Social Hatred: Efficient Multimodal Detection of Hatemongers

Tom Marzea
Ben Gurion University
tommarz@post.bgu.ac.il

Abraham Israeli
University of Michigan
isabrah@umich.edu

Oren Tsur
Ben Gurion University
orentsur@bgu.ac.il

Abstract

Automatic detection of online hate speech serves as a crucial step in the detoxification of the online discourse. Moreover, accurate classification can promote a better understanding of the proliferation of hate as a social phenomenon. While most prior work focus on the detection of hateful *utterances*, we argue that focusing on the *user* level is as important, albeit challenging. In this paper we consider a multimodal aggregative approach for the detection of hate-mongers, taking into account the potentially hateful texts, user activity, and the user network. Evaluating our method on three unique datasets X (Twitter), Gab, and Parler we show that processing a user's texts in her social context significantly improves the detection of hate mongers, compared to previously used text and graph-based methods. We offer comprehensive set of results obtained in different experimental settings as well as qualitative analysis of illustrative cases. Our method can be used to improve the classification of coded messages, dog-whistling, and racial gas-lighting, as well as to inform intervention measures. Moreover, we demonstrate that our multimodal approach performs well across very different content platforms and over large datasets and networks.

Offensive content warning: The illustrative examples throughout the manuscript, and specifically in Table 1 may be offensive to some readers.

1 Introduction

The rising popularity of social platforms coincides with the proliferation of online hate speech and a surge in hateful content targeting minorities (Waseem and Hovy, 2016; Laub, 2019). Accordingly, there is a growing body of research on the appearance and magnitude of hate speech on social media (Knuttila, 2011; Chandrasekharan et al., 2017; Zannettou et al., 2018), and on hate speech detection (Saleem et al., 2017; Waseem and Hovy, 2016; Davidson et al., 2017). Online hate is often linked to real-world violence (Munn, 2019; Malevich and Robertso, 2019; Thomas, 2019; McIlroy-Young and Anderson, 2019; Mathew et al., 2019; ADL, 2023).

Hate is often promoted by communities, rather than by isolated individuals. Shifting the focus from the utterance (post) level to the user and community level can be beneficial: it may provide a better understanding of hate group dynamics; it could improve detection of coded language (dog whistling, gas-lighting) at the post level; and it could inform intervention strategies[1] (Thomas and Wahedi, 2023).

Challenges in utterance-level hate speech classification are well-documented (Talat et al., 2018; Fortuna et al., 2022; Velankar et al., 2022; Kim et al., 2024), and even state-of-the-art LLMs show shortcomings in nuanced social tasks like hate detection (Ziems et al., 2024).

Table 1 provides examples illustrating these challenges. While explicit hate (utterance #1) is often detected, coded language and symbols (#2, the (((echo))) symbol[2]) or white supremacist slogans (#3, the '14 words'[3]) may require external knowledge or specific training data. Allusions (#4-7) and twisted slogans (#7-8) further complicate detection for both models and humans. A detailed analysis of these examples is provided in Appendix A.

Contextual information and aggregation at the user level can help mitigating challenges posed by individual posts. Multiple implicit posts from one user, or association with other hateful users/posts, can reinforce weak signals. However, aggregation requires careful consideration of prediction confi-

[1]Intervention alternatives depend on legal regulation and platform policy and range from the removal of a specific post, the flagging of a specific post, to the temporal or indefinite suspension of an account.

[2]www.adl.org/resources/hate-symbol/echo

[3]www.adl.org/resources/hate-symbol/14-words

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 284–298

	Text	DHB	JS	FTDB	GPT4	Gemini 1.5Pro
1	“Muslim terrorists did this and obama brought them here to run our government. Gotta get rid of Obammy, Omar and all these other f**king Muslims”	.57	.9	.98	✓	✓
2	“is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not.”	.22	.35	.92	✓?	✓
3	“We must secure the existence of our people and a future for white children!”	.03	.36	.02	✓	✓?
4	“If white people are going to continue existing - we must support our race”	.06	.4	.69	✓?	✓?
5	“When the youngest white generation becomes adult they will be 5% of the world population, down from 1/3 in 1900”	.06	.25	.01	✗	✓?
6	“Did you know that most dog breeds were bred by whites? We’re awesome and they’re awesome!”	.2	.25	.65	✗	✗
7	”White lives matter!”	.02	.24	.02	✗	✗?
8	“Blue lives matter!”	.02	.01	.01	✗	✗?

Table 1: Examples of hate-promoting texts. DHB: deHateBERT (Aluru et al., 2020); JS: Google’s Jigsaw; FTDB: DistillBERT fine-tuned on our datasets; GPT-4 & Gemini 1.5 Pro predictions (see Appendix B). ‘?’ indicates nuanced prediction.

dence and context, e.g., how many implicit posts equate to one explicit post? We propose a principled way to combine predictions and modalities for accurate user-level classification.

We explore three aggregation approaches: (i) binary weights with a fixed threshold, (ii) relational aggregation using social context, and (iii) distributional aggregation based on confidence levels. We combine these into a multimodal model.

Contribution The contribution of this work is threefold:

1. We propose a robust and efficient multimodal aggregative approach for detecting hate-mongers.
2. We demonstrate the benefits of contextual aggregation over three unique datasets (Twitter, Gab, Parler).
3. We share a novel annotated dataset of Parler hate users.

2 Related Work

Hate speech detection methods and challenges are surveyed by Alkomah and Ma (2022), ElSherief et al. (2021); Velankar et al. (2022); Fortuna et al. (2022). Issues include subjectivity (Khurana et al., 2022), limitations of transfer learning (Israeli and Tsur, 2022), and the need for robust evaluation (Röttger et al., 2021, 2022). Implicit hate taxonomies exist but the datasets are often unsuitable for user-level analysis (ElSherief et al., 2021).

Research is shifting towards the user level. Early work explored demographics (Waseem and Hovy, 2016) and account meta-features (Ribeiro et al., 2018b). Arviv et al. (2021) used a multi-modal approach combining predictions from a user’s posts, followers, and followees.

Textual signals combined with network diffusion models have been used to propagate initial hate labels within communities of users (Ribeiro et al., 2018a; Israeli and Tsur, 2022). Graph Neural Networks (GNNs) incorporate text and network structure (Li et al., 2021; Miao et al., 2022; Das et al., 2021; Utku et al., 2024). GNNs were used for detection of hate on post-level (Miao et al., 2022) as well as for classification on the user level (Das et al., 2021; Utku et al., 2024). Early detection of potential spreaders using author profiling techniques has also been explored (Irani et al., 2021).

Recent work explores using LLM rationales for interpretability (Nirmal et al., 2024), though fine-tuned models often outperform LLMs on socially nuanced tasks like hate detection (Ziems et al., 2024).

3 Multimodal Aggregative Approaches

3.1 Aggregative Approaches

Our multimodal framework makes relies on two modalities: the textual modality and the social modality. The user-level classification is achieved by aggregating predictions for multiple utterances made by a specific user, informed by the user social context. We therefore describe the classification of of a single utterance, before offering a number of ways to aggregate predictions in order to achieve accurate classification at the user level.

Utterance-level Classification (C^T) The foundation is single-utterance classification. Any model outputting a probability $\theta(t)$ can be used. We binarize using a threshold τ^T:

$$C^T(t) = \begin{cases} 1 & \theta(t) \geq \tau^T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We use a fine-tuned DistilBERT (Sanh et al., 2019a) following Israeli and Tsur (2022).

User-level Classification (C^U) Detecting hateful users involves aggregating signals from their posts (T^u). The generic form of C^U uses an aggregation function $\Theta(u)$ and a user threshold τ^U:

$$C^U(u) = \begin{cases} 1 & \Theta(u) \geq \tau^U \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Naive Aggregation with Fixed Threshold (Θ_F) A simple approach counts hateful posts by user u:

$$\Theta_F(u) = \sum_{t \in T^u} C^T(t) \quad (3)$$

The threshold τ^U controls sensitivity (e.g., $\tau^U = 1$ for zero-tolerance[4], higher values for repeated offenders). We denote this classification C_F.

We propose more robust aggregations considering nuance and social context, training Θ (e.g., logistic regression) on user features derived from these aggregations.

Multimodal Relational Aggregation (Θ_R) User identity is shaped by social ties (McMillan and Chavis, 1986; Wellman and Gulia, 1999), at time including affiliation with online hate groups (Gordon, 2017; Govers et al., 2023). We incorporate the hate level in a user u's ego network (followers $\overleftarrow{u}$, followees $\overrightarrow{u}$) in graph $G(V, E)$. $\Theta_R(u)$ combines u's own post count ($C_F(u)$) with the proportion of hateful neighbors:

$$\Theta_R(u) = \alpha \cdot C_F(u) + \beta \cdot \frac{1}{|\overleftarrow{u}|} \sum_{v \in \overleftarrow{u}} C_F(v) + \gamma \cdot \frac{1}{|\overrightarrow{u}|} \sum_{v \in \overrightarrow{u}} C_F(v) \quad (4)$$

Weights α, β, γ are learned.

[4]Often too harsh: remember that in the standard case $\tau^T = 0.5$ thus a user will be labeled a hate-monger even if $\theta(t) = 0.51$ for one of her posts and $\theta(t') < 0.5 \quad \forall t' \in T^u/\{t\}$.

Distributional Aggregation (Θ_D) This approach addresses variations in hate intensity (commitment, implicitness). Instead of just counting hateful posts, we consider the distribution of hate scores $\theta(t)$ for user u's posts T^u. We represent this distribution as a k-dimensional vector using bins or quantiles. For bin-based, the [0,1] range is split into k bins; the vector counts posts falling into each bin. For quantile-based, the range $[\min(\theta(t)), \max(\theta(t))]$ for $t \in T^u$ is split into k bins. We optimize:

$$\Theta_D(u) = \sum_{i=1}^{k} w_i \cdot \sigma(B_i(u)) \quad (5)$$

where $B_i(u)$ is the count for bin i, σ is softmax, and w_i are learned weights.

Combined Multimodal Aggregation (Θ) We combine relational and distributional features. Using Θ_D^b (bin-based) and Θ_D^q (quantile-based):

$$\Theta(u) = \Theta_D^b(u) + \Theta_D^q(u) + \Theta_R(u) \quad (6)$$

The final classifier learns weights for all components.

3.2 Socially-aware Baselines

We compare the multimodal aggregations against a number of strong baselines that leverage the network structure.

DeGroot's Diffusion A belief propagation model shown effective for hate detection (Ribeiro et al., 2018a; Israeli and Tsur, 2022).

Graph Neural Networks (GNNs) Methods explored by Das et al. (2021) and others for hateful user detection:

1. **Node2Vec** (Grover and Leskovec, 2016): Embeddings from biased random walks.
2. **GCN** (Kipf and Welling, 2017): Convolutional layers aggregating neighbor features.
3. **GAT** (Velickovic et al., 2018): Attention mechanism weighting neighbor importance.
4. **GraphSAGE** (Hamilton et al., 2017): Inductive learning via sampling and aggregating neighbor features.
5. **AGNN** (Thekumparampil et al., 2018): Attention-based propagation learning adaptive local summaries.

AGNN and GCN were specifically used for hate user detection by (Das et al., 2021) and (Utku et al., 2024).

4 Datasets and Annotation

We use three datasets: Twitter-Echo, Gab, and Parler. Descriptive statistics are provided in Tables 2 and 3.

Echo (Twitter) The triple parentheses, or triple brackets, also known as the (((echo))), is an antisemitic symbol that is used to highlight the names of individuals of Jewish background (e.g., actress and comedian Amy Schumer, see utterance #2 in Table 1), organizations owned by Jewish people (e.g., Ben & Jerry's), or organizations accused of promoting "Jewish globalist values" (e.g., the International Monetary Fund). The Echo dataset curated by Arviv et al. (2021) contains over 18M English tweets posted by ∼7K echo users between May and June 2016. Annotations are provided at the tweet and the user level. An important feature of this dataset is that all users have utterances containing the echo symbol, although some users use it in a non-hateful manner, e.g., to symbolize a hug. This ambiguous nature of the symbol makes hate detection challenging.

Gab Positioning itself as putting "people and free speech first", Gab welcome users suspended from other social networks. The platform permits pornographic and obscene content, as long as it is labeled *NSFW* ('not safe for work'). Posts (called *gabs*) are limited to 300 characters, and users can repost, quote, or reply to previously created gabs. The dataset (22M posts, 337K users) was collected by (Zannettou et al., 2018) from Aug 2016 - Jan 2018. An annotated subset shared by Arviv and Tsur (2021) includes 60K posts and 1K users.

Parler Alluding to the French verb 'to speak', Parler was launched on August 2018. On April 2023, the platform was acquired by Starboard and was taken offline to "undergo a strategic assessment" (Starboard announcement on Parler's landing page https://parler.com/, accessed: 5/8/2023). The platform was relaunched in February 2024, announcing it is "breaking free from the constraints of conventional platforms".

The platform branded itself as "The World's Town Square" a place to *"Speak freely and express yourself openly, without fear of being "deplatformed" for your views"*.[5] Parler's official guidelines[6] explicitly allowed "trolling" and "not-safe-for-work" (NSFW) content, including only three "principles" prohibiting "unlawful acts", citing "Obvious examples include: child sexual abuse material, content posted by or on behalf of terrorist organizations, intellectual property theft". Parler allows posts of up to 1000 characters, compared to 280/300 allowed by Twitter/Gab.

A dataset of Parler messages was introduced by Aliapoulios et al. (2021) and a subset of 10K *posts* was annotated by Israeli and Tsur (2022). As part of this research, we share the first annotated dataset of Parler *users*.

Annotation of Parler Users Annotation was done by 94 senior year Data Science B.Sc students for bonus course credit. Annotators were introduced to Parler and were given explicit instructions about the annotation task. The annotation process involved rating each account on a 1–5 scale (non-hateful – extremely hateful). We ensured that each user was annotated by three annotators. The full annotation guidelines and further details regarding the sampling method and the annotation protocol are available in Appendix F.

5 Results and Analysis

5.1 Experimental Settings

Utterance-level: We fine-tuned DistilBERT (Sanh et al., 2019b) per dataset (80/20 train/val split, batch 32, early stopping). User-level: 5-fold CV on the largest connected component (LCC) of each dataset (see details in Table 3). GNN settings follow Das et al. (2021). DeGroot settings follow Israeli and Tsur (2022). Node2Vec used implementation by Liu and Krishnan (2021). Further details about models' configurations, hyperparameters, and application can be found in Appendix C.

5.2 Results and Analysis

Utterance Level Prediction Vanilla DistilBERT performance (Table 4) reveal the challenge in the detection of hate on the utterance level, especially on Gab (F1=0.29), motivating user-level aggregation. We observe that mean hate score distributions differ across platforms (Figure 1), revalidating the need for a generic yet robust approach.

[5] Parler branding on its landing page.
[6] https://parler.com/documents/guidelines.pdf

Dataset	Source	Raw Data #Posts	Raw Data #Users	Annotated Data (Overall) #Posts	Annotated Data (Overall) % Hate	Annotated Data (Overall) #Users	Annotated Data (Overall) % Hate
Echo (Twitter)	Arviv et al. (2021)	18M	7.1K	4.6K	8.2%	1K	15.4%
Gab	Arviv and Tsur (2021)	22M	337K	5K	5.1%	1K	24.8%
Parler	This work (users)	$183M^{\dagger}$	$4.1M^{\dagger}$	$8.3K^{\ddagger}$	$32.9\%^{\ddagger}$	890	25.4%

Table 2: Overview of datasets and annotated subsets. † Raw data from (Aliapoulios et al., 2021). ‡ Post annotations from (Israeli and Tsur, 2022). User level annotations are introduced in this work.

Dataset	Largest Connected Component (LCC) Stats #Posts	#Users	#Edges	Clust. Coeff.	Gamma (γ)	Annotated in LCC #Users	% Hate
Echo (Twitter)	9.8M	3.7K	20.7K	0.190	2.8	532	26.1%
Gab	19.3M	51.2K	2.47M	0.402	4.1	982	24.5%
Parler	60.7M	643K	11.4M	0.224	2.1	881	25.2%

Clustering Coefficient and Optimal Gamma (scale-free fit) calculated on undirected graph.

Table 3: Statistics for the largest (weakly) connected component used in experiments.

Dataset	Precision	Recall	F1 Score
Echo	0.412	0.803	0.545
Gab	0.206	0.547	0.299
Parler	0.632	0.818	0.713

Table 4: Utterance-level DistilBERT performance.

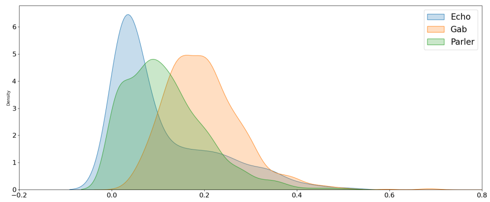

Figure 1: Density of mean utterance scores per user.

User level results Detailed results are presented in Table 5. Our Multimodal Relational approach achieves top F1 performance on Echo and Parler. Our combined Distributional (bins+quantiles) approach achieves the best results on Gab, with our Combined Multimodal approach very close third. All proposed aggregative methods (Relational, Distributional variants, Combined) consistently perform competitively, often outperforming the GNN and diffusion models.

The optimal weights for the Relational method (α, β, γ in Eq. 4) varied across datasets (more details are available in Appendix C), highlighting the influence of platform-specific network structure and dynamics, also reflected in Figure 1.

Structure-based Analysis The F-scores achieved by different aggregation methods show that our approaches consistently outperform baseline algorithms. The multimodal approach performs best on the Echo and Parler datasets, while results differ on the Gab dataset (Table 5), suggesting that Gab's size, structure, and unique characteristics influence aggregation effectiveness.

Table 3 highlights these differences. The Echo network, the smallest of the three (3.7K users, 20.7K edges), has a clustering coefficient of 0.19 and a degree exponent (γ) of 2.8, indicating a classic scale-free structure. This topology supports the multimodal approach by effectively leveraging diverse signals. Parler, the largest network (643K users, 11.4M edges), also exhibits a scale-free structure ($\gamma = 2.14$) and a higher clustering coefficient (0.224), allowing it to benefit from similar aggregation strategies. In contrast, Gab (51.2K users, 2.47M edges) has a high clustering coefficient (0.402) and a γ of 4.06, indicative of a random-like network structure (Amaral et al., 2000; Goh, 2001). This random-like structure hampers the performance of the multimodal aggregation. However, the Combined Multimodal Aggregation model mitigates this by assigning weights to sub-models, nearly matching the best performerming model.

Breaking down the combined model into its components (Relational, Distributional-bins, Distributional-quantiles) we observe that these methods are competitive, outperform the baselines, and capture the diverse facets of user behavior and network structure. For instance, the Relational method's performance varies with platform-specific parameters like α, β, and γ (see further details in Appendix C).

These results underscore the importance of properly addressing the network structure, user dynamics, and platform norms in determining aggregation method effectiveness: distinct structural features, such as clustering coefficients and degree distributions, impact performance, highlighting the need

	Method	Precision	Recall	F1	PRC AUC
Echo	DeGroot's Diffusion	0.472 ± 0.389	0.255 ± 0.261	0.320 ± 0.310	0.319 ± 0.0
	Node2Vec	0.764 ± 0.112	0.819 ± 0.148	0.788 ± 0.121	0.806 ± 0.092
	GCN	0.597 ± 0.192	0.950 ± 0.041	0.717 ± 0.127	0.819 ± 0.119
	GraphSAGE	0.661 ± 0.097	0.921 ± 0.029	0.767 ± 0.076	0.845 ± 0.113
	GAT	0.755 ± 0.048	0.914 ± 0.054	$0.825 \pm 0.025^{\dagger}$	0.874 ± 0.074
	AGNN	0.722 ± 0.108	0.914 ± 0.070	0.803 ± 0.084	0.851 ± 0.117
	Fixed-Threshold	0.654 ± 0.063	0.627 ± 0.095	0.633 ± 0.040	0.678 ± 0.052
	Multimodal Relational Aggregation	0.789 ± 0.091	0.878 ± 0.054	$\mathbf{0.826 \pm 0.046}$	0.844 ± 0.076
	Distributional (bins)	0.772 ± 0.045	0.871 ± 0.058	0.817 ± 0.042	0.874 ± 0.040
	Distributional (quantiles)	0.747 ± 0.064	0.899 ± 0.047	0.815 ± 0.053	0.878 ± 0.037
	Distributional (bins+quantiles)	0.757 ± 0.058	0.885 ± 0.052	0.815 ± 0.049	0.887 ± 0.036
	Multimodal (Relational + Bins + Quantiles)	0.766 ± 0.045	0.892 ± 0.059	$0.824 \pm 0.046^{\ddagger}$	0.894 ± 0.024
Gab	DeGroot's Diffusion	0.314 ± 0.001	0.777 ± 0.0	0.447 ± 0.001	0.240 ± 0.005
	Node2Vec	0.404 ± 0.072	0.335 ± 0.085	0.364 ± 0.075	0.386 ± 0.050
	GCN	0.296 ± 0.045	0.678 ± 0.285	0.387 ± 0.034	0.344 ± 0.060
	GraphSAGE	0.390 ± 0.086	0.469 ± 0.263	0.387 ± 0.075	0.397 ± 0.075
	GAT	0.218 ± 0.125	0.608 ± 0.389	0.316 ± 0.182	0.282 ± 0.063
	AGNN	0.337 ± 0.059	0.539 ± 0.197	0.403 ± 0.081	0.389 ± 0.029
	Fixed-Threshold	0.466 ± 0.072	0.335 ± 0.061	0.388 ± 0.061	0.468 ± 0.093
	Multimodal Relational Aggregation	0.408 ± 0.067	0.429 ± 0.094	0.414 ± 0.069	0.538 ± 0.070
	Distributional (bins)	0.461 ± 0.034	0.649 ± 0.044	$0.538 \pm 0.024^{\dagger}$	0.521 ± 0.048
	Distributional (quantiles)	0.429 ± 0.027	0.702 ± 0.056	0.532 ± 0.033	0.533 ± 0.068
	Distributional (bins+quantiles)	0.435 ± 0.026	0.714 ± 0.043	$\mathbf{0.540 \pm 0.029}$	0.524 ± 0.064
	Multimodal (Relational + Bins + Quantiles)	0.429 ± 0.021	0.718 ± 0.040	$0.537 \pm 0.025^{\ddagger}$	0.538 ± 0.070
Parler	DeGroot's Diffusion	0.395 ± 0.221	0.441 ± 0.247	0.417 ± 0.233	0.414 ± 0.001
	Node2Vec	0.408 ± 0.055	0.392 ± 0.058	0.400 ± 0.056	0.434 ± 0.039
	GCN	0.284 ± 0.054	0.760 ± 0.404	0.348 ± 0.130	0.410 ± 0.184
	GraphSAGE	0.309 ± 0.092	0.649 ± 0.189	0.394 ± 0.028	0.343 ± 0.083
	GAT	0.379 ± 0.051	0.731 ± 0.164	$0.488 \pm 0.013^{\dagger}$	0.315 ± 0.111
	AGNN	0.369 ± 0.081	0.552 ± 0.255	0.416 ± 0.082	0.449 ± 0.100
	Fixed-Threshold	0.470 ± 0.050	0.369 ± 0.040	0.412 ± 0.035	0.481 ± 0.044
	Multimodal Relational Aggregation	0.517 ± 0.057	0.523 ± 0.060	$\mathbf{0.519 \pm 0.056}$	0.382 ± 0.086
	Distributional (bins)	0.284 ± 0.037	0.500 ± 0.075	0.362 ± 0.049	0.300 ± 0.031
	Distributional (quantiles)	0.324 ± 0.018	0.734 ± 0.057	0.449 ± 0.025	0.314 ± 0.030
	Distributional (bins+quantiles)	0.324 ± 0.021	0.738 ± 0.051	0.450 ± 0.027	0.308 ± 0.031
	Multimodal (Relational + Bins + Quantiles)	0.310 ± 0.018	0.874 ± 0.038	$0.457 \pm 0.023^{\ddagger}$	0.382 ± 0.086

Table 5: 5-Fold CV results on the test sets of Echo, Gab, and Parler datasets using the best (F1-score-wise) configuration. Top results are in bold face, second and third-best results are marked with † and ‡, respectively.

for (social-) context-specific modeling.

A concrete illustrative example The benefits of the multimodal approach are evident through the actual probability models assigned to a user.[7] An illustrative example is provided in Figure 2. The figure depicts three ego networks of a specific user (black frame; gold label: hate-monger). The three ego networks, identical in nodes and structure, differ only in node colors: the color of a node reflects probability assigned by a specific model: Multimodal (left), Text-only (Fixed-Threshold; middle), and network only (node2vec; right). The figure highlights three important observations:

1. The text-only aggregative model (Fixed Threshold, optimized) fails to identify the ego node as a hatemonger. The threshold of the text-based aggregative model is optimized in order to achieve the best performance on the data. However, for this reason it can misclassify users that do not cross this threshold. For example, two of the texts posted by the focal node in Figure 2 are "We should take time to thank the left. Every time they call a totally reasonable person a White Supremacist Nazi, you grow our ranks." and "We've done it goys, the #AltRight has gone mainstream, and we will purge these c*cks from the GOP in the

[7]The results reported in Table 5 are based on a binary classification (convert probability to class assignment using the standard 0.5 decision threshold).

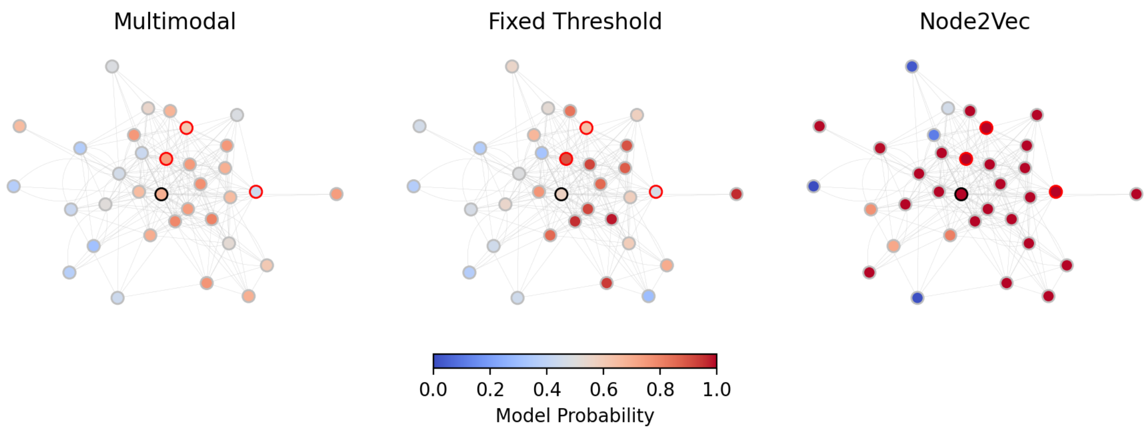

Figure 2: Ego network of user 19543095 (Echo data, gold frame=hateful). Node colors show hate probability assigned by: Multimodal (left), Text-only (Fixed-Threshold, middle), Network-only (Node2Vec, right). 0.5 is threshold.

near future". These (especially the latter one) are clear demonstrations of hate speech. The need for a threshold in text-based models can be justified by the need to give users some level of grace and/or assume that some texts are ironic.

2. The node2vec classifier indeed classifies the ego-node correctly, but it also significantly overestimates the probabilities of the other nodes, classifying most nodes as hatemongers with very high probability, while classifying only a small subset of nodes as non-hateful users with probabilities close to 0.

3. Unlike the very high/low probabilities assigned by the network only (node2vec) model, both the text-only model and the multimodal models assign probabilities closer to the 0.5 threshold. However, while the fine-tuned and calibrated text-only model still fails to cross the decision threshold, the multimodal approach used the network structure to push the decision beyond the threshold, achieving the correct classification.

Further Analysis While out of the scope of this work, it is worth noting that some models fluctuate in their performance.

The DeGroot model never achieves top performance. However, while achieving decent results over the GAB and the Parler datasets, it performs poorly over the Echo dataset (F-score: 0.31), compared to the best-performing model (F-score: 0.86) and even to the penultimate GAT (F-score: 0.574).

Similarly, the node2vec model shows decent results over the Echo and Parler datasets while performing poorly over the Gab dataset with an F-score of 0.364, compared to 0.55 and 0.264 of the best and worst performing models (extended Multimodal and Graph Attention networks, respectively). The performance of the GAT model on the Gab data is also exceptionally poor compared to decent GAT performance over other datasets.

We attribute these inconsistent performances to the marriage of the unique mechanisms employed by each of the models and the unique characteristics of each dataset in terms of size, density, network structure, betweenness, modularity, etc. (see Tables 2 and 3, Figure 1, and further details in Appendix E).

These differences may stem from the way the data were collected, as well as different norms and dynamics that emerge on different platforms. A careful study of the ways these traits interact and impact model performance is planned for future work. Yet, these fluctuations in the performance of algorithms of different modalities highlight the benefits of our multimodal approach.

6 Conclusion

In this work, we proposed a robust and efficient multimodal aggregative method, combining text and social context through relational and distributional aggregations. We demonstrated the benefits of this approach for the task of hate speech and hatemonger detection over three unique and very different datasets from three social platforms: X (Twitter), Gab, and Parler, which we curated ourselves as part of this research and for future studies to come. We evaluated our methods on both a large and unique corpus constructed around an ambiguous antisemitic meme and two other datasets based on platforms known for their free speech agenda, and demonstrated their ability to correctly detect hate mongers across different social platforms.

We showed how analyzing both the content and the network features of the users significantly improves the ability to detect hate mongers and provides insights on the promotion of hate. Our findings underscore the importance of a multi modal approach in tackling the complex issue of online hate speech. By integrating textual analysis with social network dynamics, our method achieves higher accuracy and robustness compared to traditional text-only models. Furthermore, our approach is scalable and adaptable to various social media platforms, making it a valuable tool for researchers and practitioners aiming to monitor and mitigate hate speech online. We believe that this method can assist in early detection and intervention strategies, ultimately contributing to safer online communities.

7 Limitations and Ethical Considerations

Limitations This work has a number of limitations: (i) All three aggregation procedures depend on the basic classifier $\theta(t)$. An evasive user who is careful with their words, using only coded language or consistent gas-lighting, may not be identified at all if $\theta(t) < \tau^T$ for all $t \in T^u$ for a user u. This reliance on the basic classifier means that if the classifier fails to flag any of the user's texts as problematic, the aggregation methods will also fail to detect the user as a hate monger. This limitation highlights the need for more sophisticated classifiers that can detect subtle or coded forms of hate speech. Enhancing the basic classifier's ability to recognize implicit or veiled hate speech, possibly by incorporating contextual or semantic analysis techniques, might deal with that.

(ii) We use the fixed threshold τ^U as a strong baseline; however, a more comprehensive comparison to other existing methods (e.g., diffusion-based approaches and Graph Neural Networks) should be considered in future work. Although our approach demonstrates effectiveness, benchmarking it against a broader spectrum of state-of-the-art methods would provide a more thorough evaluation of its performance. Additionally, integrating elements from diffusion-based methods or GNNs might enhance our model's ability to capture complex network dynamics and improve detection accuracy.

(iii) Our work focuses on three unique datasets from different social platforms. While unique, these platforms and datasets are mostly associated with users associating themselves with political ideology of the far-right. This ideological tent may limit the generalizability of our findings. Extending the research to include a wider variety of platforms and more diverse datasets could help validate the robustness of our method across different social contexts.

Recognizing the limitations mentioned above is crucial for the ongoing development of more effective methods for detecting hate speech and hate mongers online. Addressing these challenges in future work will be essential for improving detection capabilities and ensuring that methods remain effective against evolving strategies used to spread hate speech.

Ethical Considerations The detection of hate mongers on social platforms and the detoxification of discourse are of great importance. However, mislabeling individuals as hate mongers can lead to unintended consequences, including defamation, social ostracization, and psychological distress as well as legal implications. It may also undermine the credibility of the detection systems and erode public trust in moderation efforts. Therefore, while automated systems are valuable for processing large volumes of data efficiently, they should complement rather than replace human judgment. Furthermore, ethical considerations must guide the deployment of such detection systems. Transparency in how these systems operate and clear communication about the criteria used for labeling are essential. Providing users with avenues to appeal or contest labels can help prevent unjust consequences and maintain trust between the platform and its users.

In this work, we explore computational approaches that can be used to study the phenomena at scale, rather than propose a functional end-to-end content moderation system.

References

Center on Extremism ADL. 2023. Shooter in allen, texas, embraced antisemitism, misogyny and white supremacy.

Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An early look at the parler online social network. *ArXiv preprint*, abs/2101.03820.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13:273.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *ArXiv preprint*, abs/2004.06465.

LA Amaral, A Scala, M Barthelemy, and HE Stanley. 2000. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152.

Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It'sa thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 61–70.

Eyal Arviv and Oren Tsur. 2021. From individuals to communities : community-aware language modeling for the detection of hate speech.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *ArXiv preprint*, abs/2109.05322.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

K-I Goh. 2001. Universal behavior of load distribution in scale-free networks. *Physical review letters*, 87(27):278701.

Linda Gordon. 2017. *The second coming of the KKK: The Ku Klux Klan of the 1920s and the American political tradition*. Liveright Publishing.

Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Computing Surveys*, 55(14s):1–35.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 855–864. ACM.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Darius Irani, Avyakta Wrat, and Silvio Amir. 2021. Early detection of online hate speech spreaders with learned user representations. In *Conference and Labs of the Evaluation Forum*.

Abraham Israeli and Oren Tsur. 2022. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. *Preprint*, arXiv:2406.07886.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lee Knuttila. 2011. User unknown: 4chan, anonymity and contingency. *First Monday*, 16(10).

Zachary Laub. 2019. Hate speech on social media: Global comparisons.

Shu Li, Nayyar A Zaidi, Qingyun Liu, and Gang Li. 2021. Neighbours and kinsmen: hateful users detection with graph neural network. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May*

11–14, 2021, Proceedings, Part I, pages 434–446. Springer.

Renming Liu and Arjun Krishnan. 2021. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics*, 37(19):3377–3379.

Simon Malevich and Tom Robertso. 2019. Violence begetting violence: An examination of extremist content on deep web social networks.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Reid McIlroy-Young and Ashton Anderson. 2019. From "welcome new gabbers" to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media*, volume 13, pages 651–654.

David W McMillan and David M Chavis. 1986. Sense of community: A definition and theory. *Journal of community psychology*, 14(1):6–23.

Zhenxiong Miao, Xingshu Chen, Haizhou Wang, Rui Tang, Zhou Yang, and Wenyi Tang. 2022. Detecting offensive language on social networks: An end-to-end detection method based on graph attention networks. *ArXiv preprint*, abs/2203.02123.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. *Preprint*, arXiv:2403.12403.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018a. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018b. "like sheep among wolves": Characterizing hateful users on twitter. *ArXiv preprint*, abs/1801.00317.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *ArXiv preprint*, abs/2206.09917.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *ArXiv preprint*, abs/1709.10159.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham.

Kiran K. Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. 2018. Attention-based graph neural network for semi-supervised learning. *Preprint*, arXiv:1803.03735.

Daniel Robert Thomas and Laila A Wahedi. 2023. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24):e2214080120.

Elise Thomas. 2019. Aspi explains: 8chan.

Anıl Utku, Umit Can, and Serpil Aslan. 2024. Correction to: Detection of hateful twitter users with graph convolutional network model. *Earth Science Informatics*, 17(1):867–867.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. A review of challenges in machine learning based automated hate speech detection. *Preprint*, arXiv:2209.05294.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Barry Wellman and Milena Gulia. 1999. Virtual communities as communities. *Communities in cyberspace*, pages 167–194.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. International World Wide Web Conferences Steering Committee.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*.

A Analysis of Table 1 Examples

Here we provide a more detailed analysis of the examples presented in Table 1:

- Utterance #1: Explicit racism and Islamophobia. Detected by most models, though deHateBERT's confidence (0.57) is surprisingly low.
- Utterance #2: Antisemitism conveyed via the (((echo))) symbol targeting Amy Schumer (who is Jewish). Requires external knowledge of the symbol and target. Missed by DHB and Jigsaw; GPT-4 flags it cautiously.
- Utterance #3: The "14 words," a core white supremacist slogan. Appears innocuous without context. Only detected by the LLMs (GPT-4, Gemini).
- Utterances #4-7: Variations or allusions related to white identity and the "14 words" (#4, #5, #7) or white pride (#6). Detection is inconsistent. FTDB captures #4 and #6 but misses the original #3. LLMs are uncertain or miss these. #7 ("White lives matter") co-opts BLM language with white supremacist intent[8].
- Utterance #8: "Blue lives matter" responds to BLM, supporting police[9]. Not classified as hate speech by the models.

These examples highlight the difficulty of detecting coded language, requiring contextual understanding beyond literal text, motivating user-level aggregation.

B Prompts and Generated Responses for LLMs

Offensive content warning: Some of the prompts (based on the examples in Table 1) may be offensive to some readers. Slurs and swear words are obfuscated by using the * symbol.

```
 Which of the following
utterances may be perceived as
hate speech?

 1. Good morning boys and girls!

 2. Muslim terrorists did this and
    Obama brought them here to run
    our government. Gotta get rid
    of Obammy, Omar, and all these
    other f**king Muslims.

 3. Is there a more debased
    creature in our modern
    pop-cultural cesspool than
    (((Amy Schumer)))? Probably
    not.

 4. We must secure the existence
    of our people and a future for
    white children!.

 5. If white people are going to
    continue existing - we must
    support our race.

 6. When the youngest white
    generation becomes adult
    they will be 5% of the world
    population, down from 1/3 in
    1900.

 7. Did you know that most dog
    breeds were bred by whites?
    We're awesome and they're
    awesome!

 8. White lives matter!

 9. Blue lives matter!

10. Black lives matter!
```

GPT-4 The utterances that may be perceived as hate speech are those that explicitly or implicitly promote hate, discrimination, or hostility towards particular groups based on their ethnicity, religion, or other inherent characteristics. Based on the examples you've provided, here's an analysis of which might be considered hate speech:

2. **"Muslim terrorists did this and Obama brought them here to run our government. Gotta get rid of Obammy, Omar, and all these other fucking Muslims"** - This statement is clearly hate speech. It uses derogatory language targeting Muslims and makes unfounded, harmful generalizations about an entire religious group.

3. **"Is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not."** - This statement might be perceived as hate speech, particularly because the use of triple parentheses is often

[8] www.adl.org/resources/hate-symbol/white-lives-matter

[9] Often used alongside racial slurs but not inherently hate speech itself.

a coded way to identify someone as Jewish in an anti-Semitic context.

4. **"We must secure the existence of our people and a future for white children!"** - This phrase is widely recognized as a white supremacist slogan, known as the "Fourteen Words." It is used to promote the idea of white racial superiority and is considered hate speech.

5. **"If white people are going to continue existing - we must support our race"** - This could be interpreted as promoting racial supremacy or separatism, which might be viewed as hate speech depending on the context in which it is used.

The other statements, while potentially controversial or promoting particular views, do not explicitly convey hate against a group based on inherent characteristics in the same way that the identified statements do.

Gemini 1.5 Pro Analysis of Potentially Hateful Utterances: Here's a breakdown of the provided utterances and their potential to be perceived as hate speech:

1. **Good morning boys and girls!**
   - This is a simple greeting and does not contain any hateful content.
2. **Muslim terrorists did this and Obama brought them here to run our government. Gotta get rid of Obammy, Omar and all these other fucking Muslims.**
   - This statement employs harmful stereotypes against Muslims, associating them with terrorism and expressing a desire to remove them from the government. This is a clear example of hate speech.
3. **Is there a more debased creature in our modern pop-cultural cesspool than (((Amy Schumer)))? Probably not.**
   - The use of triple parentheses around a name is a known anti-semitic symbol online. This statement is likely intended to target Amy Schumer for her Jewish heritage and is therefore hate speech.
4. **We must secure the existence of our people and a future for white children!**
5. **If white people are going to continue existing - we must support our race**
6. **When the youngest white generation becomes adult they will be 5% of the world population, down from 1/3 in 1900**
   - These three statements, while not explicitly hateful, promote a white supremacist ideology. They focus on the preservation of white people and imply a fear of their decline, potentially inciting hostility towards other racial groups.
7. **Did you know that most dog breeds were bred by whites? We're awesome and they're awesome!**
   - This statement, while expressing pride in an accomplishment associated with white people, does not directly attack or demean any other group. It is unlikely to be considered hate speech.
8. **White lives matter!**
9. **Blue lives matter!**
10. **Black lives matter!**
    - These statements, on the surface, simply affirm the value of different groups of people. However, depending on the context and intent, they can be interpreted in different ways.
    - *White lives matter* is often used as a response to the "Black Lives Matter" movement, implying a denial of the systemic racism faced by Black people. In such cases, it can be considered hate speech.
    - *Blue lives matter* expresses support for law enforcement. While not inherently hateful, it can be used to dismiss or oppose calls for police accountability and racial justice.
    - *Black lives matter* aims to highlight the systemic racism and violence faced by Black communities. It is a call for justice and equality, not a statement of superiority or hatred towards other groups.

It's crucial to consider the context and intent behind these statements to determine if they constitute hate speech.

C Additional Experimental Details

Utterance Classification We used DistilBERT (Sanh et al., 2019b) fine-tuned on each dataset separately. Training used a batch size of 32, AdamW optimizer with learning rate 5e-5, and ran for a maximum of 20 epochs with early stopping based on validation loss (patience 5) using an 80/20 train/-validation split. The threshold τ^T was set to 0.5.

User Classification All user-level models were evaluated using 5-fold cross-validation on the largest weakly connected component (LCC) of each dataset.

- **GNN Baselines:** We followed the experimental settings from Das et al. (2021) where applicable (e.g., hidden layer sizes, dropout rates). Specific parameters might vary slightly based on library implementations (PyTorch Geometric). Node features for GNNs were derived from aggregating utterance embeddings (e.g., mean pooling of DistilBERT embeddings for user posts).
- **DeGroot's Diffusion:** We followed Israeli and Tsur (2022), running for 10 iterations. Initial beliefs were set based on a small seed set (e.g., 5
- **Node2Vec:** Used implementation from Liu and Krishnan (2021). Parameters: 50 epochs, embedding dimension 128, $p = 1, q = 1$, 10 walks per node, walk length 20, window size 10. Embeddings were fed into a Logistic Regression classifier.
- **Our Aggregative Methods:**
  - Fixed-Threshold (Θ_F): The threshold τ^U was optimized via grid search on the validation set for each fold (e.g., $\tau^U \in \{1, 3, 5, 10, 20, 50, 100\}$). Table 5 reports performance with the best τ^U.
  - Relational (Θ_R): Weights α, β, γ (Eq. 4) were learned using Logistic Regression on features derived from $C_F(u)$ and neighbor hate percentages. Best average weights across folds were approximately: Echo ($\alpha = 0.6, \beta = 0.8, \gamma = 1.5$), Gab ($\alpha = 0.8, \beta = 0.1, \gamma = 0.1$), Parler ($\alpha = 0.2, \beta = 0.3, \gamma = 0.2$).
  - Distributional (Θ_D): We used $k = 10$ bins. For bin-based, bins were [0, 0.1), [0.1, 0.2), ..., [0.9, 1.0]. For quantile-based, the range $[\min \theta(t), \max \theta(t)]$ for user u was divided into 10 equal bins. Features were the normalized counts in each bin (using softmax σ). Weights w_i (Eq. 5) were learned via Logistic Regression.
  - Combined Multimodal (Θ): Features from Θ_R, Θ_D^b, and Θ_D^q were concatenated and fed into a Logistic Regression classifier.

D Details for Figure 2 Illustrative Example

The focal user (19543095) in Figure 2 was labeled as hateful in the ground truth. The text-only aggregative model (Θ_F with optimized τ^U) assigned a score below the threshold, misclassifying the user. This occurred despite the user posting clearly hateful content, such as:

- "We should take time to thank the left. Every time they call a totally reasonable person a White Supremacist Nazi, you grow our ranks." (Implicitly aligns with those labeled Nazis/-Supremacists)
- "We've done it goys, the #AltRight has gone mainstream, and we will purge these cucks from the GOP in the near future" (Explicit use of antisemitic slur 'goys', alt-right identification, violent rhetoric 'purge')

The need for a threshold $\tau^U > 1$ in text-based models (to account for potential irony, single mistakes, or low-confidence predictions) can lead to missing users like this if their aggregated score doesn't cross the optimized threshold. The network-only model (Node2Vec) correctly classified the user but assigned extremely high probabilities (> 0.9) to most neighbors, lacking nuance. The multimodal model successfully integrated the weak textual signal (below threshold τ^U) with the strong network signal (hateful neighbors) to correctly classify the user with a more calibrated probability (just above 0.5).

E Networks Statistics

Tables 6 and 3 (in main text) provide network statistics. Table 6 shows details for the full networks (excluding users with no posts), highlighting the

Dataset	#Posts	#Users	#Edges	#Connected Comp.	#Singletons
Echo (Twitter)	18M	7.1K	21.4K	9.1K	2.9K
Gab	19.4M	61.4K	2.63M	40K	10.1K
Parler	115M	3.1M	11.1M	5.5M	2.4M

Table 6: Statistics for the full networks (users with >=1 post).

large number of components and singletons, justifying the focus on the largest connected component (LCC) for graph-based analyses (Table 3).

F Annotation Guidelines

Each annotator was assigned 50 Parler accounts, ensuring that each account is being annotated by three students. For each account, annotators were given the user name, self-description of the user's account, and 30 posts published by the user. The 30 posts are sampled out of the following: (i) 15 posts with the highest hate speech prediction values; (ii) Five random posts published by the account. The annotators showed a reasonable agreement level with an average Cohen's Kappa of 0.36. The full guidelines for the Parler dataset annotation process can be found here: https://shorturl.at/dMwfK

G Full Data and Code

The full data and code can be found here: http://bit.ly/4gMG3bl