KarmaNet: A Sentiment-Weighted Graph Approach to Analyzing Political Communities on Reddit

Andrei Iliescu
KTH Royal Institute of Technology
iliescu@kth.se

Leandro Duarte
KTH Royal Institute of Technology
leandrod@kth.se

Miguel Arroyo Marquez
KTH Royal Institute of Technology
miguelam@kth.se

Mingyang Chen
KTH Royal Institute of Technology
minchen@kth.se

Abstract

This study introduces KarmaNet, a novel sentiment-weighted network construction method, to improve the temporal analysis of political communities on Reddit. Leveraging the 'Reddit Politosphere' dataset (2008-2019), we compare KarmaNet networks, where edge weights are determined by comments votes consistency across subreddits, against a baseline co-commenting model. Our analysis, focusing on the 2016 U.S. election period, reveals that KarmaNet generates networks with improved structural stability, higher density, and greater sensitivity to real-world political events. Furthermore, graph embeddings derived from KarmaNet networks demonstrate superior performance in link prediction tasks compared to the baseline, suggesting a more meaningful representation of underlying community structures. These findings highlight the value of incorporating sentiment consistency for a more nuanced understanding of online political discourse, polarization dynamics, and the structural impact of major events. The codebase for these experiments is available at `https://github.com/Leandr0Duar7e/kth-id2211-project`.

1 Introduction

1.1 Motivation

Social media platforms have become key areas for public deliberation, where everyday users, political elites, journalists, and automated actors continuously co-produce a running commentary on current events. Though these discussions do not happen in a vacuum, there is a reciprocal relationship between online talk and real-life political discourse [2].

Reddit is an especially revealing case: it aggregates more than a decade of timestamped conversations in thousands of topical "subreddits," yet it also allows users to roam freely across communities under a single pseudonymous identity. This combination of persistent identity and voluntary association means that we can watch political communities form, splinter, and re-assemble at a resolution that surveys or news coverage cannot approach.

While previous research has extensively investigated political communities on Reddit, identifying patterns of polarization and hostility, existing network construction methods may not fully capture the subtle, sentiment-driven interactions that are crucial for a comprehensive understanding of political dynamics. Our initial research aimed to track how groups of political subreddits cluster together based on shared users and measure if these groupings become more separate (polarized) over time. However, we quickly came to the conclusion that the available graphs provided by the Politosphere [7] dataset did not portray any meaningful political interactions between users. Therefore, our study has derailed towards finding a greater network, that provides the needed information to conduct a thorough research on understanding political communities on Reddit.

1.2 Problem Definition

In this report, we aim to improve the fidelity of network-driven interactions that may be overlooked by conventional methods. If sentiment is a critical factor in the formation of political interactions and alignments, then a network model that explicitly accounts for it should, naturally, yield more accurate and insightful representations. This implies that the Politosphere method, while a valid baseline, may be insufficient to capture the full spectrum of online political dynamics, as it omits a crucial qualitative dimension.

Therefore, the core problem this study addresses is the lack of a comprehensive, temporally sensitive, and sentiment-aware network representation that can illuminate how political communities on Reddit form, evolve, polarize, and potentially fragment or converge over time. The challenge lies in constructing a more meaningful network that reflects these dynamics and supports more accurate analysis of political community behavior.

2 Related Work

Prior research into political communities on Reddit offers a foundation for our study. Soliman et al. (2019) characterized such communities, utilizing a similar dataset to identify relationships between subreddits, though with a greater emphasis on NLP techniques rather than the sentiment-driven network topology central to our approach [16]. Their work underscores the value of analyzing subreddit interactions to understand political discourse.

Investigations into online political behavior have also explored the nature of interactions within these communities. For instance, Efstratiou et al. (2023) examined hostility within political echo chambers on Reddit, revealing that negative interactions often occur within ideologically aligned groups, a complexity that purely structural network analyses might overlook and which motivates our sentiment-aware network construction [3]. This highlights the need for nuanced methods that can capture subtle interaction dynamics.

Methodologically, combining network structure with semantic information from user-generated content, as explored by researchers like Sawhney et al. [15], suggests pathways to reveal latent community properties. This resonates with our aim to enhance network representation through sentiment. Furthermore, the application of graph-theoretic metrics such as modularity, clustering coefficients, and centrality measures is well-established for quantifying community cohesion, polarization, and the influence of nodes, often in conjunction with analyses of real-world political events [5, 16, 12]. Such approaches affirm the validity of using these metrics for evaluating network evolution and linking online dynamics to offline phenomena, a strategy central to our study.

One relevant framework that we made use of is node2vec, which represents a significant algorithmic framework for learning continuous feature representations, or embeddings, for nodes within complex networks. Its core mechanism involves mapping nodes into a low-dimensional feature space, meticulously designed to maximize the preservation of network neighborhoods. This is achieved through a flexible, biased random walk procedure that intelligently explores diverse neighborhood structures within a graph [4]. The resulting embeddings are compact vector representations that encapsulate the intricate structural properties and relationships inherent in the original graph, thereby rendering them highly suitable for a broad spectrum of downstream machine learning tasks.

3 Dataset

We utilize the Reddit Politosphere dataset [7], which tracks political discussions across more than 600 subreddits spanning a 12-year period (2008-2019). This comprehensive dataset offers a unique opportunity to study the evolution of online political communities, capturing both major election cycles and significant sociopolitical events. The dataset consists of three primary components.

3.1 Network Data

The original Politosphere dataset provides pre-constructed yearly networks where nodes represent subreddits and edges represent user co-commenting patterns. The weighted graph connects subreddits when they share users who posted at least 10 comments in both communities during a year.

3.2 User and Reddit Metadata

The dataset includes anonymized user metadata (Table 1) with binary features extracted from usernames using sentiment analysis and pattern matching. These features track attributes that might correlate with political behavior. On top of that, it includes metadata related to each subreddit. The attributes provide information about political inclination, gun control, religion, etc., as seen in Table 2.

Author	automoderator	bot	gender	angry	anti	...	trump
"7W7I3"	0	0	0	0	0	...	0
"IZ5YE"	0	0	0	1	0	...	1

Table 1: Sample of user metadata attributes

Subreddit	Banned	Gun	Meta	Party	Politician	Region
HillaryForAmerica	0	0	0	dem	1	
HongKongProtest	0	0	0		0	world

Table 2: Sample of subreddit metadata attributes

3.3 Comments Data

The core of our analysis relies on the comments dataset, which contains the actual user interactions within political subreddits. Each comment includes critical metadata such as score (upvotes minus downvotes), author ID, subreddit, and timestamps (Table 3).

author	score	subreddit	body_cleaned	created_utc
"fR8On"	1	politics	"what organization? aq is a bunch..."	"1199145728"
"MtZfs"	-7	politics	"no his problem is that we have..."	"1199145817"
"s6Ti0"	3	politics	"war. n. 1. a contest between..."	"1199145849"

Table 3: Sample from comments dataset.

3.4 Initial Analysis

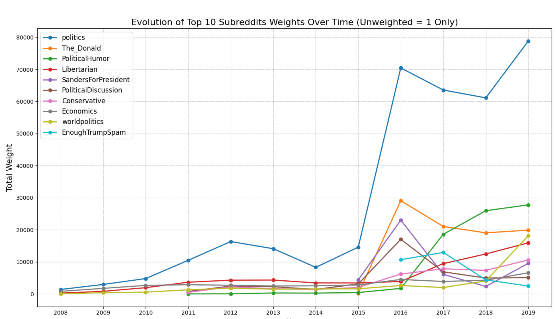


Figure 1: Evolution of connection weights for top subreddits

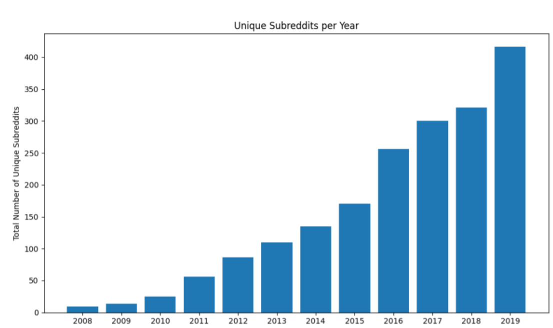
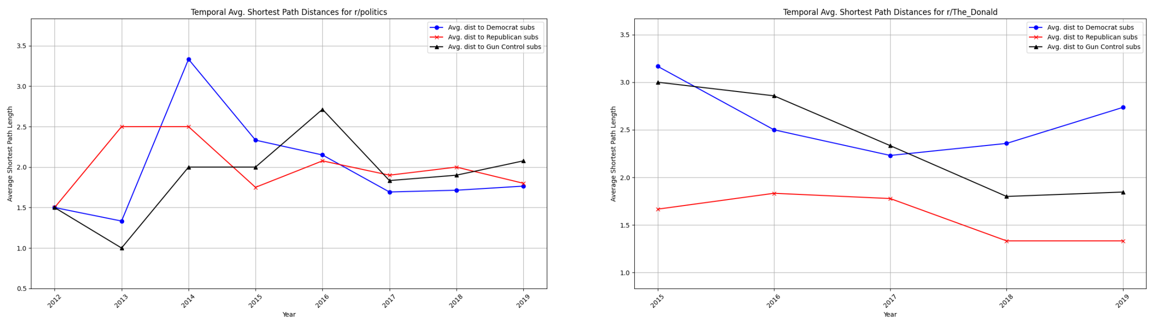


Figure 2: Growth of unique political subreddits per year

Our preliminary analysis revealed a dramatic spike in community interaction during the 2016 U.S. election cycle (Figure 1). This occurred despite a more gradual linear increase in the number of unique active subreddits (Figure 2), suggesting heightened political engagement rather than merely community expansion.

3.5 Ideological Network Structure

Analysis of shortest paths between influential subreddits and ideological groups revealed that network topology reflects real-world political dynamics.



(a) `r/politics` (b) `r/The_Donald`

Figure 3: Average shortest path distances from key subreddits to ideological groups. Lower values indicate closer proximity.

Visualization of the 2016 network (Figure 4) further demonstrates how ideological clustering manifests in the network structure.

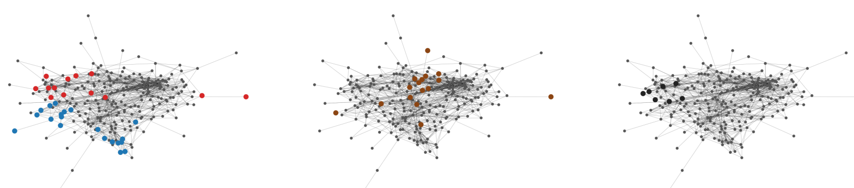


Figure 4: Network visualization of 2016 subreddits showing ideological clustering: Democratic (blue) and Republican (red) communities (left), radical subreddits later banned (center), and gun control related subreddits (right).

3.6 Data Processing for KarmaNet

Based on these initial findings, we focused our analysis on the 2016 U.S. election period, constructing monthly networks from August 2016 to February 2017. This window, frequently used in related research on political social media content [6], captures the height of campaign activity through the presidential inauguration.

For our analysis, we excluded bot accounts and "[deleted]" users, retained only specific fields including author, score, subreddit, body cleaned, id, link id, parent id and timestamps.

This processed data formed the foundation for our KarmaNet approach, which extends the original Politosphere dataset by incorporating sentiment consistency into network construction.

4 Methodology

We processed the monthly comments data to construct comparable networks using both the Politosphere commenting approach and our novel KarmaNet method. Following data cleaning we applied each method's edge weighting mechanism to the same underlying comment interactions, creating paired monthly networks for subsequent analysis.

The KarmaNet method refines the Politosphere by incorporating user sentiment consistency. An edge between two subreddits, s_i and s_j, for a given month M, is weighted based on users who not only comment in both but also create the same sentiment signal (positive or negative) in both communities. The process is defined as follows:

User Activity Variables. For a user u commenting in subreddit s during month M:

- $N_{total,u,s,M}$ is the total number of comments.
- $S_{pos,u,s,M}$ is the sum of scores from positive-scoring comments.
- $N_{pos,u,s,M}$ is the count of positive-scoring comments.
- $S_{neg\_abs,u,s,M}$ is the sum of absolute values of scores from negative-scoring comments.
- $N_{neg,u,s,M}$ is the count of negative-scoring comments.

User Sentiment Value. The sentiment value for user u in subreddit s during month M is:

$$Val_{u,s,M} = \frac{1}{N_{total,u,s,M}} \left(S_{pos,u,s,M} \cdot N_{pos,u,s,M} - S_{neg\_abs,u,s,M} \cdot N_{neg,u,s,M} \right) \tag{1}$$

User Sentiment Signal. The binary sentiment signal is determined by:

$$Sig_{u,s,M} = \mathrm{sgn}(Val_{u,s,M}) \tag{2}$$

Edge Weight. The weight of the edge between subreddits s_i and s_j for month M is:

$$W_{s_i,s_j,M} = \sum_{u \in \text{Users}} \mathbb{I}\Big((N_{total,u,s_i,M} > 0) \wedge (N_{total,u,s_j,M} > 0) \wedge (Sig_{u,s_i,M} \neq 0) \wedge (Sig_{u,s_i,M} = Sig_{u,s_j,M}) \Big) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise.

This KarmaNet approach aims to capture more nuanced ideological alignments by filtering interactions for consistent sentiment.

4.1 Construction of the training and testing data set

For each calendar month we created undirected weighted graphs ($G_t = (V_t, E_t)$) using both edge functions: Politosphere and KarmaNet, outlined in Section 3.1 and Section 4. Let the chronologically ordered sequence of graphs be $\langle G_1, G_2, \ldots, G_T \rangle$.

Expanding-window split. At test month $t > 1$ we form a cumulative *training* graph

$$G^{(t)}_{\text{train}} = \bigcup_{i=1}^{t-1} G_i,$$

Positive and negative instances. All edges in $E^{(t)}_{\text{test}}$ constitute the positive class. Because true non-edges vastly outnumber edges, we subsample an equal-sized set of negative pairs $(u, v) \notin E^{(t)}_{\text{train}}$ using three increasingly difficult strategies:

1. **Uniform random**—pairs are drawn uniformly at random from $V^{(t)}_{\text{train}} \times V^{(t)}_{\text{train}}$.
2. **Distance-2**—pairs whose endpoints share at least one common neighbour in $G^{(t)}_{\text{train}}$.
3. **Hard negatives**—pairs ranked highest by the common-neighbour count yet still absent from $G^{(t)}_{\text{train}}$.

4.2 Link Prediction using Supervised Learning

Our central objective is to build a graph that captures Reddit's political interactions more faithfully than the original Politosphere network. A standard way to test whether a new construction genuinely adds value is supervised link prediction: if the structure we learn today can accurately forecast which edges will appear tomorrow, it must encode the mechanisms that drive ties. Because this turns "graph quality" into a label-free classification problem, link prediction has become the benchmark across network science and machine learning with hundreds of papers reporting AUC or Average-Precision[17]. Graphs that excel at link prediction also deliver better results on downstream analyses such as community detection, influence estimation, and recommendation [18]. If we can prove that KarmaNet outperforms the baseline we prove that we have created a more predictive representation of Reddit's political landscape—and therefore a stronger foundation for any subsequent study of discourse, polarisation, or community dynamics.

4.2.1 Feature Selection — Based on Connectivity

Before fitting a predictive model, we must decide how to represent each candidate pair of vertices as a fixed-length feature vector. A long line of work in link prediction shows that local connectivity heuristics [10], counts or normalized counts of shared neighbours and degree statistics, carry much of the signal needed to infer missing or future edges. These measures focus on the topology of the graph and ignore extra information about the nodes, so they work even when outside data are scarce or messy. They are quick to compute and easy to explain, which makes them a strong choice for our feature set.

1. Common neighbors (CN) [11]
2. Jaccard Index (JC) [8]
3. Preferential Attachment (PA) [13]
4. Adamic–Adar Index (AA) [1]

4.2.2 Model Selection

For the binary classification problem, Logistic Regression is a strong candidate as it allows us to provide calibrated probabilities that can later be threshold or plugged into ranking-based evaluation. It also normally yields easily interpretable coefficients that reveal the relative importance of each topological cue, which will help us compare both graphs. Finally, it remains robust when the number of predictive features is moderate and highly correlated, as is often the case with graph similarity scores.

4.2.3 Training the Logistic Regression

For every candidate pair, we compute four local-topology features on $G^{(t)}_{\text{train}}$ as outlined above: Common neighbours (CN), Jaccard coefficient (JC), Adamic–Adar (AA), and preferential attachment (PA). The resulting matrix is concatenated with a binary target vector (1 for positives, 0 for negatives) to form the month-specific data set.

We evaluate an unsupervised scorer that uses CN directly and a supervised logistic-regression model fitted on an *independent* balanced sample drawn from $G^{(t)}_{\text{train}}$. Both models output a probability (or score) for every test pair, which we evaluate with AUC, average precision, and $Precision@k/Recall@k$ for $k \in \{50, 100\}$.

These metrics allow us to assess the ranking quality of the model, particularly its ability to prioritize relevant results within a limited set of recommendations. This procedure is repeated for every month, yielding a longitudinal matrix of performance metrics.

4.3 Graph Embeddings with Node2vec

Graph embedding creation involves several key steps: graph loading and preparation, negative sampling, Node2Vec embedding and scoring, and temporal evaluation. The experimental setup begins by defining several configuration parameters. For node2vec, the embedding dimension is set to 64, the walk length to 30, and the number of walks to 200. The return p and in-out q parameters, which control the random walk behavior, are both set to 1. As seen prior for link prediction, we chose some negative sampling strategies: uniform and distance-2. This aims to create more challenging negative examples, as these nodes are structurally closer than uniformly sampled pairs.

5 Evaluation

5.1 Networks Topology

The comparative analysis of Politosphere and KarmaNet networks employed graph-theoretic metrics to evaluate both static structural properties and temporal evolution patterns.

For static network characterization, we computed standard topological metrics. Network density quantified graph completeness, with higher values indicating greater interconnectedness among subreddits. Path-based metrics as average shortest path length and network diameter assessed the network's communication efficiency and information diffusion potential. Cohesion measures included the average clustering coefficient, evaluating local triadic closure propensity, and modularity, which measured community structure strength.

To analyze temporal dynamics, we employed stability metrics focused on monthly transitions. Jaccard indices for nodes and edges measured structural persistence between consecutive time points. Turnover rates quantified network volatility by measuring the fraction of components appearing or disappearing between months. Changes in modularity were tracked to detect community restructuring, particularly in response to world political events such as the 2016 U.S. presidential election.

5.2 Graph Embeddings

In the assessment of the graph embeddings, distance metrics such as Euclidean distance and cosine similarity are frequently employed for what is termed intrinsic evaluation. Intrinsic evaluation serves as a method to assess the quality of embeddings by directly analyzing the internal structure of the embedding space itself. The primary objective of this approach is to determine whether the learned embeddings accurately reflect human intuition regarding data relationships or effectively preserve the theoretical properties derived from the original graph structure. The fundamental justification for employing these metrics for intrinsic evaluation stems directly from node2vec's design principle. Since node2vec's objective is to preserve network neighborhoods, nodes that are topologically close in the original graph are intended to be spatially close in the embedding space. Thus, applying distance metrics directly tests this intended proximity preservation.

Similar to the way cosine similarity was used in [9], where they create a comprehensive evaluation methodology of structural embedding methods, we will employ cosine similarity as a scoring metric for edge existence. Using this score, we can conclude our evaluation using some common metrics in the machine learning field: area under the ROC (AUC) and average precision (AP). AUC measures the ability of the model to distinguish between positive and negative links. AP reflects the precision-recall trade-off and is particularly useful for imbalanced datasets.

5.3 Edge Prediction

We will use **Uniform**, **Distance 2** and **sample hard negatives**. For the model we have decided to use Logistic regression due to its solid out of the box performance, speed and intepretability.

For both we evaluated how well would it predict the next month, given prior months. So if given April, May, June, see how it would predict July.

6 Results

6.1 Networks Topology

The comparative analysis of network metrics revealed consistent structural and dynamic differences between networks generated by the Politosphere and KarmaNet methods. Table 4 summarizes key average monthly values.

Networks constructed using the KarmaNet method consistently exhibited higher density, shorter average path lengths, and higher average clustering coefficients. This suggests that sentiment-consistent connections foster more tightly-knit and efficient network structures. For instance, the average density for KarmaNet networks was 0.263, compared to 0.116 for Politosphere networks. Similarly, the APL for KarmaNet networks was 2.049, versus 2.456 for Politosphere.

In terms of temporal stability, KarmaNet networks demonstrated significantly higher node and edge persistence. The average Node Jaccard Index for KarmaNet networks was approximately 0.90, compared to 0.71 for

Politosphere networks. The distinction was more pronounced for Edge Jaccard Index, with KarmaNet networks averaging 0.50 while Politosphere networks averaged only 0.05. This indicates a substantially lower turnover of connections in sentiment-weighted networks. Both methods showed changes in modularity around the November 2016 US election, but the KarmaNet network exhibited a more pronounced drop and subsequent recovery, suggesting greater sensitivity to event-driven structural realignments.

Metric	KarmaNet	Politosphere
Density	0.263	0.116
Average Shortest Path	2.049	2.456
Average Clustering Coeff.	0.758	0.684
Average Nodes	347	245
Average Edges	15734	3517
Node Jaccard Index (avg.)	0.896	0.712
Edge Jaccard Index (avg.)	0.499	0.047
Modularity (election drop Oct-Nov)	-0.100	-0.050

Table 4: Summary of Average Monthly Network Metrics (Aug 2016 - Feb 2017)

6.2 Graph Embeddings

The core of our evaluation simulates a realistic temporal scenario. We iterate through the monthly graphs, building a cumulative training graph for each time step. For example, to predict links in month i, the training graph is composed of all graphs from month 1 to month $i-1$.

Graph	**Strategy**	**AUC – mean**	**std**	**AP – mean**	**std**
Politosphere	distance-two	0.4867	0.0734	0.5446	0.0463
	uniform	0.4936	0.0618	0.5496	0.0410
KarmaNet	distance-two	0.7623	0.0339	0.7742	0.0240
	uniform	0.7650	0.0289	0.7773	0.0186

Table 5: Comparison of AUC and AP for node2vec using different negative sampling strategies on two graphs.

After training, the similarity between node pairs is calculated using cosine similarity on their learned embeddings. The performance is assessed using the previously mentioned metrics: AUC and AP. Finally, we compute the mean and standard deviation of AUC and AP for each negative sampling strategy across all temporal splits, as seen in Table 6.

The table compares the performance of the Node2Vec algorithm on two graphs, Politosphere and KarmaNet. For the Politosphere graph, both strategies yield relatively low AUC and AP scores, with AUC values below 0.50 and AP values around 0.54. These low AUC values, bellow 0.5. The uniform strategy performs slightly better than distance-two, achieving an AUC of 0.4936 and an AP of 0.5496 compared to 0.4867 and 0.5446, respectively. Standard deviations for AUC are moderately high.

In contrast, the KarmaNet graph shows significantly higher performance across both metrics. AUC scores are approximately 0.76, and AP scores are around 0.77 for both strategies. Again, the uniform strategy slightly outperforms distance-two, with marginally higher mean values and lower standard deviations.

6.3 Edge prediction

These results provide compelling evidence in support of KARMANET's central claim: sentiment-weighted ties yield a more informative representation of Reddit's political interactions. KARMANET attains AUC values of 0.90 (for both uniform and distance-2 samplers) and 0.70 even under "hard" negatives, whereas POLITOSPHERE remains at or below the random-guessing threshold for two samplers (AUC = 0.17 and 0.03) and reaches only 0.34 under the easiest setting. Average precision shows a parallel gap (0.79–0.93 for KARMANET vs. 0.39–0.50 for POLITOSPHERE), confirming that KARMANET's ranking maintains discriminative power deep into the candidate list. The modest recall values (≈0.02–0.05) underscore the sparsity of the prediction task rather than any flaw in the model, and they suggest a clear path for future work that trades a small amount of

Graph	Sampler	AUC	AUC – AP	P@50	R@50	P@100	R@100
Politosphere	distance2	0.170988	0.453082	0.362778	0.057319	0.315556	0.100230
	hard	0.025512	0.392805	0.080000	0.011855	0.061389	0.018748
	uniform	0.344570	0.501359	0.469444	0.077884	0.403333	0.133052
KarmaNet	distance2	0.898208	0.919546	1.000000	0.023208	1.000000	0.046416
	hard	0.697693	0.790476	1.000000	0.023208	1.000000	0.046416
	uniform	0.907461	0.926354	1.000000	0.023208	1.000000	0.046416

Table 6: Logistic Regression Results using different samplings on both graphs

precision for broader coverage. Crucially, KARMANET's resilience to harder negative draws demonstrates that its advantage is not an artifact of sampler choice, but reflects genuine explanatory structure.

7 Discussion

The results indicate that incorporating sentiment consistency into network construction, as done by the KarmaNet method, yields networks that are structurally distinct and dynamically more stable yet responsive to major political events. The higher density, clustering, and efficiency of KarmaNet networks suggest they capture more cohesive communities of shared sentiment. The significantly greater temporal stability, particularly in edge persistence, implies that sentiment-aligned connections are more enduring than those based purely on co-commenting volume. This stability is proven by our much higher AUC score within the edge prediction where a simple model was able to quickly learn and accurately predict the following months. This enhanced stability provides a more robust baseline for tracking long-term ideological shifts. Furthermore, the heightened sensitivity of the KarmaNet network's modularity to the 2016 US election period suggests it more accurately reflects the impact of real-world events on the structure of online political discourse. These findings underscore the value of sentiment-aware approaches for a nuanced understanding of political dynamics on social media.

Considering the Node2Vec embeddings evaluation, in the case of Politosphere, its low AUC, which is below 0.5, could be interpreted as not better than random. This is concerning and could incorporate some doubt in our methodology. In addition, standard deviations for AUC are moderately high, indicating some variability in performance. These results suggest that while Node2Vec struggles to capture meaningful relationships in the Politosphere graph, potentially due to its structural properties, it performs much more effectively on the KarmaNet graph. Overall, the uniform strategy demonstrates a consistent advantage in both settings, although the performance gap is modest, indicating more consistent and accurate embeddings.

8 Conclusion

In this study, we proposed KarmaNet, a sentiment-weighted network construction method, to analyze political communities on Reddit. By incorporating sentiment consistency into edge formation, KarmaNet was able to generate networks with higher structural cohesion, greater temporal stability, and increased sensitivity to real-world political events compared to baseline co-commenting networks. Our evaluations, through both topological metrics and graph embedding-based link prediction tasks, consistently demonstrated KarmaNet's superior performance.

However, several limitations must be acknowledged. First, the dataset itself imposes constraints: although we filtered out users labeled as bots (0.16% of users), voting behaviors possibly influenced by bots could not be fully eliminated due to limitations in the underlying metadata. Additionally, while KarmaNet demonstrates strong results on Reddit, its direct applicability to other social media platforms remains uncertain. Platforms like Twitter, Facebook, or newer decentralized networks have different user interaction dynamics, identity structures, and sentiment expression mechanisms. Extending KarmaNet to these ecosystems would likely require adaptations to account for these platforms.

Despite these limitations, our work offers promising insights into how sentiment-aware network construction can improve the understanding of online political discourse and polarization. KarmaNet offers meaningful contributions to the study of online political communities and the broader field of computational social science. By embedding sentiment consistency into network construction, our approach provides a more nuanced representation of user interactions, revealing structural and temporal patterns that traditional methods might overlook.

9 Future Work

Building upon the insights gained from this study, several directions remain for future exploration:

Cross-validation with Topic Extraction: To strengthen the interpretability of sentiment-based edges, future work will involve cross-validating KarmaNet's sentiment-weighted connections against topic modeling results. This will help verify whether communities defined by sentiment consistency also exhibit thematic coherence.

Our team started implementing a prototype sentiment analysis pipeline using RoBERTa (social media optimized Deep Learning classifier) and LDA topic modeling, processing all comments through a GPU accelerated Google Cloud VM. While this successfully classified comment sentiment (-1/0/1) and identified political topics, the computational demands of processing all comments forced prioritization in our core methodology. The initial implementation remains available for future integration.

Density-Controlled Replication: While KarmaNet demonstrates high performance, further analysis could involve controlling for network density by systematically pruning edges. By reducing the volume of edges while maintaining structural properties, isolate the effects of topology from sheer interaction volume and test the robustness of our link prediction results (e.g., maintaining high AUC scores under sparser conditions).

Community Evolution Time Analysis: A finer-grained temporal analysis is needed to trace how individual communities evolve, split, or merge in response to major political events. This will involve dynamic community detection over monthly or even weekly snapshots, shedding more light on the lifecycle of political discourse clusters.

Non-Political Subreddit Expansion: Although this study focused on political subreddits, previous research has shown that a significant portion of political discussion occurs in non-political spaces [14]. Future work should expand KarmaNet to include non-political subreddits from the original Pushshift Reddit dataset, enabling a more comprehensive mapping of political discourse across the platform.

By pursuing these directions, we would be able to refine the KarmaNet model further and broaden its applicability, ultimately contributing to a deeper understanding of how online communities form and respond to societal events.

References

[1] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo. Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the italian referendum. *arXiv preprint arXiv:1702.06016*, 2017.

[3] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. Non-polar opposites: Analyzing the relationship between echo chambers and hostile intergroup interactions on reddit. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 17:197–208, 2023.

[4] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[5] Amira Haji, Benjamin Rosenbaum, Jordan Hartmann, and David Aldous. A statistical analysis of network data from reddit, 2017. Available at: https://www.causeweb.org/usproc/sites/default/files/usresp/2017-1/statistical-analysis-network.pdf.

[6] William Hobbs, Lisa Friedland, Kenneth Joseph, Oren Tsur, Stefan Wojcik, and David Lazer. "voters of the year": 19 voters who were unintentional election poll sensors on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 11:544–547, 05 2017.

[7] Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. The Reddit Politosphere: A large-scale text and network resource of online political discourse. In *Proceedings of the International AAAI Conference on Web and Social Media 16*, 2022.

[8] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272, 1901.

[9] Junchen Jin, Mark Heimann, Di Jin, and Danai Koutra. Toward understanding and evaluating structural node embeddings. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3):1–32, 2021.

[10] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

[11] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

[12] Deepanshu Malhotra and Rinkaj Goyal. Supervised-learning link prediction in single layer and multiplex networks. *Machine Learning with Applications*, 6:100086, 2021.

[13] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[14] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. Political discussion is abundant in non-political subreddits (and less toxic), 2021.

[15] Kartik Sawhney, Marcella Cindy Prasetio, and Suvadip Paul. Community detection using graph structure and semantic understanding of text. Department of Computer Science, Stanford University, 2025.

[16] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 259–263, 2019.

[17] Tao Zhou. Progresses and challenges in link prediction. *iScience*, 24(11):103217, 2021.

[18] Ruiqi Zhu, Alan Bundy, Jeff Pan, Kwabena Nuamah, Fangrong Wang, Xue Li, Lei Xu, and Stefano Mauceri. Assessing the quality of a knowledge graph via link prediction tasks. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 124–129, 2023.

10 Appendix

1. Common neighbors (CN): The common neighbors [11] between a pair of vertices is the length of the set of vertices that have a connection to both the given vertices. The formal definition is given below where $\Gamma(x)$ and $\Gamma(y)$ represent the neighbors of nodes and respectively.
$$CN = |\Gamma(x) \cap \Gamma(y)| \tag{4}$$
2. Jaccard Index (JC): The jaccard index [8] is the ratio of the number of common neighbors to the total neighbors of the gives nodes. It is calculated using the equation below.
$$JC = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{5}$$
3. Preferential Attachment (PA): The preferential attachment [13] index calculates the score for the existence of a link between two nodes by utilizing their degree information. It is computed as below.
$$PA = |\Gamma(x)| * |\Gamma(y)| \tag{6}$$
4. Adamic–Adar Index (AA): Adamic–Adar [1] uses the information of the degree of the common neighbors between the given node. Its formal definition is given below.
$$AA = \Sigma_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{7}$$

Precision@k measures the proportion of relevant items among the top-k retrieved items:
$$\text{Precision@}k = \frac{|\{\text{relevant items}\} \cap \{\text{top-}k \text{ items}\}|}{k} \tag{8}$$

Recall@k measures the proportion of relevant items that are successfully retrieved in the top-k:
$$\text{Recall@}k = \frac{|\{\text{relevant items}\} \cap \{\text{top-}k \text{ items}\}|}{|\{\text{relevant items}\}|} \tag{9}$$

These metrics allow us to assess the ranking quality of the model, particularly its ability to prioritize relevant results within a limited set of recommendations.