulation of hateful content. The eSafety Commissioner should issue guidance to platforms recommending the development of network-level analytical capacity, including the identification of structurally central accounts whose influence on hate flows exceeds what content-volume metrics alone would reveal.

- R3-B. The narrative categories documented in this report (terrorism association, collective blame, replacement theory, depictions of gender oppression, moral corruption framing, and violence/crime association) represent the dominant discursive forms through which anti-Muslim hate circulates online in Australia. The eSafety Commissioner should develop an evidence-based narrative typology framework and issue it as moderation guidance to platforms, providing both definitional clarity and graduated severity indicators. Replacement theory content warrants particular attention given its documented association with real-world violent extremism.

- R3-C. The evidence that anti-Muslim hate surges are reliably associated with identifiable trigger categories (local incidents, geopolitical escalation, and symbolic anniversaries) means major platform operators now have evidence to justify anticipatory rather than purely reactive moderation. Platforms operating in the Australian market should develop documented protocols for pre-positioning moderation capacity, reducing algorithmic amplification of at-risk content categories, and activating counter-narrative promotion strategies in advance of foreseeable trigger windows.

R4. Establish a whole-of-society framework to mitigate online hate in Australia

A whole-of-society framework is needed that treats online hate as a long-term social cohesion and public health challenge requiring coordinated prevention, monitoring, and response strategies. This approach should involve collaboration between governments, digital platforms, researchers, civil society organisations, and communities.

Further discussion on R4.

- R4-A. The evidence in this submission cuts across the mandates of the eSafety Commissioner, the Department of Home Affairs, the Attorney-General's Department, the Australian Federal Police, and ASIO's social cohesion functions, as well as state-level multicultural and community safety agencies. There is currently no formal mechanism for these bodies to coordinate responses to online hate escalation in real time. We recommend the Federal Government establishes a standing multi-agency taskforce with a specific mandate to coordinate early warning monitoring, trigger-event responses, and strategic counter-narrative activity across agency boundaries.

- R4-B. Muslim community organisations, interfaith bodies, and civil society represents one of the most effective long-term partners to tackle hate in society. However, this capacity is currently critically under-resourced, and often ad hoc. Government should fund long-term partnerships with community organisations for both rapid- and long-term response and proactive mitigation of online hate.

- R4-C. Algorithmic recommendation and engagement-maximisation systems may be materially contributing to the elevation of harmful actors within hate network hierarchies. Platform operators should commission and publish audits of their recommendation systems with specific attention to whether these systems disproportionately amplify accounts or content that is structurally central in identified hate networks.

INTRODUCTION

1.1. Aims

This report has four primary aims:

1. Map online discussions manifesting anti-Muslim hate and anti-Palestinian hate in Australian online discourse;
2. Track trends in online hate, identifying how trigger events reshape baseline levels of hostility;
3. Examine the relationship between online anti-Muslim discourse and verified offline incidents, assessing whether activity in one domain amplifies the other;
4. Investigate how hate spreads through user interactions, with a focus on influence, concentration of activity, and changes in network structure following major events.

1.2. Data

This report analyses a large dataset of online content captured between 1 January 2023 and 19 January 2026. The dataset contains **1,299,734** individual entries (posts and interactions) authored by 63,680 unique accounts. The dataset was collected using a query designed by the research team and consisting of keywords, hashtags, and phrases designed to retrieve posts containing several overlapping types of conversation. First, it captures content related to protests, marches, strikes, vigils, encampments, and other forms of collective action connected to Palestine and Gaza, including student movements and trade union activism. Second, it retrieves discussions that explicitly link Islam or Muslim identity to crime or terrorism, for example posts that mention Islam or Muslims alongside references to terrorism, criminality, or security threats. Third, it includes explicit hashtags and slogans commonly used in hostile or derogatory content targeting Muslims, Islam, Palestinians, or pro-Palestinian supporters, as well as posts that frame Palestinians as illegitimate, criminal, or genocidal. These terms were included in the query to capture how such framings appear in online conversations. Taken together, this query captures a broad discursive environment in which anti-Muslim, anti-Palestinian hostility, political mobilisation, factual reporting, and counter-speech coexist. Most entries come from X, accounting for **936,327** posts geolocated in Australia. Content from X includes original posts, replies, and shares, allowing us to observe not only how anti-Muslim hate is expressed, but also how it is amplified through retweeting and reactive exchanges. X is particularly important for this analysis because it plays a central role in real-time political debate in Australia and allows users to rapidly circulate emotionally charged or polarising content. While X is used by about 17% of Australians, around 4.7 million people,[1] its main value lies in scale, openness, and near real time access. Researchers can collect large volumes of data, including posts, replies, reposts, likes, and geolocation metadata, which allows identification of Australian content. Other platforms such as Facebook and Instagram restrict programmatic access, limiting large scale analysis. Since 2022, reduced moderation has also made X a key space for political discourse, including fringe content, making it useful to track the emergence and spread of hate narratives over time. The dataset includes a substantial volume of news content (**292,965** entries) drawn from 1,586 Australian media outlets. These include national and metropolitan sources such as The Sydney Morning Herald, Herald Sun, Sky News, and MSN, as well as numerous online news sites, regional and local newspapers across Australia. This diversity allows us to capture how discussions about Muslims, Palestine, and related issues appear not only in opinionated commentary, but also in news reporting, headlines, and comment sections, which can shape public interpretation of events. Smaller but analytically important portions of the dataset come from online forums and blogging platforms. Forum content (**10,547** entries) is primarily drawn from Australian discussion boards such as BigFooty and HotCopper, where users often engage in longer, less moderated discussions around politics, national identity, and current affairs. Tumblr (**57,572** entries) contributes a distinct form of content, characterised by highly expressive posts, visual material, and the circulation of memes and slogans, which can play a role in spreading both overt and coded forms of hate. Blogs and other minor sources account for a small number of entries (1,564 entries) but provide additional context for how narratives circulate across different parts of the online ecosystem. Across all platforms, the dataset includes original posts (394,635 entries), replies (143,472 entries), and shares (761,627 entries). This mix is critical, as hate does not spread only through original messages, but also through repetition, endorsement, and contestation, with reposting and replying often extending the reach and lifespan of hostile content. The full list of news outlets, blogs, forums and other sites included in the analyses is available upon request.

The second dataset comprises **309** Anti-Muslim incidents reported by community members to the Islamophobia Register Australia between 1 January 2023 and 30 November 2024 (Carland et al., 2025). Incidents were reported through the Register's online platform, which collects information on the timing, location, context, and characteristics of both victims and perpetrators. To ensure data quality, only completed and formally submitted reports were included. Register staff conducted follow-up contact with reporting individuals to verify incidents, and also to obtain additional details where required. Each verified case was then individually reviewed to remove duplicate, out-of-period, or vexatious reports, ensuring a reliable and rigorously validated dataset.

1.3. Approach to classification

We used a range of tools to classify online data. Firstly, we used existing tools such as Perspective API, which provides automated measures of harmful language at scale, with a focus on general toxicity and identity-based attacks.

Toxicity captures the degree to which a comment is rude, disrespectful, or likely to make others leave a conversation. Identity attack focuses more specifically on content that targets individuals or groups based on characteristics such as religion, ethnicity, or gender. These measures are useful for large-scale monitoring of online discourse, but they rely on general definitions of harm and are not tailored to how specific communities experience hate. Once this classification was applied across the dataset, we used Qwen, an advanced artificial intelligence system trained to interpret human language, to identify who or what each piece of harmful content was directed at. The model was applied consistently across all posts to determine whether the content targeted Muslims or Palestinians. Each target group was coded separately, allowing us to distinguish between different forms of targeting and report them clearly in the results.

[1] https://sproutsocial.com/insights/social-media-statistics-australia/

Second, we used a machine learning classifier developed by the Tackling Hate Lab to identify anti-Muslim content based on how it is perceived by members of the Australian Muslim community. The model was trained on social media text annotated by contributors recruited through the Islamophobia Register Australia. As a result, the classifier reflects community-informed understandings of anti-Muslim hate. The classifier analyses text and determines whether a post should be classified as anti-Muslim hate or not.

The model was developed from an initial dataset of 2,700 unique social media text samples. Four trained community annotators reviewed the posts, with each sample assessed by multiple annotators. To maximise reliability, we applied a strict data-cleaning process that removed posts with unusable annotations and excluded cases where only two annotators reviewed a post but disagreed. Final labels were assigned using a majority-decision rule. The resulting dataset contained 2,093 unique text samples, including 871 labelled as hateful and 1,222 labelled as not hateful. We tested a range of contemporary language models and training approaches. The best-performing model was RoBERTa-Large, a state-of-the-art language model adapted to this task using partial fine-tuning. The final classifier performed strongly across standard evaluation measures. Overall, it correctly classified 92.6% of posts. When the model identified a post as hateful, it was correct in 89.1% of cases, indicating a relatively low rate of false positives. It also successfully detected 93.7% of hateful posts, meaning relatively few were missed. A combined performance score (F1) of 91.3% demonstrates a strong balance between these measures. The classifier also showed a high capacity to distinguish between hateful and non-hateful content across different decision thresholds (AUC: 97.6%). Taken together, these results indicate that the classifier is both technically robust and closely aligned with how anti-Muslim hate is understood by the community members who contributed to its development.

We then used topic modelling to distinguish between anti-Muslim hate and anti-Palestinian within the content identified by our classifier. Using BERTopic, we grouped similar posts into clusters based on shared language patterns, and applied GPT-4o to generate descriptive labels for each topic.

2. UNDERSTANDING ONLINE DISCUSSIONS

To map online discourse (Aim 1), we conducted a thematic analysis of the language contained in the dataset capturing conversations related to Muslims, Islam, Palestine, Gaza, and associated protests and political debates in Australia. This analysis was conducted using BERTopic, a machine-learning tool that groups large volumes of text into clusters based on semantic similarity. In simple terms, BERTopic identifies patterns in how people talk about an issue and groups together posts that use similar language or express similar ideas. To ensure that these clusters were interpretable and aligned with the aims of the study, we used a large language model to generate clear, structured labels and descriptions for each topic.

The thematic analysis revealed a highly uneven distribution of narratives. The most prominent category was Pro-Palestinian discourse, which accounted for nearly 35% of all posts.

This category includes expressions of solidarity with Palestinians, calls for protest or political action, and emotional responses such as grief, anger, or empathy. The next largest categories were Anti-Palestinian discourse, Terrorism and Anti-Israel discourse, which include highly polarising discussions. A substantial portion of the dataset (about 10%) also consisted of factual reporting, including posts that present information in a descriptive or informational manner without explicit evaluative language, reflecting the circulation of news articles, images, videos, and descriptive updates without overt opinion. Smaller but analytically important categories include meta-discussions of antisemitism and Islamophobia, debates about Western governments and Australian politics, and explicitly pro-Israel discourse. The full list of categories is presented in Table 1.

Table 1. List of topics identified in the dataset capturing conversations about Muslims, Islam, Palestine and Gaza

Category	Description	Frequency	Percentage
Pro-Palestinian Discourse	Posts expressing support for Palestinians, including political mobilisation, protest, donations, or emotional expressions such as grief or outrage.	459,082	35.5%
Anti-Palestinian Discourse	Posts expressing hostility or opposition to Palestinians and associated groups, including accusations of terrorism, war crimes, or opposition to pro-Palestinian activism.	192,785	14.9%
Terrorism	Posts discussing terrorism. They often associate Muslims or Islam with crime, violence, or security threats.	184,247	14.2%
Anti-Israel Discourse	Posts expressing hostility or opposition to Israel and associated groups or institutions, including accusations of war crimes or genocide.	164,313	12.7%
Factual Reporting	Posts sharing descriptive or documentary information such as news updates, images, videos, or statistics without clear opinion.	136,201	10.5%
Meta-Discussion of Antisemitism	Posts where antisemitism itself is the object of discussion or analysis.	49,663	3.8%
Western Countries Actions	Posts discussing Western governments, leaders, institutions, and Australian politics.	48,530	3.7%
Meta-Discussion of Islamophobia	Posts where Islamophobia itself is the object of discussion or analysis.	17,332	1.3%
Pro-Israel Discourse	Posts expressing support for Israel, including political or emotional expressions of solidarity.	15,083	1.2%
Unrelated Content	Posts unrelated to Islam, Muslims, Palestine, Gaza, antisemitism, or Islamophobia.	10,180	0.8%
Non-Western Countries Actions	Posts discussing actions of non-Western states and actors.	4,924	0.4%
Global conflicts	Posts discussing conflicts unrelated to Israel–Palestine (e.g. Ukraine)	259	0.1%
Other	Residual category capturing related content not covered above.	11,867	0.9%

We then analysed the subset of posts classified as anti-Muslim hate using our community-informed classifier, which captures perceptions of our community annotators. Within this content, we distinguished analytically between anti-Muslim hate (targeting Muslims or Islam more broadly) and anti-Palestinian hate (targeting Palestinians or pro-Palestinian supporters, often through Anti-Muslim tropes).

Among posts classified as anti-Muslim hate, the most common subtype involved linking Muslims or Islam to terrorism (about 76% of all the content classified as anti-Muslim hate), followed by forms of collective blame, where all Muslims are held responsible for violence committed by individuals, groups, or states. Other recurrent themes included conspiratorial claims about demographic "replacement", efforts to delegitimise or suppress pro-Palestinian protest, and depictions of Islam as inherently immoral, violent, or oppressive. The full breakdown of anti-Muslim and anti-Palestinian hate subtypes is reported in Table 2.

Table 2. Categories of anti-Muslim and anti-Palestinian hate

Category	Description	Frequency	Percentage
Terrorism Link	Associating Muslims or Islam with terrorism (e.g. ISIS, Al-Qaeda).	155,115	76.7%
Collective Blame	Holding all Muslims responsible for actions of individuals or groups.	18,038	8.9%
Terrorist Support*	Claims that pro-Palestinian supporters are terrorists or support terrorism.	6,383	3.2%
Protest Suppression*	Claims that pro-Palestinian protests are illegitimate and should be shut down.	5,260	2.6%
Replacement Theory	Claims that Muslims aim to replace White or European populations through migration.	4,971	2.5%
Moral Corruption	Portrayals of Islam as inherently evil, corrupt, or immoral.	3,311	1.6%
Women Oppression	Depictions of Islam or Muslims as inherently oppressive toward women.	3,377	1.7%
Moral Corruption	Portrayals of Islam as inherently evil, corrupt, or immoral.	3,311	1.6%
Violence and Crime	Associations between Muslims and violence, crime, or toxic masculinity.	3,121	1.5%
Criminality Claims*	Claims that pro-Palestinian supporters are inherently criminal or dangerous.	1,828	0.9%
Other	Residual category not captured above.	880	0.4%

*These categories together form the broader, superordinate category of anti-Palestinian hate.

Together, these findings show that anti-Muslim hate in Australian online discourse is heavily concentrated in narratives that frame Muslims as violent, collectively culpable, or fundamentally incompatible with social order, and that such narratives are closely intertwined with debates about security and migration.

3. ONLINE ANTI-MUSLIM HATE TRENDS

To track trends in online hate (Aim 2), we examined changes in the volume and intensity of anti-Muslim content in the online discussions captured by our queries before and after two major events: 7 October 2023, the date of the Hamas-led terror attack in Israel, and 14 December 2025, the date of the Bondi terror attack in Sydney. In addition to overall counts of anti-Muslim and anti-Palestinian hate, we also examined two related indicators of hostility: toxicity targeting Muslims and Palestinians, and identity attacks targeting Muslims and Palestinians.

The data show two distinct ruptures. The first occurred after 7 October 2023, when anti-Muslim hate shifted from low daily counts to a sustained baseline in the low hundreds. The second, following Bondi, was far larger in scale, with daily anti-Muslim hate reaching 7,786 posts on 15 December 2025 and other hostility indicators also rising into the thousands. This indicates that Bondi did not merely extend the post-October 2023 pattern, but produced a much sharper and more concentrated escalation.

Figures 1 to 6 present 7-day rolling mean trends for all six hate and toxicity indicators across the observation period.

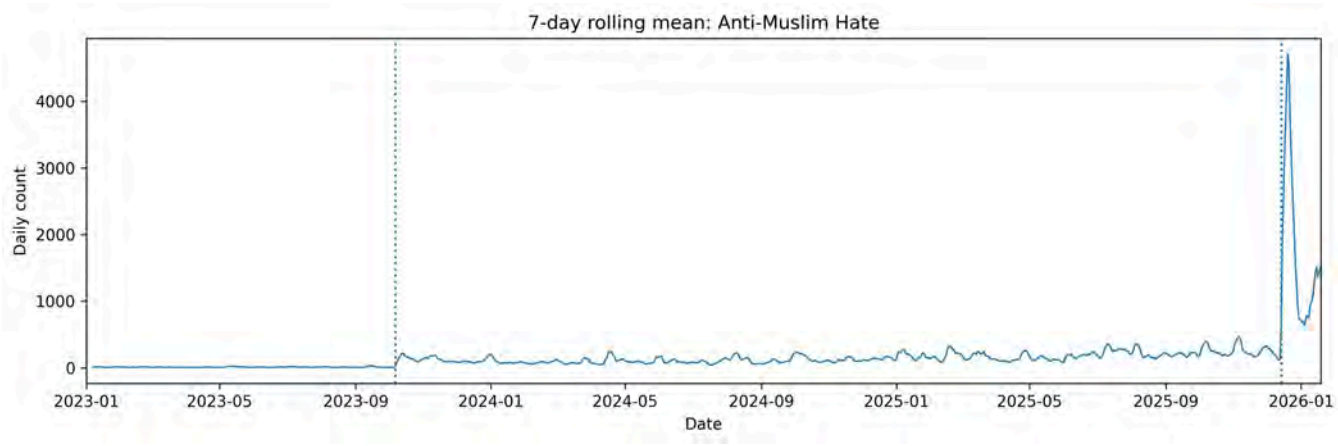


Figure 1. Seven-day rolling mean of anti-Muslim hate posts across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

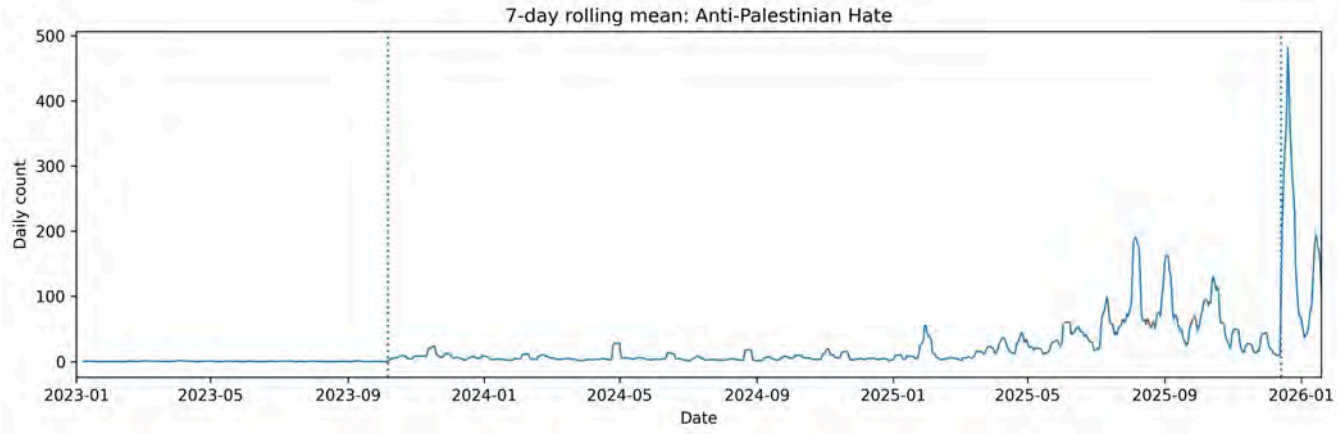


Figure 2. Seven-day rolling mean of anti-Palestinian hate posts across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

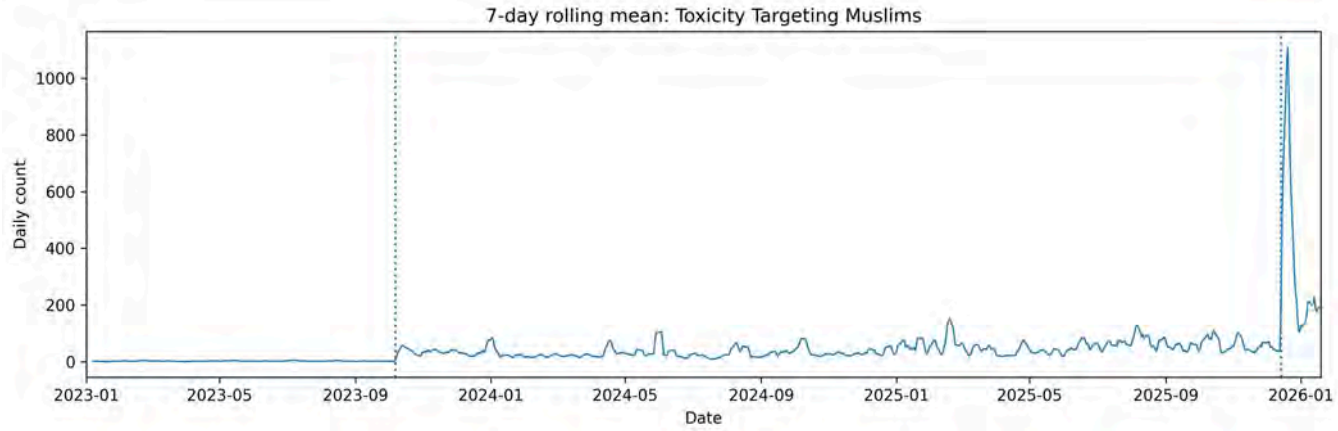


Figure 3. Seven-day rolling mean of toxicity targeting Muslims across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

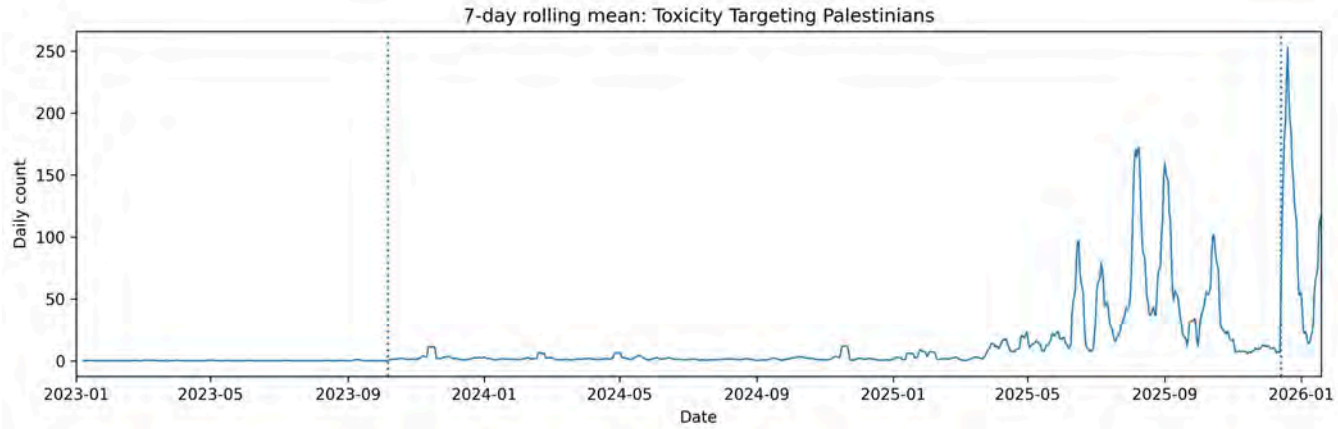


Figure 4. Seven-day rolling mean of toxicity targeting Palestinians across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

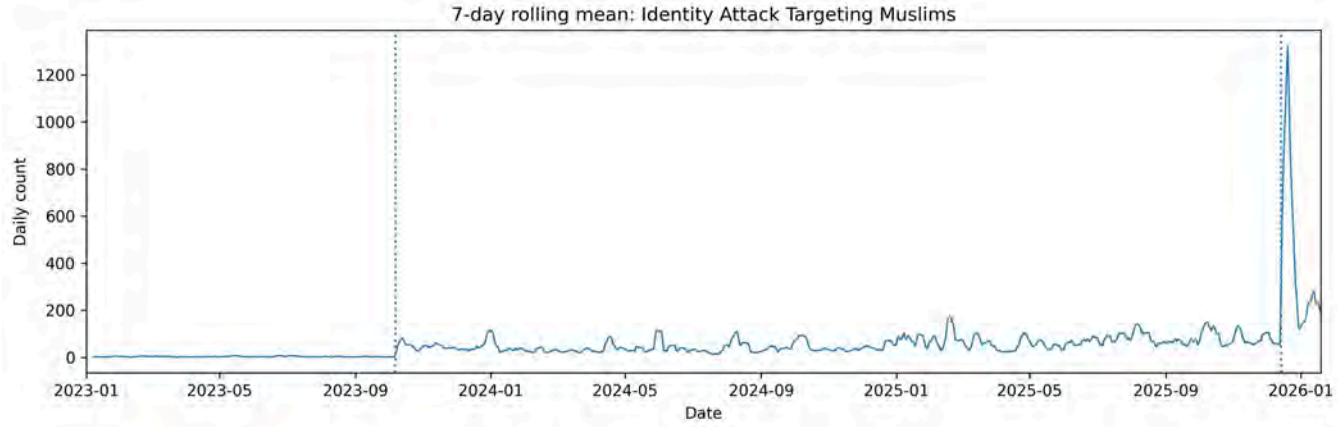


Figure 5. Seven-day rolling mean of identity attacks targeting Muslims across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

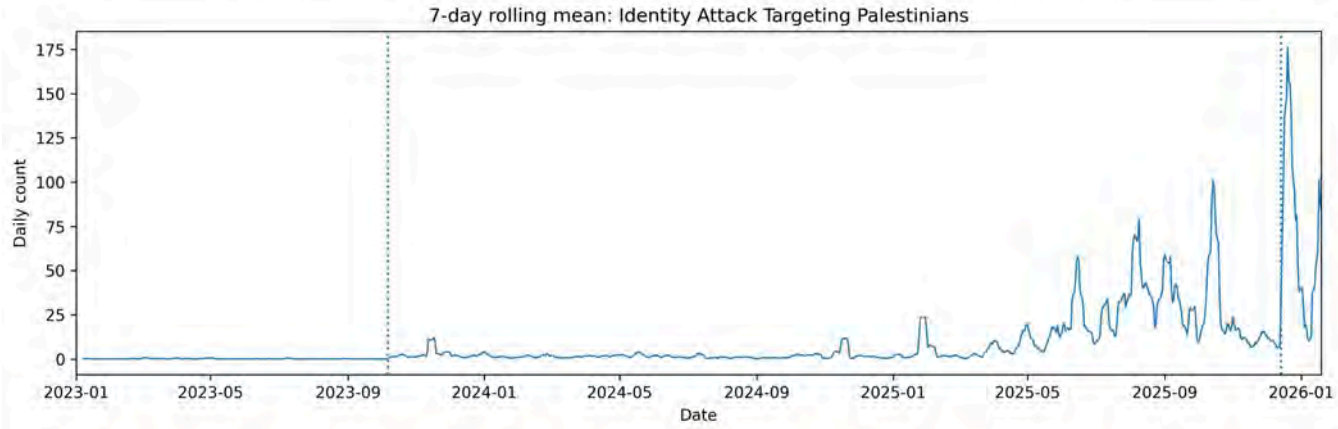


Figure 6. Seven-day rolling mean of identity attacks targeting Palestinians across all platforms (1 January 2023 – 19 January 2026). The first dotted line marks 7 October 2023, and the second marks 14 December 2025.

3.1. Trends before and after 7 October 2023

October 7 2023 marked a clear structural break and produced durable shifts in domestic online discourse. Between 1 January 2023 and 6 October 2023, anti-Muslim hate, as captured by our community-informed classifier (Figure 1), averaged 18.2 posts per day (standard deviation = 17.3). From 7 October 2023 to 22 March 2025, that average increased to 121.3 posts per day (standard deviation = 82). This represents a more than sixfold increase in daily levels. The rise is not only statistically visible but also substantively large and sustained over time: in late 2023 and early 2024, the 7-day average typically sits between 60 and 120 posts per day. In later months, it frequently exceeds 150 and at times rises above 200 or even 300 following major peaks. A negative binomial model estimates an increase of approximately 3 percent per month if compounded. Both the statistical coefficient and the visual trend indicate a meaningful upward drift.

Anti-Palestinian hate (Figure 2) also increased after 7 October 2023, but from a much lower baseline. Before 7 October 2023, it averaged 0.7 posts per day (standard deviation = 1.3). After 7 October 2023, it averaged 7 posts per day (standard deviation = 12.6). While this is a tenfold increase in proportional terms, the absolute scale remains far below that of anti-Muslim hate. When smoothed using a 7-day moving average, the baseline remains relatively stable, fluctuating between roughly 3 and 8 posts per day. A negative binomial time trend model detects a small but statistically significant upward trend (p-value = 0.046). Importantly, there was a visible increase in anti-Palestinian hate around the third quarter of 2025 (roughly between July and September 2025), which then returned closer to baseline before the new peak triggered by the Bondi attacks.

Similar trends are visible across all other classifications used to capture anti-Muslim and anti-Palestinian hate: anti-Muslim toxicity (Figure 3) and anti-Muslim identity attack (Figure 5), as well as anti-Palestinian toxicity (Figure 4) and anti-Palestinian identity attack (Figure 6).

3.2. Trigger Events and Symbolic Dates Associated with Spikes in Online Hate

Before the Bondi terror attack, the largest spikes in online anti-Muslim and anti-Palestinian hate were linked to two types of trigger events: local incidents and symbolic anniversaries.

First, local events produced sharp surges. For example, on 16 April 2024, the day after the Wakeley church stabbing in Sydney, 568 posts recorded that day contained anti-Muslim hate. This represents one of the highest absolute volumes in the dataset and reflects the rapid attribution of violence to Muslim identity. Similarly, the clashes between pro-Palestinian and pro-Israel protesters on 12 November 2023 in Caulfield (Melbourne), during which a fire destroyed a burger shop and a synagogue had to be evacuated, were associated with a spike in anti-Palestinian hate (75 posts contained anti-Palestinian hate).

Second, symbolic anniversaries reactivated historical grievance frames. On 11 September 2023, 109 of the 309 posts recorded that day (35.3 percent) were hateful. Unlike event-driven spikes, this reflects anniversary-based reactivation of 9/11 narratives. Importantly, other symbolic dates, such as 25 April (Anzac Day) and 29 January (anniversary of the liberation of Auschwitz), which often trigger polarised online discussions, were also associated with spikes in anti-Palestinian hate (respectively 129 and 155 posts containing anti-Palestinian hate).

For policy and programs, this suggests the need for anticipatory monitoring around predictable trigger dates, rapid-response coordination following violent incidents, and proactive counter-narrative strategies during periods of geopolitical escalation.

3.3. Before and after 14 December 2025

The Bondi terror attack on 14 December 2025 represents the largest escalation in online anti-Muslim hate observed in the dataset.

In the month prior to the attack (14 November 2025 to 13 December 2025), anti-Muslim hate averaged 205.6 posts per day. In the month following the attack (14 December 2025 to 13 January 2026), that average increased to approximately 1,917.9 posts per day. This represents more than a ninefold increase in daily anti-Muslim hate volumes over an already elevated post-October 2023 baseline. The escalation was immediate and extreme. On 15 December 2025, the day after the attack, anti-Muslim hate reached 7,786 posts, the highest level observed in the dataset. Unlike the October 2023 break, which shifted baseline discourse into the low hundreds, the post-Bondi period produced sustained daily counts in the high hundreds and thousands.

The increase was not limited to overall hate volume. Toxicity targeting Muslims (Figure 3) and identity attacks targeting Muslims (Figure 5) also rose sharply after Bondi. In the month before the attack, toxicity targeting Muslims averaged 52.1 posts per day and identity attacks targeting Muslims averaged 72.8 posts per day. In the month afterwards, these averages increased to 409.5 and 538.8 posts per day respectively. Peak levels were substantially higher. On 15 December 2025, toxicity targeting Muslims reached 2,114 posts and identity attacks targeting Muslims reached 2,151 posts. These figures indicate that the post-Bondi escalation involved not only greater discussion volume, but a major intensification in explicitly hostile and identity-directed content targeting Muslims.

Anti-Palestinian hate also increased following Bondi, although at a substantially lower scale. In the month before the attack, anti-Palestinian hate averaged 20.9 posts per day. In the month afterwards, it averaged 196.8 posts per day, representing roughly a ninefold increase. Toxicity targeting Palestinians and identity attacks targeting Palestinians also rose after Bondi, with several spikes reaching into the low hundreds. However, the absolute scale of these increases remained far below equivalent anti-Muslim indicators.

4. THE RELATIONSHIP BETWEEN ONLINE LANGUAGE AND OFFLINE ANTI-MUSLIM INCIDENTS

To examine the relationship between online Anti-Muslim language and offline incidents (aim 3), we used a statistical modelling approach known as a Hawkes process. This method estimates whether one type of event (for example, an Anti-Muslim incident offline) triggers a surge in anti-Muslim hateful posts on social media. In practical terms, this allows us to trace how hate propagates over time and to assess whether real-world incidents amplify online hostility, or whether intensified online discourse itself contributes to further surges in hate.

To conduct this analysis, we examined the relationship between online anti-Muslim hate and the 309 Anti-Muslim incidents reported by community members to the Islamophobia Register Australia (Carland et al., 2025), which we further characterised into 15 non-exclusive categories for descriptive purposes. Table 3 provides an overview of the incident types recorded in our dataset, which range from verbal abuse to physical violence. We analysed their relationship with the overall volume of Anti-Muslim activity in our online dataset, allowing the model to identify both self-effects (within the same stream) and cross-effects (between online and offline streams). The model covers a subset of our data, from 1 January 2023 to 22 March 2025, enabling us to identify patterns both before and after October 7 2023.

Type	Before October 7 2023	After October 7 2023	Total
Verbal Abuse	18	196	214
Non-verbal Abuse	4	58	62
Discrimination Not Otherwise Categorised	4	29	33
Property Damage	0	25	25
Physical Assault	4	32	36
Threat of Violence	2	37	39
Incidents containing anti-Palestinian bias indicators	0	107	107
Anti-Palestinian Verbal Remarks	0	77	77
Anti-Muslim Verbal Remarks	16	134	150
Anti-Immigrant Verbal Remarks	9	40	49
Anti-Arab Verbal Remarks	1	9	10
Pro-Palestinian Symbols	0	55	55
Pro-Palestinian Vigil Protest	0	36	36
Target Religious Buildings	1	12	13
Gendered Insults	3	30	33
Total	62	877	939

Table 3. Distribution of offline Anti-Muslim incidents before and after October 7 2023. Classes are non-exclusive, as a single incident can be classified in multiple classes of attacks. Because incidents may involve multiple forms of abuse, the total number of classified incidents across categories (939) exceeds the number of unique incidents recorded (309).

4.1. Statistical analyses

To assess the impact of the October 7 2023 attacks, we divided the analysis into two periods. We treated the online and offline streams of events as separate "timelines" within a multidimensional system. The Hawkes model then estimated how much activity in one timeline (e.g. offline events) influenced activity within the same stream and in the other stream (e.g. one Anti-Muslim tweet triggering additional tweets).

We used a non-parametric version of the model, meaning we did not impose strong assumptions about how influence unfolds over time. Instead, we allowed the data to determine how often one type of event occurred shortly after another. We assumed that any given tweet could influence other tweets for up to 10 hours, with the effect gradually fading over time. This 10-hour window was divided into 50 equal intervals to track how influence rises and decays.

The only difference between the two sets of figures is the timeframe: whether the data include the full period up to October 7 2023 or the period after that date.

4.2. Results

Figure 7 displays the estimated influence relationships between anti-Muslim online activity and offline incidents in the period before October 7 2023. Each cell in the matrix represents how much one type of event triggers additional events of the same or the other type, with darker shading indicating a stronger causal link. The colour scale adopted is relative to the figure only. Prior to October 7, online anti-Muslim activity was largely self-exciting: each hateful post generated on average 0.77 additional posts, while cross-influence between online discourse and offline incidents was negligible. Critically, cross-influence between the two domains was negligible. Offline anti-Muslim incidents had virtually no measurable effect on online activity, and online posts did not trigger offline events. This suggests that, before the attacks, the online and offline spheres of anti-Muslim activity operated largely independently.

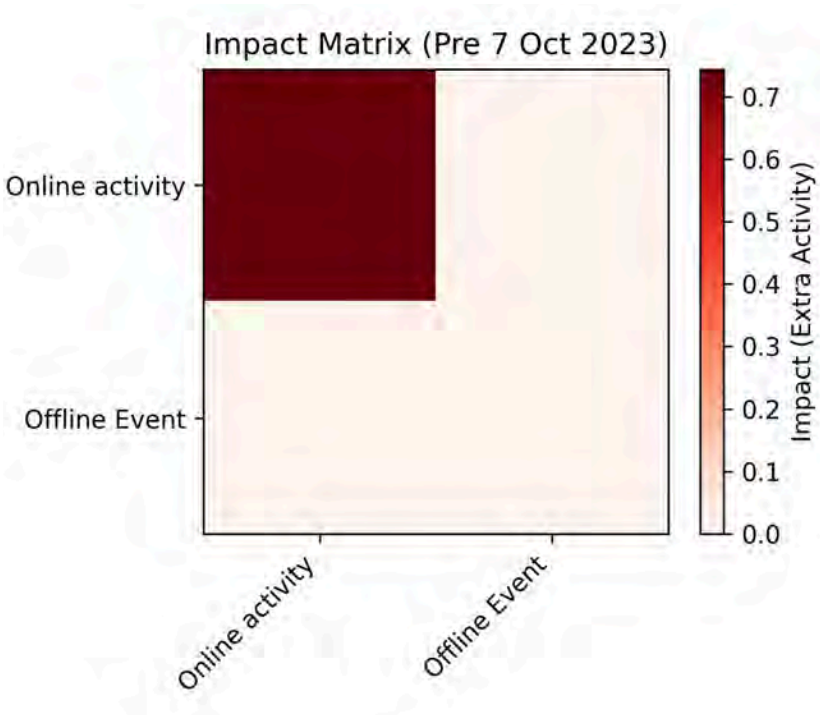


Figure 7. Pre-7th October Attacks. Influence Matrix

Figure 8 shows how the triggering effects identified in Figure 7 evolved over the hours following an event and contextualises the total effect, which was negligible in magnitude. The self-excitation of online anti-Muslim activity —the dominant effect in this period—peaked sharply within the first one to two hours after a post, before gradually decaying. By approximately five to six hours, this already modest influence had largely dissipated. Cross-stream effects between online posts and offline incidents showed a delayed build up after 2h and sharp drop at 5h. This temporal profile confirms that, prior to October 7, online anti-Muslim discourse was a self-sustaining but short-lived process: hateful posts tended to trigger further posts in quick succession, but this cascading effect faded rapidly, with no evidence of meaningful spillover between online and offline domains.

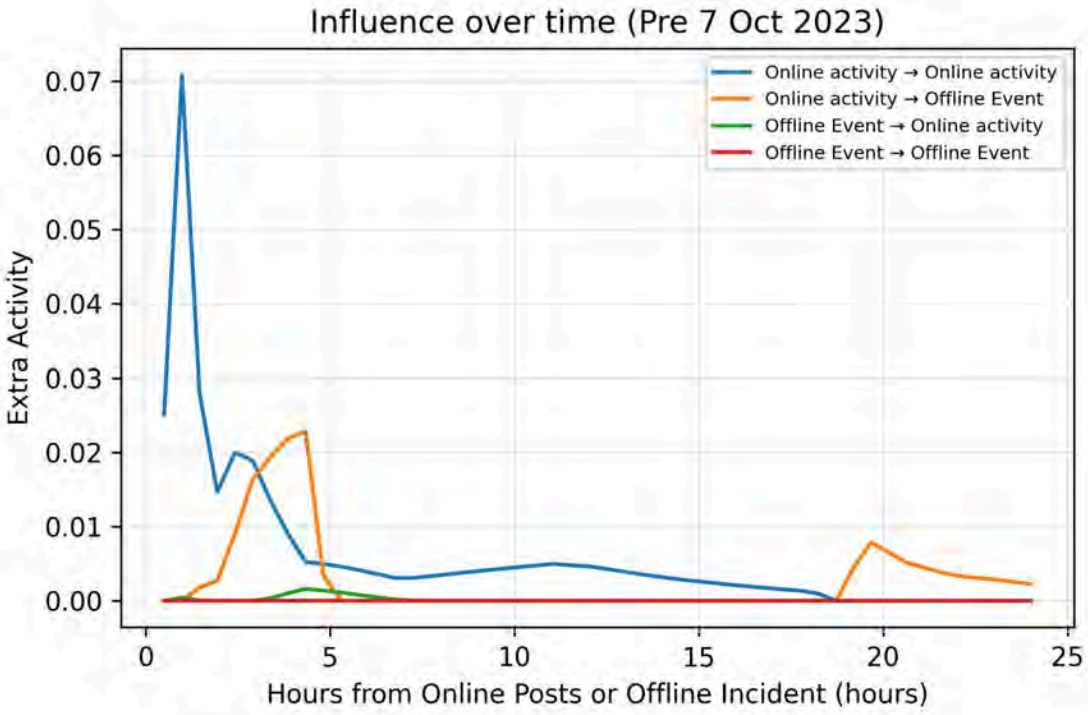


Figure 8. Pre-7th October Attacks. Influence over Time

Figure 9 presents the same influence analysis for the period following the October 7 2023 attacks. The contrast with the pre-attack period is clear. The most significant change is the emergence of a substantial cross-influence from offline anti-Muslim incidents to online activity. Each offline event now triggered, on average, approximately 12 additional online hateful posts—a large increase from near zero in the earlier period, particularly in light of the much higher number of offline incidents taking place in the post-October 7 period. Online self-excitation also intensified, moving from 0.07 to 2.5 extra activity, thus indicating that hateful content became even more effective at triggering further posts.

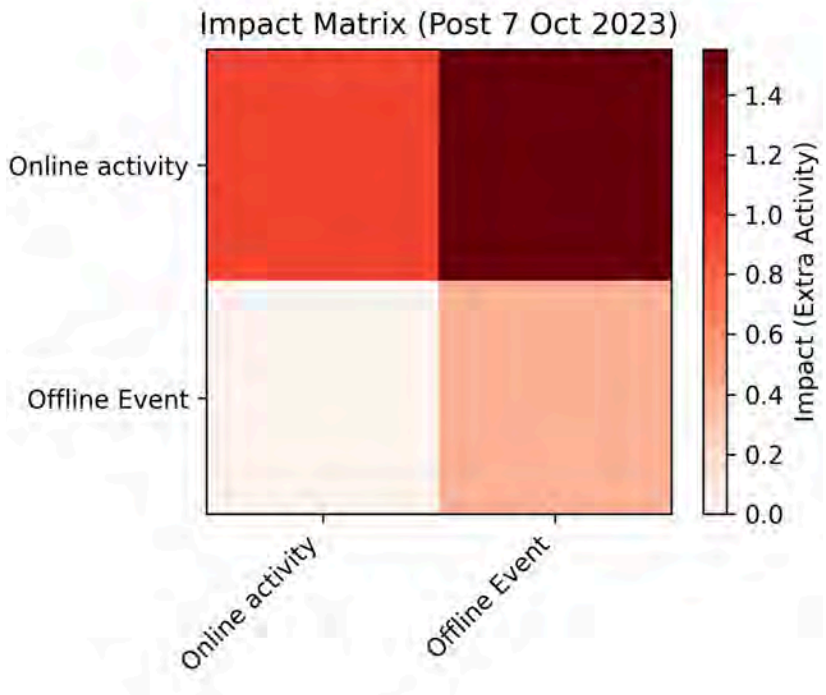


Figure 9. Post-7th October Attacks. Influence Matrix

Diagnostic measures of system-wide dynamics reinforce this finding. In particular, the spectral radius of the system (a measure of how close the overall process is to becoming self-sustaining or explosive) increased from 0.77 to 0.83, approaching the critical threshold of 1.0, where a dynamical process transitions from a stationary regime (i.e. spikes in activity will converge back to 0) to an explosive one (i.e. spikes in activity cumulate exponentially). In other words, after October 7, the ecosystem of anti-Muslim hate was operating close to a tipping point at which activity could theoretically sustain itself indefinitely. In practical terms, the online and offline spheres became tightly interconnected, with real-world Anti-Muslim incidents serving as powerful catalysts for surges in online hate.

Figure 10 shows how the post-attack triggering dynamics unfolded over time, and the pattern differs markedly from the pre-attack period. Rather than a modest early peak followed by rapid decay, the influence profiles for both online-to-online and offline-to-online patterns after October 7 were characterised by noticeable surges. The self-excitatory profile of the online-to-online influence, develops almost simultaneously with the posting of new content, and while the decay is now sharper than in the pre 7th regime, it peaks at 2.5 extra posts in the 2 hours following a tweet, thus signalling a somewhat more vehement and erratic influence of the any single online message. At the same time, the offline-to-online influence becomes substantially stronger. Although the build-up is slower than in the pre-October period, the effect persists much longer, lasting roughly 6 to 12 hours after an incident and plateauing at approximately 2.5 additional online posts. Such influence is recurring as a similar pattern repeats after 15 hours from the offline events. Overall, this suggests that after the attacks, the online ecosystem became far more reactive. A single event, whether a hateful post or a real-world incident, could trigger waves of anti-Muslim content that reverberated for up to a full day. In contrast to the brief, self-limiting bursts observed before October 7, the post-attack environment was characterised by prolonged and more volatile cascading effects, indicating a fundamentally more connected and escalatory relationship between online and offline anti-Muslim activity.

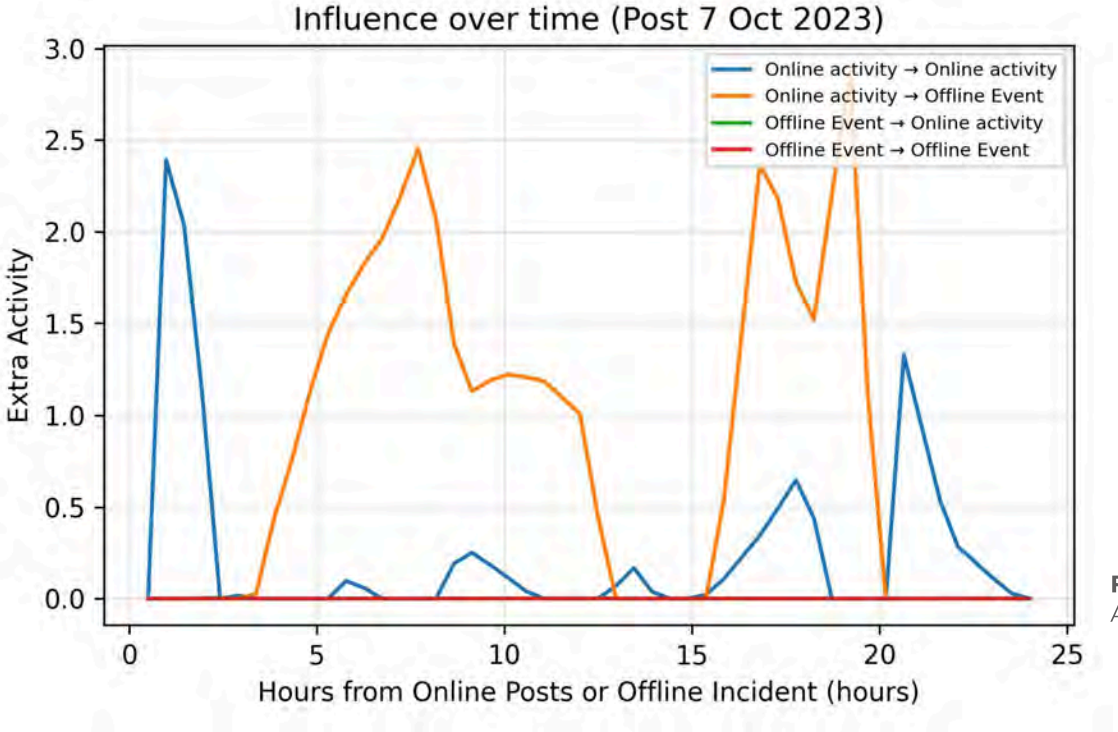


Figure 10. Post-7th October Attacks. Influence over time

Taken together, these results indicate that the relationship between online discourse and offline incidents changed fundamentally after October 7 2023. Whereas the two domains previously operated largely independently, they now form a tightly coupled system in which real-world incidents can trigger sustained waves of online anti-Muslim hate. Previous research (Wiedlitzka et al., 2023) suggests that this dynamic is often mediated by media visibility and public attention surrounding highly salient incidents. Offline attacks and hate incidents can create what researchers describe as "compound retaliation", where extensive news and social media coverage increases the visibility and perceived legitimacy of hostile discourse, providing opportunities and social permission for individuals to express hateful views online. Research has also shown that offline trigger events frequently produce broader surges in online hate speech, including against groups not directly connected to the original event, particularly when events receive sustained political and media attention (Lupu et al, 2023). We suspect a similar process is present in our dataset, where a relatively small number of highly visible incidents may have amplified online anti-Muslim discourse well beyond the incidents themselves, even though most individual offline events would never receive national media coverage.

Figure 10 shows how the post-attack triggering dynamics unfolded over time, and the pattern differs markedly from the pre-attack period. Rather than a modest early peak followed by rapid decay, the influence profiles for both online-to-online and offline-to-online patterns after October 7 were characterised by noticeable surges. The self-excitatory profile of the online-to-online influence, develops

5. NETWORKS OF ANTI-MUSLIM HATE

To understand how anti-Muslim hate circulates online, we constructed and analysed a network of hateful interactions within the dataset.

5.1. Network analyses

In this network, each node represents a user account and each link represents a hateful interaction, such as a post, reply, mention, or share directed at another account that was classified as anti-Muslim hate by our community-informed classifier. To isolate the structural change caused by the Attacks of the October 7 2023, we begin by distinguishing between two time periods: T1 (January 1 2023 to October 6 2023) and T2 (October 7 2023 to March 22 2025). A key concept in this analysis is influence, measured as a user's out-degree. This is the number of hateful interactions they trigger, regardless of whether the message they post is hateful or not, and the number of unique users they reach through those interactions. A consequence of the definition of our measure is that a user will appear highly central if their posts triggered hateful responses from others, and it is not directly associated with the kind of activity done by the user (and also the opposite situation is possible: a very hateful user might be peripheral if not triggering other users). As a result, the network therefore captures structural involvement in hate circulation, not only authorship of hate.

5.2. Findings

Across this analysis, three key findings emerge.

Finding 1. Under the T1-T2 split introduced above, the first finding is that anti-Muslim hate became structurally amplified after October 7 2023. While the volume of hateful interactions increased markedly in T2, the change was not limited to scale. As shown in Figure 11, which compares the distribution of user influence in T1 and T2, the post-October period exhibits a much heavier upper tail. This means that a greater share of users generated high levels of hateful interactions after October 7 2023, and those already active expanded their reach substantially. The figure plots the proportion of users exceeding a given level of hateful out-degree; the curve for T2 sits consistently above that of T1 at higher values, indicating more highly influential actors in the later period. In practical terms, the environment after October 7 2023 enabled both existing and newly active users to scale up their activity and reach wider audiences. For policymakers, this suggests that crisis periods do not merely increase the amount of hateful speech; they can also alter the structural conditions under which hate spreads, making it more efficient and more deeply embedded.

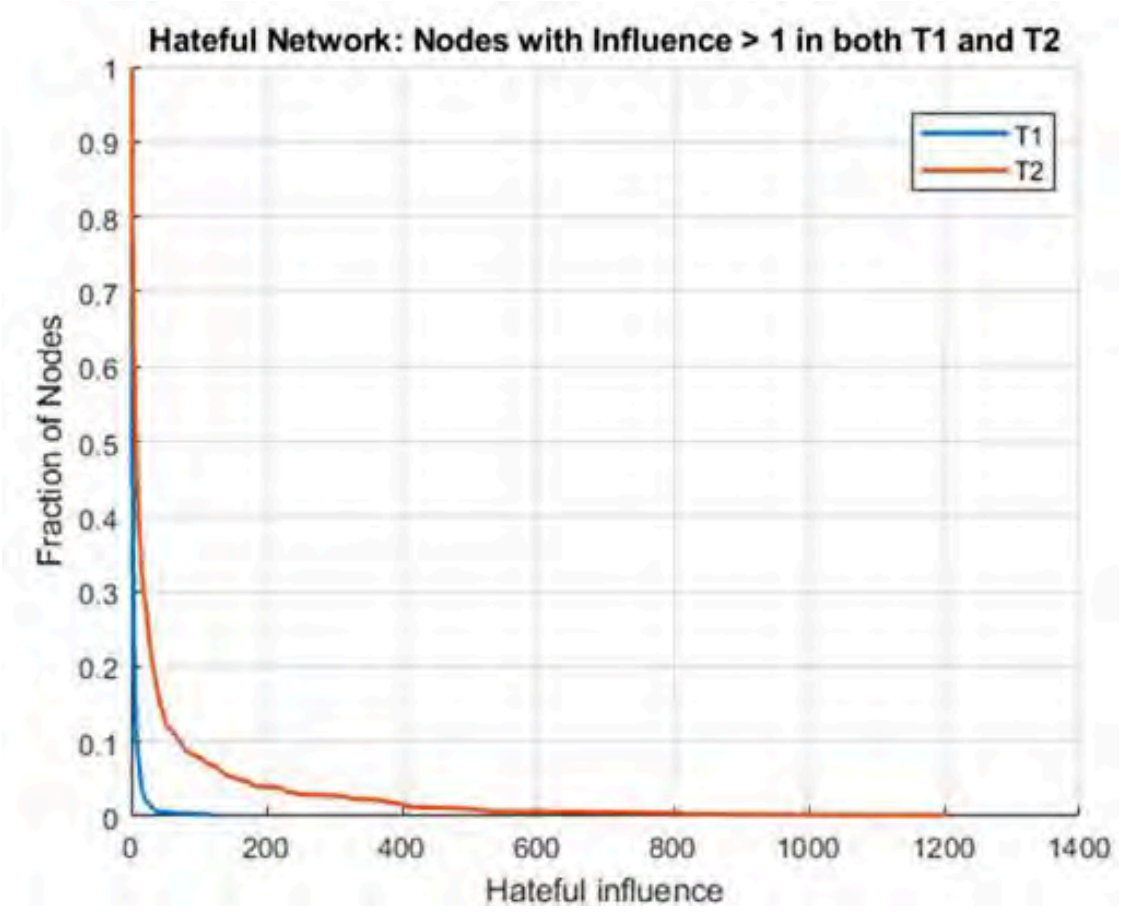


Figure 11. The figure compares the influence distributions of users in the HN across T1 and T2. For each user, out-degree is the number of outgoing hateful edges (?) they produced. The y-axis reports the share of users with out-degree => k. The distribution is truncated to users with out-degree ≥ 1 in both T1 and T2, so differences reflect changes among persistently active hateful accounts, and not entry/exit of accounts.

Finding 2. To introduce our next finding, we adopt a finer temporal structure, splitting the data flows in 5 periods. The first window (Period 1) corresponds to T1 introduced above, whereas the remaining amount of time is split in four evenly distributed periods (Period 2-Period 5). Under this time structure, we find that influence within the hateful network became increasingly concentrated among a small set of users. Figure 12 illustrates this dynamic by tracking, period by period, the share of total hateful influence held by the top 20 most central users in the network in that given period. The blue line shows the proportion of all hateful interactions attributable to these top 20 users, while the orange line shows their share of total anti-Muslim posts. Over time, the influence share of the per-period top 20 users increases (nearly doubling across all periods), while the respective share of total hateful posts does not rise proportionally. Taken together, these results indicate that with the massive inflow of new users and the overall evolution of the hate network (which is a function itself of changing preferences and information flows) centrality becomes less dependent on sheer volume of posting and more dependent on structural position within the network. In other words, a small number of actors are shaping a larger portion of hate flows, thus providing first-hand evidence of what can be described as "a reputational premium" enjoyed by these users. This concentration matters for policy and programs because it suggests that **online hate ecosystems can become increasingly hierarchical. Targeted interventions directed at structurally central accounts may therefore have greater impact than strategies focused solely on high-volume posting.**

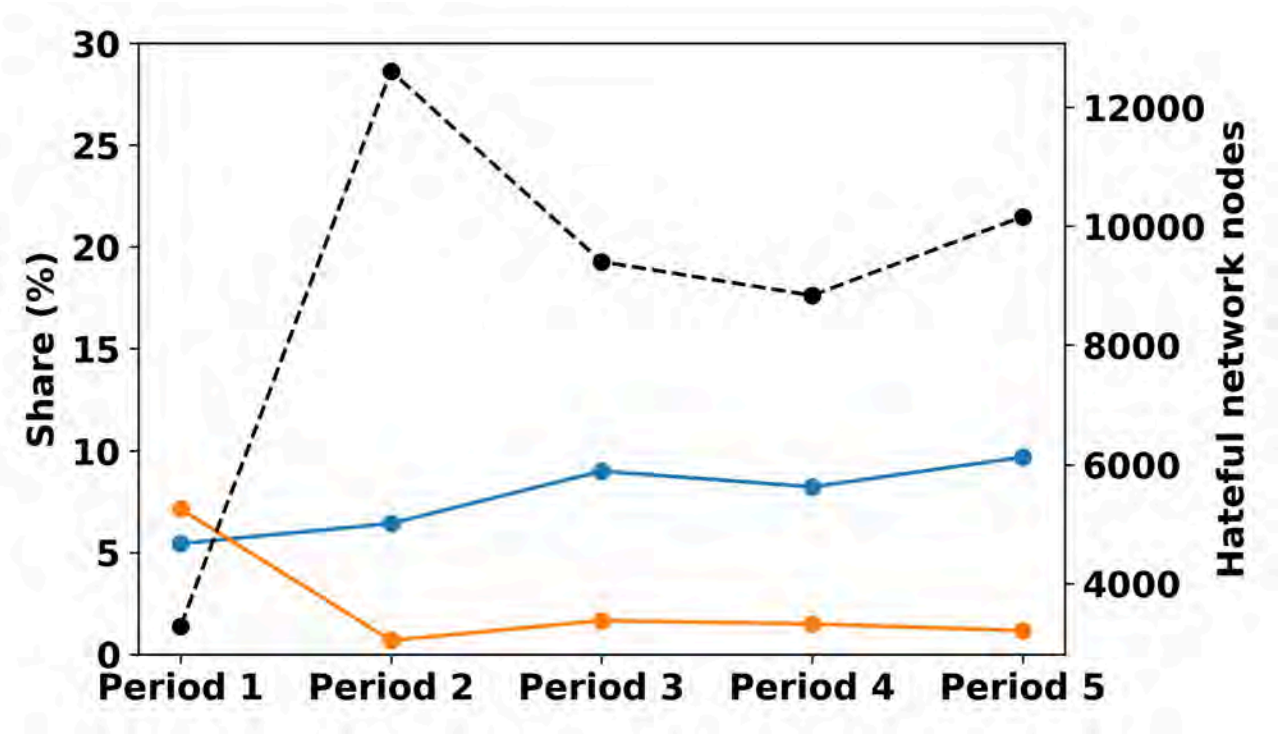


Figure 12. (left axis) each point shows the percentage share held by the top 20 users in the HN computed for that period. The blue line is the share of influence in the HN. The orange line captures the share of anti-Muslim posts or sharing made by those top-20 users over all users. The dashed black line (right axis) shows the number of users in the hateful network for each period. From the figure, we see that in each period, the top 20 users become more central (almost doubling up) as the network becomes overall richer and more active.

Finding 3. The third finding concerns the growth and consolidation of the network on Twitter/X. To achieve this finding, we further refine the temporal structure of the analysis by performing weekly analysis. Figure 13 shows, week by week, the share of users belonging to the largest connected component of the hateful network on each platform. Values closer to one indicate that most users are connected within a single large component, while lower values indicate fragmentation into smaller clusters. On Twitter, the curve follows a U-shaped pattern. Before October 7 2023, the network is relatively cohesive, suggesting a core group of entrenched participants. In the immediate aftermath of October 7 2023the influx of new users fragments the network, reducing the share in the largest component. Over time, however, the network reconsolidates, with a growing proportion of users again belonging to a dominant component. This pattern indicates a three-stage process: initial entrenchment, crisis-driven expansion, and subsequent reassembly around central hubs. For policymakers, this underscores the importance of timing. Interventions during the expansion phase may prevent reconsolidation and long-term normalisation of hateful narratives.

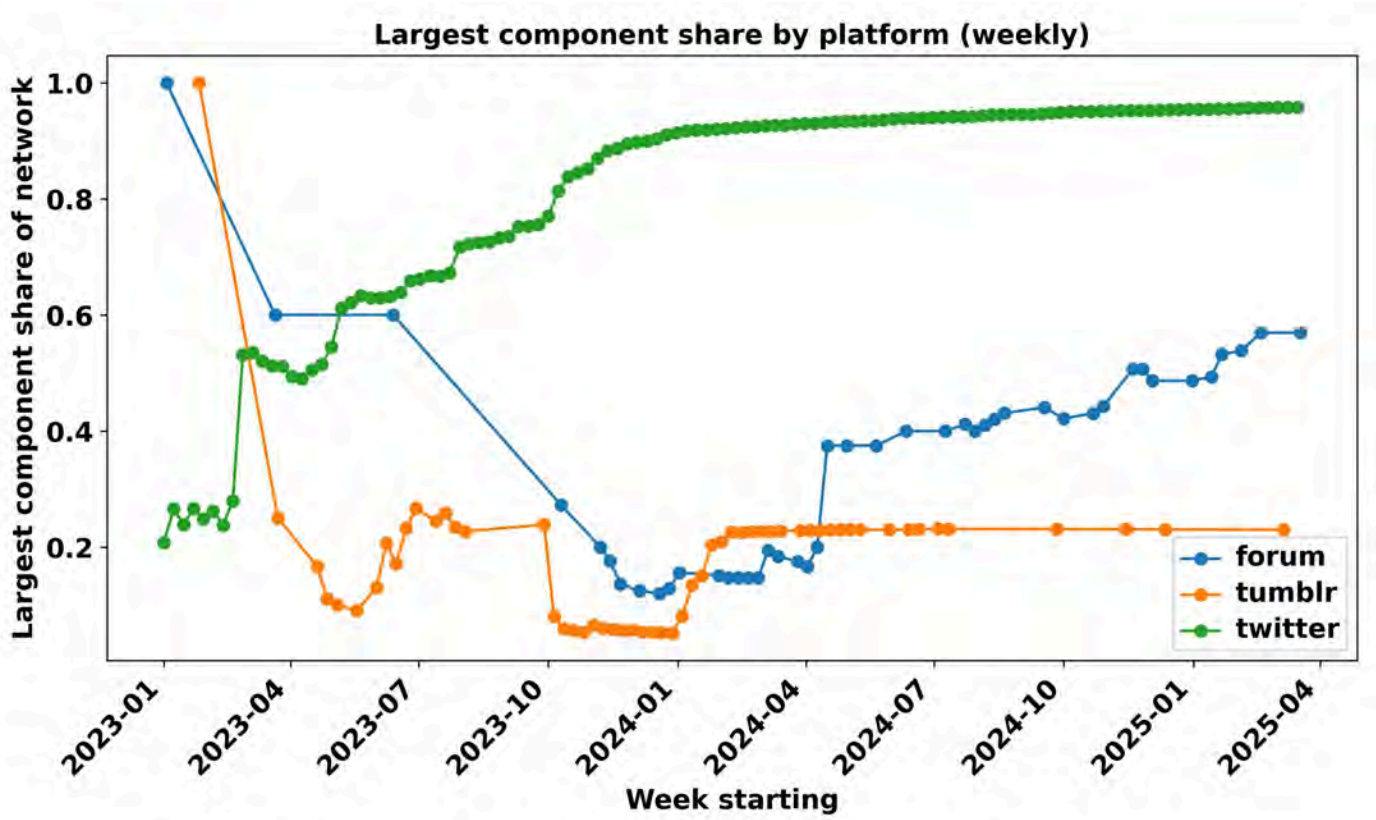


Figure 13. Each line shows, week by week (captured by dots), the share of users that belong to the largest connected component of the hateful network for a given platform. Values closer to 1 indicate a more connected network where most users are linked to other through one single component, while lower values indicate a more fragmented network where users are organised in multiple non-communicating components.

Taken together, these findings show that anti-Muslim hate online after October 7 2023 became larger, more concentrated, more cohesive, and more structurally organised. The network did not simply grow; it reorganised. Influence became more centralised, participation expanded and then reconsolidated, and hate circulated within identifiable communities. Understanding these structural dynamics is essential for designing proportionate and effective responses. Policies and programs that address only content volume risk overlooking the underlying architecture that sustains and amplifies hate.

CONCLUSION

This report shows that anti-Muslim hate in Australia underwent two major transformations during the period examined. The first followed the start of the Gaza war on October 7 2023. The second, substantially larger escalation followed the Bondi terror attack on 14 December 2025. Together, these events reshaped the scale, intensity, and organisation of anti-Muslim discourse in Australian online environments.

Before October 7 2023, in the dataset retrieved by our queries, anti-Muslim hate existed at relatively low and fluctuating levels. Online and offline anti-Muslim activity operated largely as separate systems. Hateful posts triggered further posts in short bursts, but these cascades were self-limiting and quickly dissipated. Offline incidents did not measurably amplify online discourse, and hateful networks remained comparatively smaller and less structurally concentrated.

After October 7 2023, this pattern changed. The baseline level of anti-Muslim hate increased substantially and continued to rise over time. Trigger events generated not only spikes in activity, but prolonged waves of amplification. Offline incidents began to act as catalysts for online surges, while online self-excitation intensified to near-critical levels. Network analysis showed that the ecosystem became more centralised, more cohesive, and more structurally organised. Anti-Muslim hate was increasingly embedded within dominant narratives linking Muslims to violence, terrorism, threat, and collective responsibility.

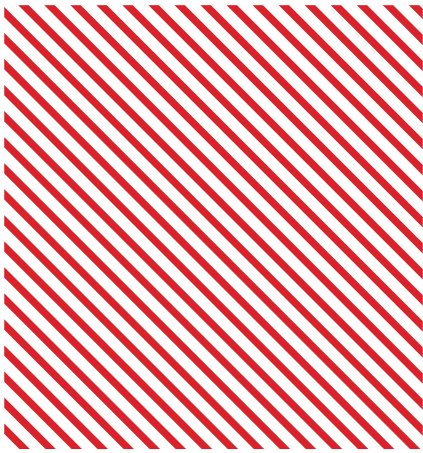
The Bondi terror attack marked a further escalation beyond this already heightened environment. Unlike the post-October 2023 shift, which raised baseline levels into the low hundreds, the post-Bondi period produced sustained daily anti-Muslim hate counts in the high hundreds and thousands. On 15 December 2025, anti-Muslim hate reached 7,786 posts in a single day, while toxicity targeting Muslims and identity attacks targeting Muslims also rose into the thousands. This represented not simply a continuation of earlier trends, but a qualitative intensification in the scale and aggressiveness of online hostility.

Importantly, the post-Bondi escalation demonstrates how rapidly anti-Muslim discourse can expand following highly salient domestic incidents. The findings suggest that an already elevated online hate ecosystem can become massively amplified when local acts of violence are interpreted through existing geopolitical and identity-based narratives. In these periods, hostility shifts beyond ordinary political disagreement into widespread identity-directed abuse, collective blame, and dehumanising discourse targeting Muslims as a group.

Anti-Muslim hate and anti-Palestinian hate also followed different trajectories throughout the study period. Anti-Palestinian hate was more volatile and event-driven, with sharp spikes linked to specific incidents and phases of the conflict. This distinction is critical for policy and program development. Episodic volatility can be addressed through event-based response mechanisms. Structural escalation requires deeper interventions targeting the underlying online ecosystem.

The network evidence further confirms that hate circulates through identifiable clusters and is increasingly shaped by a relatively small number of structurally central actors. Over time, influence became more concentrated, hateful interactions became more interconnected, and networks reconsolidated following periods of crisis-driven expansion. These dynamics suggest that online hate ecosystems are not random or diffuse, but organised around identifiable pathways of amplification and influence.

Taken together, the findings show that online anti-Muslim hate in Australia became more reactive, more interconnected with offline events, more structurally concentrated, and more systemically embedded over time. The post-October 2023 environment established a persistently elevated baseline of hostility, while the Bondi attack demonstrated the capacity for that environment to escalate rapidly into mass online mobilisation. Addressing these dynamics requires not only moderation of individual content, but coordinated strategies capable of responding to both long-term structural escalation and acute crisis-driven surges in hate.



REFERENCES

Carland, S. Alziyadat, N., Vergani, M. & O'Brien. K. (2025) Islamophobia in Australia Report V, Sydney: Islamophobia Register Australia, available here: https://islamophobia.com.au/wp-content/uploads/2025/03/Islamophobia-in-Australia-Report-5.pdf

Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, N. J., Goldberg, B., & Johnson, N. F. (2023). Offline events and online hate. PLoS one, 18(1), e0278511.

Wiedlitzka, S., Prati, G., Brown, R., Smith, J., & Walters, M. A. (2023). Hate in word and deed: the temporal association between online and offline islamophobia. Journal of quantitative criminology, 39(1), 75-96.

AUTHORS BIO

Matteo Vergani

Matteo is director of the Tackling Hate Lab and Associate Professor in Sociology at Deakin University and specialises in radicalisation and hate crime, publishing in leading international journals and securing large research grants. He collaborates with numerous institutions and government agencies in Australia and Canada. His research advances the systematic consolidation of knowledge in hate and extremism studies through large-scale systematic reviews and the development of rigorous measurement tools of online and online hate and radicalisation. His research programme fosters multidisciplinary collaboration across social sciences, data science, econometrics and engineering, leveraging advanced technologies for analysing digital archives and social media big data.

Susan Carland

Susan is Senior Lecturer in Sociology at Monash University. She is a sociologist of religion, specialising in the intersection of gender, religion, discrimination, and inclusion. She is an ARC DECRA and Churchill Fellow. Her research focuses on Muslims, sexism, and Islamophobia. Dr Carland has led national and international studies employing quantitative and qualitative research methods exploring the connection between gender, religion, and discrimination. She has spoken about her research to the UN in Geneva, Chatham House in London, to DFAT, State & Federal government ministers, and across media.

Andrea Giovannetti

Dr Andrea Giovannetti is Co-Director of the Tackling Hate Lab, Assistant Professor of Economics at the Australian Catholic University and a member of the Violence Research Centre at the Institute of Criminology of the University of Cambridge, where he previously held a Marie Curie Postdoctoral Fellowship. His research on organised crime, contemporary extremism and social cohesion combines machine-based econometrics with advanced computational methods in network theory to support policymakers and security agencies on a large spectrum of inter-connected issues. Andrea's collaborations with public stakeholders on complex social threats include London Metropolitan Police, Merseyside Police, Home Office and Home Affair.

Stephanie Zi Xin Ng

Stephanie holds a PhD in Engineering from Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI). Her research focuses on advancing computational methods that leverage Natural Language Processing (NLP) and Large Language Models (LLMs) to address pressing social challenges. Her work spans a variety of topics, including the policy framing of contemporary social issues, stance detection in parliamentary debates, and agent-based modelling of inoculation strategies against online extremism. Her recent work explores video and image processing, including multimodal analysis of the privacy paradox in short videos and Computer Vision (CV) for hazard detection.

Muhammad Sakib Khan Inan

Sakib holds a PhD in Data Science from Deakin University. He is a data science and artificial intelligence researcher specialising in advanced AI methods and large language models. His work focuses on applying AI techniques to solve complex interdisciplinary research problems across domains such as natural language processing, computer vision, and time-series analysis. Sakib has worked on projects involving IoT sensor data, biomedical research, and geotechnical engineering, with an emphasis on developing innovative, data-driven solutions to real-world social and technological challenges. His research combines methodological innovation with practical applications aimed at improving the reliability and impact of AI systems.

Kewen Liao

Kewen is co-director of the Tackling Hate Lab and Associate Professor in Data Analytics and Machine Learning at Deakin University. Kewen is an accomplished algorithms researcher with a PhD in Computer Science. His expertise spans data science, machine learning, and theoretical computer science, with a focus on algorithm design and analysis, clustering and graph mining, time series and streaming analytics, and image and text data analysis. He is a Chief Investigator on an Australian Research Council Discovery Project and previously led a Defence Next Generation Technologies Fund Project. He also holds a US patent with Canon Inc. for a novel object-matching method in computer vision. Kewen is committed to cross-disciplinary research, leveraging data science and AI to drive significant societal and economic impact. He also holds key leadership and professional roles in leading national and international data science and AI conferences.

Huu Phuc Hong (Felix)

Felix is a data scientist and research assistant working at the intersection of computational social science and natural language processing. He was a highest-achieving graduate of Deakin's Master of Data Science (across both the Standard and Professional programs), and his work focuses on building reliable, transparent measurement tools for analysing online harms at scale. Felix develops hate speech detection approaches that combine large language models with smaller supervised classifiers, and creates end-to-end tooling to collect, organise, and monitor social media data. He also applies statistical validation methods and graph-based network analysis to map communities, narratives, and information flows, with an emphasis on clear, defensible evaluation.

Yinsong Cheng

Yinsong is an AI researcher and software engineer with a PhD in Computer Science and Electrical Engineering from Deakin University. His research focuses on trustworthy machine learning, large language model alignment, hate speech detection, explainable AI, uncertainty-aware learning, and scalable data analytics. He has developed LLM alignment methods for detecting implicit and context-dependent harmful language, with a particular focus on the Direct Preference Optimization method family. He also has experience across computer vision, time-series forecasting, probabilistic deep learning, and real-time data analytics. His work aims to build reliable, scalable, and interpretable AI systems for safety-critical and socially impactful applications.

Haily Tran

Haily is a mixed-methods researcher in social psychology, focussing on psychological drivers of online radicalisation, violence, and hate-based ideologies. Her PhD examined the role of masculinity and victimhood in the mobilisation of Australian men toward far-right extremism. Experienced in experimental research designs and evidence-based practice, Haily has also contributed to multiple projects in hate crime prevention and countering violent extremism (P/CVE). She currently serves as HDR/ECR coordinator for the AVERT Research Network and is a member of the Australian Psychological Society's College of Forensic Psychologists.

Amiee Taylor

Amiee is an Honours researcher at Deakin University whose work focuses on radicalisation, far-right extremism, and new religious movements. Her research examines how Christian nationalism, premillennialist beliefs, and conspiracy narratives contribute to religiously motivated extremism and political violence. Her thesis investigates the geographical and social conditions shaping Neo-Nazi recruitment and ideological commitment, with particular attention to the structural and cultural factors sustaining extremist movements. Alongside her academic work, Amiee contributes to research projects exploring the use of artificial intelligence and machine learning tools for detecting and analysing online hate and extremist content.

Dan Goodhardt

Dan is a researcher at Deakin University, where he has collaborated for over a decade on studies of hate crime and violent extremism in Australia and Southeast Asia. He has extensive experience in collecting anti-Jewish hate data through the Jewish Community's Community Security Group and currently works as a senior manager with Victoria Police. As co-convenor of the Practitioners Working Group on Tackling Hate in Victoria at the Centre for Resilience and Inclusive Societies, he connects security practitioners with researchers to advance efforts in countering hate and extremism.

www.tacklinghate.org