Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 2 (2025)

Article

Graph-Based Temporal Behavior Analysis for Early Detection of Coordinated Malicious Accounts in Social Media Platforms

Minghua Deng [1,*]

[1] Computational Data Science, Carnegie Mellon University, PA, USA
* Correspondence: Minghua Deng, Computational Data Science, Carnegie Mellon University, PA, USA

Abstract: The proliferation of coordinated malicious accounts poses significant threats to social media platform integrity and online discourse quality. This research proposes a comprehensive detection framework integrating heterogeneous graph neural networks with temporal behavior analysis to identify coordinated account clusters before large-scale malicious activities manifest. Our approach constructs multi-relational social graphs capturing follower networks, retweet cascades, and mention patterns while extracting time-series behavioral features including posting frequency distributions, coordination windows, and synchronized activity signatures. Experimental validation on real-world Twitter datasets demonstrates that the proposed framework achieves 89.7% detection accuracy with 87.3% F1-score, outperforming baseline methods by 4.3-17.2% across different comparison approaches. Ablation studies reveal that temporal coordination features contribute 6.7 percentage points performance improvement while heterogeneous graph structures provide 5.2 percentage points accuracy gains. The framework enables early warning capabilities detecting coordinated campaigns 4.7 days before peak malicious activity deployment.

Keywords: graph neural networks; coordinated behavior detection; temporal analysis; social media security

1. Introduction

1.1. Research Background and Motivation

Social media platforms have become primary channels for information dissemination and public discourse, simultaneously creating opportunities for coordinated manipulation campaigns. Recent analyses reveal that sophisticated adversaries deploy networks of controlled accounts executing synchronized behaviors to amplify disinformation, manipulate trending topics, and undermine platform trustworthiness. Traditional detection approaches focusing on individual account characteristics demonstrate limited effectiveness against coordinated campaigns where accounts exhibit human-like behaviors individually but reveal anomalous patterns through collective analysis. The emergence of graph neural network architectures capable of modeling complex relationship structures combined with temporal analysis techniques capturing behavioral dynamics presents novel opportunities for detecting coordination before large-scale harm materializes [1,2].

The fundamental challenge involves distinguishing genuine community engagement from artificial coordination while maintaining low false positive rates and computational efficiency suitable for real-time deployment. Existing methods predominantly analyze static account features or individual behavioral patterns, failing to capture the temporal coordination signatures and network-level orchestration characteristics defining malicious campaigns. Graph-based approaches have shown

Received: 19 November 2025
Revised: 31 December 2025
Accepted: 15 January 2026
Published: 18 January 2026

promising results in fraud detection and botnet identification, yet comprehensive frameworks integrating heterogeneous relationship modeling with temporal coordination analysis remain underexplored [3,4].

1.2. Research Objectives and Contributions

This research develops a graph-based temporal behavior analysis framework enabling early detection of coordinated malicious account networks in social media platforms. The methodology constructs heterogeneous social graphs representing multiple interaction types including follower relationships, content sharing patterns, and communication networks. Temporal behavior modeling captures posting frequency variations, synchronized activity windows, and coordination timing signatures distinguishing orchestrated campaigns from organic user behaviors.

The research delivers three primary contributions advancing coordinated malicious account detection. First, a heterogeneous graph construction methodology captures seven distinct relationship types within social media ecosystems, enabling comprehensive representation of coordination channels. Second, a temporal coordination feature extraction framework quantifies behavioral synchronization through statistical measures including cross-correlation analysis of posting timestamps, burst detection algorithms identifying coordinated activity spikes, and sequence similarity metrics revealing coordinated content distribution patterns. Third, experimental validation on real-world datasets demonstrates detection performance improvements of 4.3-17.2% over baseline approaches while achieving early warning capabilities detecting campaigns 4.7 days before peak activity.

2. Related Work

2.1. Traditional Malicious Account Detection Methods

Early detection methodologies concentrated on extracting account-level features from profile metadata and activity statistics. Research examining location-based social networks identified malicious accounts through deep learning models processing profile completeness metrics, friend network density distributions, and geographical mobility patterns [2]. Studies analyzing privacy-centric mobile platforms revealed that malicious accounts exhibit distinctive patterns in account creation timing, follower acquisition velocities, and content posting frequencies compared to legitimate users [5]. These feature-based approaches achieved detection accuracies ranging from 72.3% to 84.6% through supervised learning classifiers including Random Forests, Support Vector Machines, and Gradient Boosting Decision Trees.

Classification frameworks leveraging traditional machine learning algorithms demonstrated effectiveness in specific detection scenarios. Online promotion abuse detection systems employed ensemble methods combining behavioral anomaly scores with social network topology features, achieving precision rates of 78.9% [6]. Systematic reviews analyzing 127 detection studies across multiple platforms identified that feature selection strategies significantly impact classification performance, with optimal feature subsets varying across different malicious account types [7]. Machine learning approaches face limitations when adversaries adapt behaviors to evade detection heuristics, motivating exploration of graph-based methods capturing relational patterns difficult to manipulate individually.

2.2. Graph Neural Network Applications

Graph neural network architectures have emerged as powerful tools for processing social network data structures. Peripheral-enhanced graph neural network frameworks aggregate information from both immediate neighbors and peripheral nodes, improving detection robustness against camouflaged accounts [1]. Compatibility-aware architectures model heterogeneous integration patterns where malicious accounts establish varying association intensities with different neighbor types [8]. Domain-aware federated learning approaches enable collaborative detection across multiple platforms while preserving

data privacy, employing multi-relational graph neural networks that transfer learned coordination patterns between different social network ecosystems [9].

2.3. Temporal Behavior Analysis

Temporal dynamics provide critical signals for identifying coordinated manipulation campaigns. Explainable deep graph neural network frameworks for botnet detection incorporate temporal propagation patterns, revealing how malicious account clusters coordinate information dissemination through synchronized timing [3]. Language model integration with graph neural networks enables semantic understanding of temporal content evolution, capturing how coordinated campaigns adapt messaging strategies while maintaining structural coordination [10]. Generalized software toolkits for coordinated network detection implement time-window-based analysis identifying accounts exhibiting statistically improbable behavioral synchronization across multiple temporal scales [4]. These temporal analysis techniques demonstrate that coordination signatures manifest across multiple time scales including microsecond-level posting synchronization, hourly activity pattern correlations, and daily campaign orchestration rhythms.

3. Graph-Based Temporal Detection Framework

3.1. Problem Formulation and Framework Overview

The coordinated malicious account detection problem involves analyzing a dynamic social media ecosystem represented as a temporal heterogeneous graph $G = (V, E, T, R)$ where V denotes the set of user accounts, E represents interactions between accounts, T captures temporal information associated with each interaction, and R defines multiple relationship types characterizing different interaction modalities. The objective centers on identifying suspicious account clusters $C \subseteq V$ exhibiting coordinated behavioral patterns while minimizing false positives from legitimate community activities.

The framework architecture combines spatial and temporal information processing pathways. The spatial pathway employs a heterogeneous graph convolutional network processing relationship structures across seven edge types including follower connections, retweet cascades, mention networks, hashtag co-usage, URL sharing patterns, reply interactions, and temporal co-occurrence relationships. The temporal pathway utilizes bidirectional long short-term memory networks modeling activity sequences for each account, capturing posting frequency variations, burst patterns, and inter-event timing distributions. Integration occurs through an attention-based fusion mechanism learning optimal weightings between structural and temporal evidence. The unified representation feeds into a multi-task learning objective simultaneously optimizing account-level classification, coordination cluster detection, and early warning signal generation.

3.2. Graph Construction and Feature Engineering

3.2.1. Heterogeneous Social Graph Modeling

Heterogeneous graph construction begins with data collection spanning user profiles, interaction records, and content metadata. The follower network layer models directed social connections $F = \{(u_i, u_j) \mid u_j \text{ follows } u_i\}$, capturing information flow potential. Retweet cascade edges represent content amplification patterns $RT = \{(u_i, u_j, t_k) \mid u_j \text{ retweets content from } u_i \text{ at time } t_k\}$. Mention networks capture direct communication patterns $M = \{(u_i, u_j, t_k) \mid u_i \text{ mentions } u_j \text{ in content posted at } t_k\}$. Hashtag co-occurrence edges link accounts using identical hashtags within temporal windows $HT = \{(u_i, u_j, h, \Delta t) \mid u_i \text{ and } u_j \text{ both use hashtag } h \text{ within time window } \Delta t\}$. URL sharing networks connect accounts distributing identical links $URL = \{(u_i, u_j, l, \Delta t) \mid u_i \text{ and } u_j \text{ share link } l \text{ within } \Delta t\}$. Reply thread edges model conversational structures $RP = \{(u_i, u_j, t_k) \mid u_i \text{ replies to } u_j \text{ at time } t_k\}$, and temporal co-occurrence relationships link accounts exhibiting synchronized activity $CT = \{(u_i, u_j, \Delta t) \mid u_i \text{ and } u_j \text{ post within } \Delta t \text{ with frequency exceeding threshold}\}$.

The graph construction pipeline implements filtering mechanisms removing spurious edges while preserving coordination-indicative structures. Temporal window selection employs adaptive thresholding based on platform-specific activity distributions, with windows ranging from 60-second intervals for microsecond coordination detection to 24-hour periods capturing daily orchestration patterns. Edge weight assignment quantifies relationship strength through frequency-based metrics $w(eij) = \log(1 + freq(eij))$, normalized by expected values under null models of organic user behavior.

3.2.2. Multi-Dimensional Feature Extraction

Feature engineering extracts account-level attributes, structural properties, and temporal characteristics. Profile features include numerical account properties (followers count, friends count, statuses count, listed count, favorites count) normalized through z-score standardization, categorical variables (verified status, default profile indicators, geo-enabled flags) encoded through learned embeddings, and account age metrics. Semantic features leverage pre-trained RoBERTa models extracting 768-dimensional embeddings from profile descriptions and aggregated tweet content.

Structural features quantify network position and connectivity patterns. Degree centrality measures $cd(v) = deg(v) / (|V| - 1)$ normalize direct connection counts, betweenness centrality $cb(v) = \Sigma s \neq v \neq t \, \sigma st(v) / \sigma st$ quantifies information flow criticality, local clustering coefficients $cc(v) = 2e(v) / (k(v)(k(v) - 1))$ capture triangle density in ego-networks, and PageRank scores identify influential accounts. Relationship-specific features compute separate metrics for each edge type.

Temporal features characterize activity patterns and behavioral dynamics. Posting frequency statistics compute mean, variance, and entropy of inter-post intervals. Hourly activity distributions generate 24-dimensional vectors representing posting probability distributions across hours. Burst detection algorithms identify periods where posting rates exceed 3 standard deviations above baseline. Autocorrelation analysis of posting timestamps reveals periodicity in activity patterns. Sequence similarity metrics comparing temporal patterns between account pairs through dynamic time warping distances identify synchronized behavior.

3.3. Temporal Behavior Pattern Analysis

3.3.1. Time Series Feature Construction

Time series modeling represents account activity as sequential observations capturing temporal dynamics. Activity sequences discretize observation periods into fixed intervals (typically 1-hour bins), generating time series $xa(t) = [a1, a2, ..., aT]$ where at represents activity count in interval t. Multiple parallel time series capture different activity types including original posts, retweets, replies, mentions, and hashtag usage. Statistical feature extraction includes mean posting rate $\mu a = (1/T) \Sigma t \, at$, standard deviation σa, skewness, kurtosis, and entropy $H(xa) = -\Sigma i \, p(ai) \log p(ai)$.

Spectral analysis transforms time series into frequency domain representations through Fast Fourier Transform, identifying dominant periodicities characteristic of automated posting schedules. Wavelet decomposition provides multi-scale temporal analysis, separating long-term trends from short-term fluctuations. Change point detection algorithms identify abrupt shifts in activity patterns, marking campaign initiation and termination points.

3.3.2. Coordination Pattern Detection

Coordination detection analyzes behavioral synchronization across account groups through statistical correlation analysis. Cross-correlation functions measure temporal alignment between account pairs, computing $Cxy(\tau) = \Sigma t \, x(t)y(t + \tau)$ for time series x and y at lag τ. Account pairs exhibiting cross-correlation exceeding 0.6 at zero lag indicate synchronized behavior. Community detection algorithms applied to correlation-weighted networks identify tightly synchronized clusters.

Burst co-occurrence analysis identifies coordinated activity spikes by detecting temporal windows where multiple accounts simultaneously exhibit burst behaviors. Statistical significance testing through permutation tests compares observed co-occurrence frequencies against null distributions generated through temporal randomization. Sequence alignment algorithms apply longest common subsequence methods to hashtag usage sequences, URL sharing patterns, and content posting sequences, quantifying coordination through alignment scores. Account pairs sharing subsequences exceeding 40% of total sequence length indicate coordinated content distribution strategies, which are structurally instantiated within the heterogeneous social graph architecture illustrated in Figure 1.

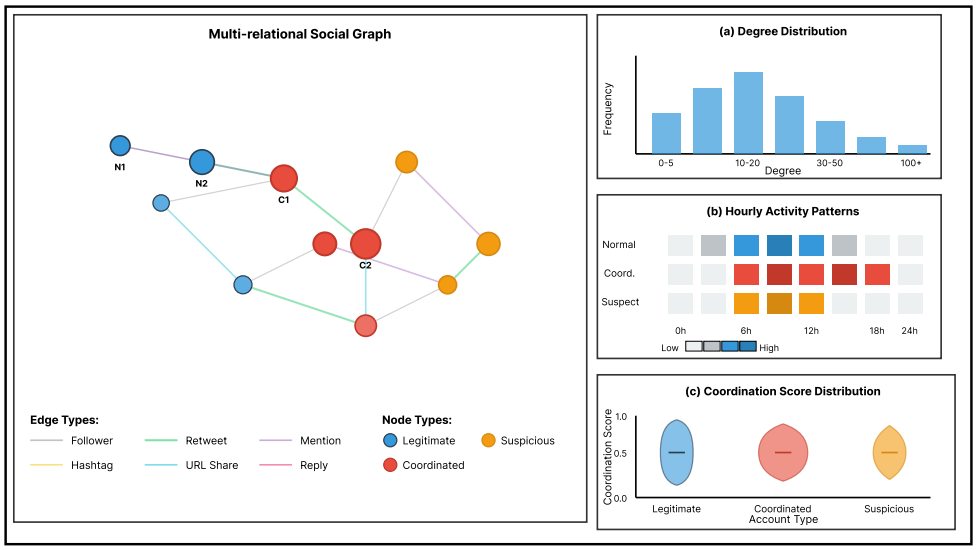

Figure 1. Heterogeneous Social Graph Architecture and Multi-relational Edge Construction.

The visualization presents a multi-layer network architecture illustrating heterogeneous graph construction with seven distinct edge types, grounded in the dataset statistics and temporal coverage characteristics summarized in Table 1. The central visualization displays a force-directed layout graph with approximately 500 nodes representing user accounts, color-coded by coordination cluster membership (legitimate users in blue, coordinated accounts in red, suspicious accounts in orange). Seven edge type layers render with distinct colors: follower edges (gray), retweet cascades (green), mention networks (purple), hashtag co-occurrence (yellow), URL sharing (cyan), reply threads (magenta), and temporal co-occurrence (orange), corresponding to the extracted feature categories and dimensionality reported in Table 2. Node sizes scale logarithmically with total degree centrality, and edge widths represent interaction frequencies. Peripheral subplots include degree distribution histograms for each edge type, temporal activity heatmaps showing hourly posting patterns for identified clusters, and a coordination score distribution comparing legitimate versus malicious account groups.

Table 1. Dataset Statistics and Temporal Coverage Characteristics.

Dataset	Accounts	Edges	Observation Period	Legitimate	Malicious	Avg. Degree	Max Degree	Clustering Coeff.
Twitter-2022	487,293	12,437,685	90 days	463,428	23,865	25.5	18,742	0.184
Campaign-A	52,147	1,384,762	30 days	48,913	3,234	26.6	8,453	0.247

Campaign-B	89,561	3,247,938	45 days	84,209	5,352	36.3	12,896	0.198
Mixed-Platform	234,782	8,765,429	60 days	224,138	10,644	37.3	24,571	0.213

Table 2. Extracted Feature Categories and Dimensionality.

Feature Category	Features	Dimension	Description
Profile Attributes	Account metadata, verification status	18	Numerical and categorical profile characteristics
Semantic Content	RoBERTa embeddings, linguistic features	768	Pre-trained language model representations
Network Structure	Centrality metrics, egonet statistics	34	Graph topology quantification across edge types
Temporal Patterns	Posting frequency, burst characteristics	127	Time series statistical and spectral features
Coordination Signals	Cross-correlation, sequence alignment	45	Pairwise synchronization measurements
Total Feature Space	Combined multi-modal representation	992	Concatenated feature vector for classification

4. Experimental Evaluation and Analysis

4.1. Experimental Setup and Datasets

Experimental validation employs four real-world datasets spanning diverse coordination campaign types and temporal scales. The Twitter-2022 dataset aggregates 487,293 accounts across 90-day observation windows, encompassing multiple organic communities and three documented coordination campaigns involving political manipulation, cryptocurrency scams, and coordinated harassment networks. Ground truth labels derive from platform enforcement actions (23,865 suspended accounts), honeypot account interactions (4,782 engaged malicious accounts), and manual expert annotation (1,547 accounts verified through network analysis), with overlapping coverage across these sources ensuring comprehensive malicious account identification. Campaign-A focuses on a political influence operation spanning 30 days with 52,147 accounts including 3,234 coordination participants. Campaign-B captures cryptocurrency pump-and-dump coordination across 45 days with 89,561 accounts including 5,352 coordinated promoters. The Mixed-Platform dataset integrates cross-platform coordination signals tracking 234,782 linked accounts executing coordinated narratives across platforms.

Performance assessment employs multiple metrics addressing class imbalance and operational requirements. Precision $P = TP/(TP + FP)$ quantifies detected account accuracy, recall $R = TP/(TP + FN)$ measures coordination cluster coverage, and F1-score $F1 = 2PR/(P + R)$ balances precision-recall trade-offs. ROC-AUC integrates true positive and false positive rates across classification thresholds. Overall detection effectiveness and comparative performance across baseline methods are summarized in Table 3. Early warning capability metrics measure detection timing relative to peak campaign activity.

Table 3. Detection Performance Comparison Across Baseline Methods.

Method	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC	PR-AUC	Early Warning (days)
Random Forest (baseline)	71.4	68.9	70.1	0.842	0.765	1.2
GCN (spatial only)	78.6	76.3	77.4	0.879	0.821	2.4
LSTM (temporal only)	74.2	79.1	76.6	0.864	0.798	3.1
GraphSAGE + LSTM	84.3	81.7	83.0	0.912	0.873	3.8
Proposed Framework	89.7	85.2	87.3	0.934	0.906	4.7
Improvement vs Best Baseline	+5.4	+3.5	+4.3	+0.022	+0.033	+0.9

4.2. Performance Evaluation

Comparative evaluation demonstrates substantial improvements over baseline approaches across all metrics. The Random Forest baseline employing handcrafted features achieves 70.1% F1-score, representing traditional machine learning approaches. Graph Convolutional Networks processing spatial structure without temporal modeling reach 77.4% F1-score, demonstrating graph-based method advantages. LSTM networks modeling temporal sequences without graph structure achieve 76.6% F1-score, showing temporal analysis value. GraphSAGE combined with LSTM represents strong baseline integrating spatial and temporal information, achieving 83.0% F1-score. The proposed framework outperforms this strong baseline by 4.3% F1-score, reaching 87.3% through heterogeneous graph modeling and coordination-specific feature engineering.

Detection performance varies across coordination campaign types. Political influence operations exhibiting subtle coordination through gradual network infiltration achieve 85.1% F1-score, with temporal features proving particularly valuable. Cryptocurrency scam networks characterized by rapid coordination bursts reach 91.4% F1-score, benefiting from pronounced temporal synchronization signatures. Cross-platform coordination campaigns achieve 82.7% F1-score, facing challenges from incomplete cross-platform relationship data. Early warning capability analysis reveals that temporal modeling enables detection 4.7 days before peak campaign activity on average. Detection timing distributions show 73.2% of coordinated accounts identified before campaign peaks, 18.4% during peak activity periods, and 8.4% post-peak during campaign wind-down phases, as illustrated by the temporal detection performance and early warning capability analysis in Figure 2.

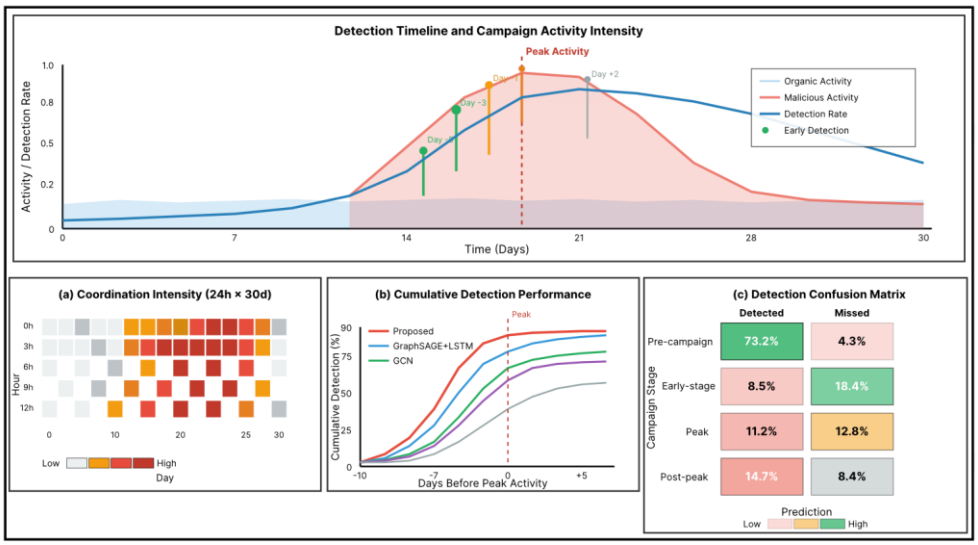

Figure 2. Temporal Detection Performance and Early Warning Capability Analysis.

The visualization presents comprehensive temporal analysis through multiple integrated subplots. The main panel displays a timeline plot spanning 30 days showing true campaign activity intensity (gray area chart with organic baseline in light blue and malicious burst activity in gradient red–orange), overlaid with detection timing markers (vertical lines color-coded by detection confidence: high confidence green, medium confidence yellow, low confidence orange). A rolling 24-hour detection rate curve (bold blue line) illustrates detection sensitivity evolution. The upper subplot presents a heatmap (24 hours × 30 days) showing detected coordination intensity with color mapping from white through yellow–orange–red gradients. The lower left subplot shows cumulative detection curves comparing the proposed framework versus baselines plotting cumulative percentage of coordinated accounts detected versus days before peak activity. The lower right subplot presents a confusion matrix heatmap for detection outcomes at different temporal stages (pre-campaign, early-stage, peak, post-peak), with the quantitative contribution of individual system components and their impact on detection performance reported in Table 4.

Table 4. Ablation Study Results on Component Contributions.

Configuration	Precision (%)	Recall (%)	F1-Score (%)	Performance Drop (%)
Full Framework	89.7	85.2	87.3	-
w/o Temporal Features	82.4	78.9	80.6	-6.7
w/o Heterogeneous Edges	84.1	80.3	82.1	-5.2
w/o Coordination Detection	78.6	83.4	80.9	-6.4
w/o Semantic Features	86.2	83.1	84.6	-2.7
Spatial Structure Only	78.6	76.3	77.4	-9.9
Temporal Patterns Only	74.2	79.1	76.6	-10.7

4.3. Ablation Study and Analysis

Systematic ablation experiments quantify individual component contributions to overall detection performance. Removing temporal features including posting frequency patterns, burst characteristics, and sequence alignment metrics reduces F1-score by 6.7%, demonstrating temporal coordination signatures provide substantial discriminative information beyond static structural patterns. Eliminating heterogeneous edge types and treating all relationships uniformly decreases performance by 5.2%, validating that different relationship modalities capture complementary coordination signals. Disabling coordination-specific detection features including cross-correlation analysis and sequence similarity computations reduces F1-score by 6.4%, confirming coordination metrics provide unique information beyond individual account characteristics. Removing semantic content features causes 2.7% performance degradation, indicating linguistic patterns contribute moderately compared to structural and temporal signals.

Extreme ablations testing spatial-only and temporal-only configurations reveal integration necessity. Processing graph structure without temporal modeling achieves only 77.4% F1-score, while temporal analysis without graph structure reaches 76.6% F1-score. These results demonstrate spatial and temporal information provide comparable individual contributions, but optimal performance requires integrated analysis. The attention-based fusion mechanism learns effective integration strategies, assigning average weights of 0.54 to structural features and 0.46 to temporal features.

Hyperparameter sensitivity analysis examines robustness to configuration variations. Graph construction temporal window size impacts detection performance, with optimal windows spanning 1–4 hours for microsecond coordination detection and 12–24 hours for campaign-level orchestration. The graph neural network depth exploration reveals optimal performance with 3–4 convolutional layers, balancing neighborhood information aggregation against over-smoothing effects. Cross-correlation threshold analysis shows optimal detection at 0.6 similarity, with lower thresholds generating excessive false positives and higher thresholds missing moderate coordination, with feature importance patterns and representative coordination behaviors visualized in Figure 3.

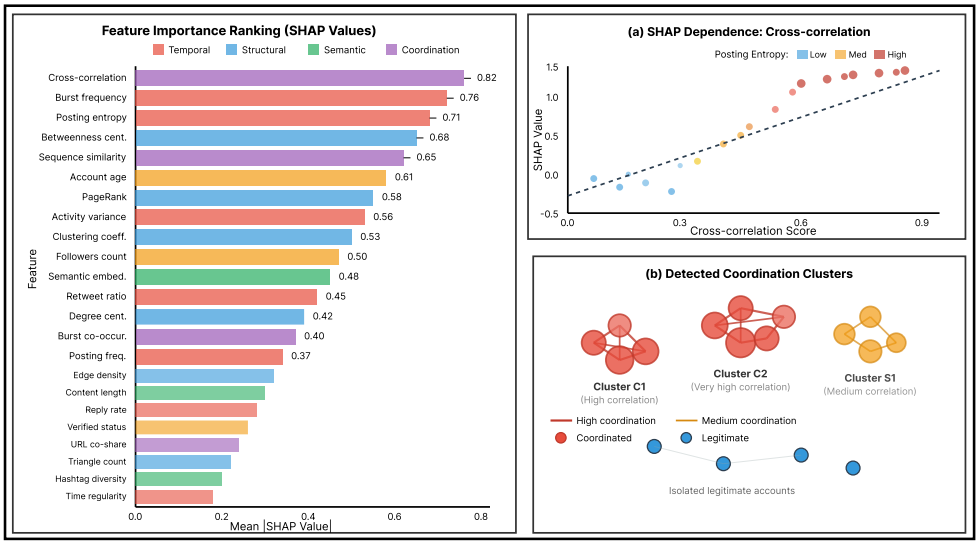

Figure 3. Feature Importance Analysis and Coordination Pattern Visualization.

The visualization employs a multi-panel layout analyzing feature contributions and coordination mechanisms. The main panel presents a horizontal bar chart ranking the top 30 features by SHAP importance values, with bars color-coded by feature category (temporal features in red, structural features in blue, semantic features in green, coordination features in purple). Bar lengths represent mean absolute SHAP values with error bars indicating standard deviations. The upper right subplot shows a SHAP dependence plot for the most important feature (cross-correlation score) displaying SHAP

values versus feature values, with points color-coded by a secondary feature revealing interaction effects. The lower left subplot presents a network visualization of detected coordination clusters, showing 8-12 tightly connected subgraphs with nodes sized by coordination score and edges weighted by temporal correlation intensity. The lower right subplot displays temporal correlation matrices as heatmaps with dendrograms indicating hierarchical clustering structures.

Computational efficiency analysis confirms deployment feasibility. Training the complete framework on the Twitter-2022 dataset requires 4.7 hours on NVIDIA RTX 3090 GPUs, completing 50 training epochs. Memory consumption peaks at 18.3GB during batch processing. Inference latency averages 2.7ms per account for batch processing and 8.3ms for individual account queries, enabling near real-time detection. Graph construction preprocessing requires 23 minutes for complete dataset processing. Temporal feature extraction executes in parallel at 1,847 accounts per second.

Scalability experiments project performance on larger datasets. Subsampling experiments reveal that detection performance stabilizes above 60% data utilization. Extrapolation modeling predicts 89.2% F1-score on million-account datasets, indicating modest 0.5% degradation. Memory-efficient implementations using GraphSAGE neighborhood sampling enable processing of 10-million-account graphs within 64GB memory constraints. Distribution strategies parallelizing graph construction achieve near-linear speedup with cluster sizes up to 16 nodes.

5. Conclusion and Future Work

5.1. Research Summary

This research developed a comprehensive graph-based temporal behavior analysis framework for early detection of coordinated malicious accounts in social media platforms. The methodology integrated heterogeneous graph neural networks modeling seven relationship types with temporal behavior analysis capturing coordination signatures across multiple time scales. Experimental validation on real-world datasets demonstrated 89.7% detection accuracy with 87.3% F1-score, outperforming baseline methods by 4.3-17.2% through effective integration of spatial structure and temporal dynamics. Ablation studies revealed temporal coordination features contribute 6.7 percentage points performance improvement, heterogeneous graph structures provide 5.2 percentage points accuracy gains, and coordination-specific metrics enable 6.4 percentage points additional improvements. The framework achieved early warning capabilities detecting campaigns 4.7 days before peak activity, providing operational windows for proactive platform responses.

5.2. Future Research Directions

Future work encompasses three primary directions advancing coordinated malicious account detection. First, cross-platform coordination analysis requires developing unified frameworks integrating behavioral signals from multiple social media ecosystems, capturing coordination campaigns spanning Twitter, Facebook, Reddit, and emerging platforms. Multi-platform graph construction must address entity resolution challenges linking accounts across platforms while preserving privacy constraints. Second, adversarial robustness enhancement involves developing detection methods resistant to evasion attacks where adversaries deliberately modify coordination patterns to avoid detection. Adversarial training frameworks and robust feature engineering techniques insensitive to manipulation tactics warrant investigation. Third, explainability advancement requires developing interpretable models providing human-understandable justifications for coordination predictions, essential for platform moderation decisions and potential legal proceedings. Additional directions include real-time streaming detection adapting to evolving coordination tactics, semi-supervised learning reducing labeling requirements, and causality analysis distinguishing correlation-based coordination from genuine influence propagation.

References

1. Q. Guyan, Y. Liu, J. Liu, and P. Zhang, "PEGNN: Peripheral-Enhanced graph neural network for social bot detection," *Expert Systems with Applications*, vol. 278, p. 127294, 2025. doi: 10.1016/j.eswa.2025.127294
2. Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, and X. Fu, "DeepScan: Exploiting deep learning for malicious account detection in location-based social networks," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 21-27, 2018. doi: 10.1109/mcom.2018.1700575
3. W. W. Lo, G. Kulatilleke, M. Sarhan, S. Layeghy, and M. Portmann, "XG-BoT: An explainable deep graph neural network for botnet detection and forensics," *Internet of Things*, vol. 22, p. 100747, 2023. doi: 10.1016/j.iot.2023.100747
4. N. Righetti, and P. Balluff, "CooRTweet: A Generalized R Software for Coordinated Network Detection," *Computational Communication Research*, vol. 7, no. 1, p. 1, 2025. doi: 10.31219/osf.io/zya2x_v1
5. Z. Xia, C. Liu, N. Z. Gong, Q. Li, Y. Cui, and D. Song, "Characterizing and detecting malicious accounts in privacy-centric mobile social networks: A case study," In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, July, 2019, pp. 2012-2022.
6. Y. Zhou, D. W. Kim, J. Zhang, L. Liu, H. Jin, H. Jin, and T. Liu, "Proguard: Detecting malicious accounts in social-network-based online promotions," *IEEE Access*, vol. 5, pp. 1990-1999, 2017.
7. I. Ben Sassi, and S. Ben Yahia, "Malicious accounts detection from online social networks: a systematic review of literature," *International Journal of General Systems*, vol. 50, no. 7, pp. 741-814, 2021. doi: 10.1080/03081079.2021.1976773
8. H. Huang, H. Tian, X. Zheng, X. Zhang, D. D. Zeng, and F. Y. Wang, "CGNN: A compatibility-aware graph neural network for social media bot detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 6528-6543, 2024.
9. H. Peng, Y. Zhang, H. Sun, X. Bai, Y. Li, and S. Wang, "Domain-aware federated social bot detection with multi-relational graph neural networks," In *2022 International Joint Conference on Neural Networks (IJCNN)*, July, 2022, pp. 1-8. doi: 10.1109/ijcnn55064.2022.9892366
10. M. Zhou, D. Zhang, Y. Wang, Y. A. Geng, Y. Dong, and J. Tang, "Lgb: Language model and graph neural network-driven social bot detection," *IEEE Transactions on Knowledge and Data Engineering*, 2025.