EDCOC: Early Detection of Coordinated Online Community using Graph Neural Networks

Hodaka Matsuzaki[1], Isao Karube[1], and Junichi Hirayama[1]

Hitachi, Ltd. Research and Development Group, Tokyo, Japan
{hodaka.matsuzaki.cs,karube.isao.ag,junichi.hirayama.qq}@hitachi.com

Abstract. Social media platforms like X (formerly Twitter) play a central role in public discourse but are also exploited for influence operations (IO) through coordinated inauthentic behavior (CIB). This study proposes a method to detect IO-related coordinated communities during the August 2023 release of Advanced Liquid Processing System (ALPS)-treated water from the Fukushima Daiichi Nuclear Power Plant. Using reposting data, we construct a graph of user communities based on network science techniques, incorporating both intra- and inter-community features. A Graph Neural Networks (GNN) is trained on these structures to classify communities as abnormal or normal. The model achieves F1 = 0.97, outperforming baseline methods. By automating early detection of coordinated communities, our method supports timely countermeasures in IO. While effective, further work is needed to capture communities with varied intents and improve attribution. The results demonstrate the potential of graph-based learning in real-time monitoring of influence activities on social platforms.

Keywords: coordinated online behavior · influence operations · graph neural networks.

1 Introduction

In today's digital society, social media serves as a dynamic platform for public discourse on politics, economics, social justice, and culture. For example, misinformation during U.S. presidential elections raised serious concerns about democratic integrity[1], while the GameStop stock surge demonstrated how coordinated actions by individual investors can affect financial markets[2]. The #MeToo movement accelerated global discussions on gender equality[3], and debates surrounding COVID-19 vaccines have been shaped by both scientific evidence and conspiracy theories[4]. While online platforms facilitate information sharing, they also enable the viral spread of disinformation and extreme views—sometimes escalating into geopolitical tensions. Efforts to deliberately shape the information environment and manipulate public perception fall within the scope of influence operations(IO). Among these, Coordinated Online Behavior (COB) plays a central role in the rapid organization of online activism[5],

boycotts[6], and protests[7]. Social media, for instance, was instrumental in mobilizing France's 2018 Yellow Vest protests[8]. However, coordination has increasingly been exploited for malicious purposes. Disinformation campaigns often involve networks of actors working together to amplify false narratives. These tactics are also used in information manipulation and astroturfing to steer public opinion toward specific political or commercial goals. Additionally, social bots and trolls amplify such content algorithmically, distort trending topics[9], and foster echo chambers that intensify online polarization[10]. These phenomena underscore the complexity and societal impact of coordinated activity in digital spaces.

Since Facebook (now Meta) introduced the concept of Coordinated Inauthentic Behavior (CIB)[11] in 2018, interest in detecting coordinated account communities has grown rapidly. Current approaches typically employ either network science or machine learning. While network methods effectively reveal latent coordination patterns, they often require manual analysis and are computationally intensive. Machine learning offers scalability but tends to rely on binary assumptions and lacks contextual nuance. Most existing work focuses on individual-level behavior, limiting its ability to evaluate the maliciousness or intent of communities due to sparse labeled data.

Rapid decision-making is crucial for countering IO, such as limiting the spread of disinformation or initiating counter-narratives. Achieving this requires the early identification of coordinated communities involved in IO. Traditional network science approaches primarily focus on detecting clusters of coordinated accounts. However, they often depend on manual analysis to assess the intent, behavior, or organizational structure behind the detected communities. Community characterization from user, content, and network perspectives can yield insights into the level of maliciousness, objectives, and underlying actor profiles[12,13]. Nevertheless, current practices remain semi-automated and heavily reliant on human interpretation. As the number of communities increases, such manual processes become a bottleneck, limiting the feasibility of rapid response. To overcome this limitation, we propose an integrated detection framework that combines network science with machine learning for the real-time classification of IO-related coordination, particularly Foreign Information Manipulation and Interference (FIMI). The framework follows a scalable two-stage pipeline. First, we construct a coordination network based on repost relationships and apply community detection methods. Second, we construct a higher-order graph where each node represents a community, characterized by aggregated user, content, and structural features. Edge features reflect inter-community behavioral similarity. To classify whether a community is maliciously coordinated, we employ Graph Neural Networks (GNN) trained on these graph-based representations.

Our approach enables effective analysis even with limited labeled data. Unlike traditional machine learning models, which typically rely on linear assumptions and content- or user-level attributes, our method captures complex non-linear dependencies among communities. A key innovation of our work is the shift from individual user-level analysis to community-level graph modeling. Exist-

ing approaches commonly construct graphs where each node represents a user and edges represent interactions. In contrast, our method defines each node as a community—a group of coordinated users—with features aggregated from user behaviors, content characteristics, and network structure. This abstraction enables higher-level analysis of coordination dynamics and better reflects real-world organized behavior, which often operates at the community level rather than the individual level. GNN are particularly well-suited for this task due to their ability to perform semi-supervised learning on structured graph data. This reduces the dependence on large labeled datasets while maintaining high detection performance. GNN also model intricate structural and relational patterns among communities, enabling accurate detection of anomalous coordination. Additionally, their scalability allows for application to large-scale social network data, facilitating the analysis of numerous communities simultaneously. While several studies have explored hybrid approaches combining network science and machine learning, few have specifically focused on modeling community-level coordination using graph-based representations for IO detection. The novelty of our approach lies in automating and scaling the detection process through a GNN-based framework that captures higher-order coordination patterns among communities. This offers a promising direction for future counter-IO strategies.

Our main contributions can be summarized as follows:

- We present a novel framework for detecting IO-related coordinated communities by integrating network science and GNN-based learning.
- Our method shifts the analysis focus from individual users to community-level nodes, enabling higher-order characterization of coordination.
- The end-to-end detection pipeline requires no human intervention.
- Our method enables effective detection even with scarce labeled data through semi-supervised learning.
- The framework is scalable and suitable for real-time analysis of large social media networks.

2 Related Work

We begin by reviewing prior research on the detection of COB on social media platforms. We then discuss existing approaches for characterizing detected coordinated communities, followed by machine learning-based COB detection methods. Finally, we examine the application of GNN—a form of semi-supervised learning—in similar problem settings, highlighting the novel contributions of our work.

Many COB detection methods rely on network science frameworks that model users and content as nodes, with interactions represented as edges[14,15]. These approaches effectively reveal group structures and visual patterns indicative of coordination. However, they often assume a binary distinction between coordinated and non-coordinated behavior, thereby overlooking the ambiguity inherent in such activities. In contrast, our study adopts the method proposed by Nizzoli et al.[16], which detects communities based on co-reposting ties and

estimates the degree of coordination, rather than imposing strict binary labels. While this method offers a more nuanced understanding, it still requires subjective manual evaluation and lacks community-level characterization, thereby limiting its capacity to identify truly malicious IO.

Characterizing detected communities is essential for interpreting behavioral patterns and inferring potential intent. This can be achieved through automated metrics derived from user, content, and network features. Lorenzo et al. [17] proposed four key evaluation dimensions: Authenticity, Harmfulness, Orchestration, and Time-variance. Authenticity assesses how accurately users present themselves, while Harmfulness evaluates the potential negative impact of their coordinated activities. Orchestration measures the degree of internal coordination within a group, and Time-variance captures temporal shifts in behavioral patterns. Although combining these metrics can yield deeper insights into coordination strategies, interpreting complex interactions often requires human judgment—especially in large-scale monitoring scenarios where rapid decision-making is critical. Our research seeks to support early detection of abnormal communities by advancing automated feature characterization techniques.

Machine learning methods offer significant advantages for representing diverse user behaviors and enabling automated decision-making. Unsupervised learning has been employed to identify coordination patterns through behavior clustering[18], with models such as textClust[19] detecting orchestrated messaging campaigns. However, clustering approaches often suffer from poor interpretability and struggle to capture non-linear patterns in high-dimensional data. Supervised learning methods, including ensemble classifiers[20], have also been explored, but their reliance on large labeled datasets limits scalability and generalizability. Our approach addresses these limitations by detecting abnormal communities through feature characterization and modeling non-linear inter-community relationships.

Recently, the application of GNN in social network analysis has garnered increasing attention. GNN effectively model relational structures among users and are widely used for tasks such as user classification and community detection[21]. They have also been integrated into recommendation systems, where they learn complex user–content interactions to deliver personalized suggestions[22]. Additionally, GNN have been applied to model and predict information diffusion across social networks[23]. Building on these insights, our study leverages GNN to learn interaction patterns within coordinated communities. We aim to efficiently extract complex inter-community relationships using only a small amount of labeled data, thereby advancing the scalability and accuracy of coordinated behavior detection.

3 Methods

We propose a method to detect coordinated communities on X involved in FIMI targeting Japan. Using the Advanced Liquid Processing System (ALPS)-treated water dataset, we perform community detection and evaluate them across four

dimensions. We also introduce a GNN-based framework for learning abnormal coordination patterns. An overview is shown in Fig. 1.

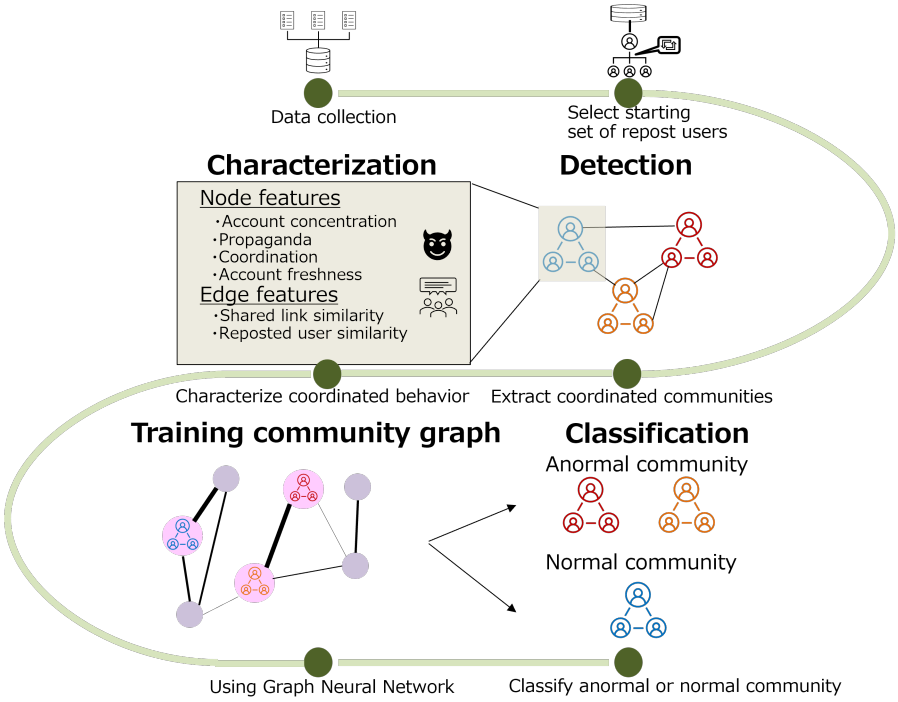


Fig. 1: Overview of the framework for detecting IO-related coordinated community.

3.1 Data Collection

In August 2023, Japan's release of ALPS-treated water from the Fukushima Daiichi plant sparked widespread debate on X. China strongly opposed the move[24], banning Japanese seafood imports, while Russia echoed similar concerns. Reports also suggested that state-led disinformation campaigns were contributing to public outrage[25]. In contrast, the IAEA deemed the discharge safe[26], a view supported by Western nations. Nonetheless, reports suggested state-led disinformation shaped critical sentiment toward Japan. To analyze this discourse, we collected X posts from April 2021 (when the release was announced) to December 2023, using search terms like "treated water," "contaminated water," and "Fukushima water" in Japanese, English, Chinese, and Korean. The dataset includes 354,882 posts, of which 255,162 are reposts by 129,996 users. We focus on repost behavior as a key signal for detecting coordinated communities.

3.2 Detecting coordinated communities

We adopt the coordination detection framework by Nizzoli et al.[16], which identifies cooperative user behavior based on repost activity. Each user's repost history is represented as a TF-IDF vector, and cosine similarity is used to build a user similarity network. This method reduces the impact of widely shared viral content and emphasizes the significance of sharing less common posts, allowing for the extraction of subtle coordination patterns. The network is refined via multi-step backbone extraction to retain only meaningful structures, and coordinated communities are identified through iterative Louvain[27] clustering. In this study, we applied this method to 255,162 reposts made by 129,996 users to detect communities involved in the ALPS-treated water discourse.

3.3 Characterizing coordinated communities

Our method builds on the multi-indicator community characterization framework proposed by Tardelli et al.[13], who evaluated coordination, automation, suspension, and bias using radar charts. We adopt a more focused structure based on Lorenzo's four key dimensions: Authenticity, Harmfulness, Orchestration, and Time-variance[17].

For Authenticity, we introduce the Account Creation Date Concentration Score (ACDCS). Marcel et al[28] found that 18 of 62 accounts suspected of inauthentic coordination—some linked to PRC diplomats—were batch-created within minutes over five days in 2020. We interpret clustered creation times as signs of inauthenticity. ACDCS is defined as:

$$ACDCS = 1 - \frac{std(t_1, t_2, ..., t_n)}{max(t_i) - min(t_i)} \tag{1}$$

This score is normalized to a range of 0 to 1, with values closer to 1 indicating that account creation is concentrated over a narrow period of time. where t_i denotes account creation times in UNIX seconds. The high score means more concentrated account creation periods.

Next, we adopted the propaganda level of repostings as a feature from the perspective of Harmfulness. Coordinated communities intended to manipulate influence often amplify posts containing propaganda. Our research group has developed a technology to detect propaganda techniques contained in text information such as SNS using natural language models[29]. This study uses this propaganda detection method to evaluate the harmfulness of the entire community. For each community, we calculate the probability (0–1) of nine propaganda techniques appearing in repost (Appeal to fear-prejudice, Causal Oversimplification, Name Calling Labeling, Whataboutism Straw Men Red Herring, Bandwagon Reductio ad hitlerum, Black-and-White Fallacy, Doubt, Exaggeration Minimisation, Loaded Language) for reposted posts in the analyzed community. The third quartile of these scores is used per technique, yielding a nine-dimensional vector per community.

To evaluate coordination strength, we employ a continuous Coordination Degree score aligned with Tardelli's framework. Each user receives a coordination score (0–1); we aggregate community-level orchestration as the area under the coordination curve. This provides a standardized metric to compare coordination intensity across communities.

Finally, we introduce an Account Freshness Age Score (AFAS) to capture the Time-variance dimension, which assesses how actor attributes change over time. This score quantifies how recently an account was created, based on the number of days since its creation t_n. A high score indicates many accounts were created shortly before the reference date d_{today}, suggesting orchestrated behavior. the score is defined by the following formula:

$$AFAS = 1 - \frac{d_{today} - t_n}{D_{max}} \tag{2}$$

where D_{max} is the referenced oldest time. We set d_{today} to September 30, 2023, and D_{max} to April 1, 2021.

In addition to node features, we define two edge-level features to capture inter-community behavioral similarity: (i) Shared Link Similarity: Jaccard index of unique URLs shared across communities. (ii) Repost Source Similarity: Normalized overlap of original accounts reposted by each community. These edge features reveal shared information sources or coordinated amplification patterns. The final edge weight is a weighted average: 60% link similarity, 40% repost similarity. Unlike node features, these capture relational dynamics crucial for detecting anomalous coordination with limited labeled data.

3.4 Labeling and training abnormal communities

Detected coordinated communities were labeled to train the GNN classifier. Communities suspected of engaging in FIMI were labeled as abnormal, while those with minimal or neutral influence were labeled as normal. Labeling criteria for abnormal communities included:

- Reposts containing anti-Japan narratives or malicious framing
- Shared links to state-controlled media or malicious image, movie, and article
- Source accounts linked to state actors or anti-Japanese influencers;
- Presence of abnormal user profiles in the community (e.g., suspended accounts, low follower/following counts).

Conversely, normal communities typically:

- Lacked such suspicious characteristics
- Reposted content defending Japan or refuting FIMI narratives
- Reposted from credible, pro-Japanese sources (e.g., Embassy accounts, Kyodo News)

The labeling process was conducted by two independent analysts. To ensure consistency and reliability, a third analyst was involved to review all labels and

resolve any discrepancies through discussion and consensus. Among the detected cooperative communities, the size (number of users) of the community is about 20-90. This was set in consideration of the realistic man-hours of labeling work and the size of the influential community. Communities with fewer than 20 users were excluded based on the assumption that their influence on public discourse was minimal. Likewise, communities with more than 90 users were excluded for both practical and methodological reasons. Extremely large communities—often comprising hundreds or thousands of users—tended to result from viral reposts and included many irrelevant or peripheral users, making coordinated behavior harder to interpret. Additionally, annotating such large communities posed significant resource challenges, and we prioritized labeling within a manageable human workload.

To classify communities as abnormal or normal, we employed a Graph Convolutional Networks (GCN)-based model. The model consists of three GCN layers with batch normalization, ELU activation, and dropout for regularization. The graph $G = (V, E)$ consists of nodes V (communities) and edges E (inter-community relationships). Each node includes aggregated features from user, content, and structural data; each edge encodes behavioral similarity between communities. GCN are particularly well-suited for this task because they effectively model relational dependencies in graph-structured data. In our case, communities are not isolated; their behavior often exhibits correlated or antagonistic patterns across the network (e.g., mirroring narratives, oppositional discourse, or coordinated propagation). By aggregating information from neighboring communities, GCN can capture these complex, non-local interactions that would be missed by traditional classifiers. Additionally, the semi-supervised nature of GCN allows for learning with limited labeled data, which is beneficial in high-cost labeling annotation scenarios like COB. The GCN comprises three layers with batch normalization, ELU activation, and dropout. The forward propagation follows:

$$\mathbf{H}^{(1)} = \text{Dropout}\left(\sigma\left(\text{BN}_1\left(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}_0\right)\right)\right) \tag{3}$$

$$\mathbf{H}^{(2)} = \text{Dropout}\left(\sigma\left(\text{BN}_2\left(\hat{\mathbf{A}}\mathbf{H}^{(1)}\mathbf{W}_1\right)\right)\right) \tag{4}$$

$$\hat{\mathbf{Y}} = \log \text{Softmax}\left(\hat{\mathbf{A}}\mathbf{H}^{(2)}\mathbf{W}_2\right) \tag{5}$$

where X is the input feature matrix with N nodes and F features per node, $\hat{\mathbf{A}}$ is normalized adjacency matrix with edge weights, using the GCN normalization rule, BN_i is batch normalization at layer i, σ is ELU activation function, W_i is trainable weight matrix at layer i.

4 Results

4.1 Detection of Coordinated communities

The repost network comprised 29,899 nodes and 339,876 edges. From this, 3,118 communities were extracted using the Louvain method. To ensure analytical relevance, we focused on 89 communities with sizes ranging from 20 to 90 accounts.

Of these, total 17 communities were labeled as abnormal, and total 72 were labeled as normal.Figure 2 shows the graph structure of these communities based on the constructed node and edge features. Red and blue nodes represent labeled abnormal and normal communities, respectively. Notable connections between abnormal nodes suggest potential cross-community coordination, while several appear isolated.

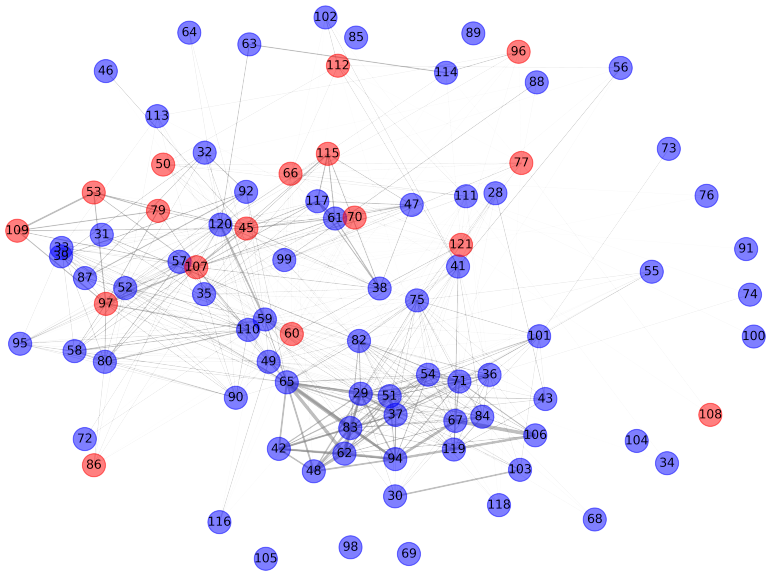


Fig. 2: Graph structure of each detected coordinated community.

4.2 Accuracy of GNN model

We evaluated classification performance using Leave-One-Out Cross Validation (LOO-CV) over the 89 labeled communities. Our GNN model achieved high accuracy, with a precision of 1.00, recall of 0.94, and F1 score of 0.97. To benchmark performance, we conducted the same classification task using other machine learning models: XGBoost, CNN, Logistic Regression, Polynomial SVM, and Random Forest. Table 1 summarizes their respective metrics. GNN outperformed all baselines, achieving the best overall scores. XGBoost and Random Forest showed relatively strong F1 scores (0.94) but lagged in either precision or recall. CNN and Polynomial SVM performed moderately, while Logistic Regression struggled to capture non-linear relationships (F1: 0.11). These results

demonstrate the advantage of graph-based learning for IO-related coordinated communities detection. By directly modeling relational structures, GNN effectively capture both intra- and inter-community patterns indicative of coordinated behavior.

Table 1: Performance comparison of learning models on classification of communities

Model	Precision	Recall	F1
GNN (our)	1.0	0.94	0.97
XGboost	0.94	0.94	0.94
CNN	0.86	0.71	0.77
Logistic Regression	1.0	0.059	0.11
Polynomial SVM	0.89	0.47	0.62
Random Forest	1.0	0.88	0.94

4.3 Features verification

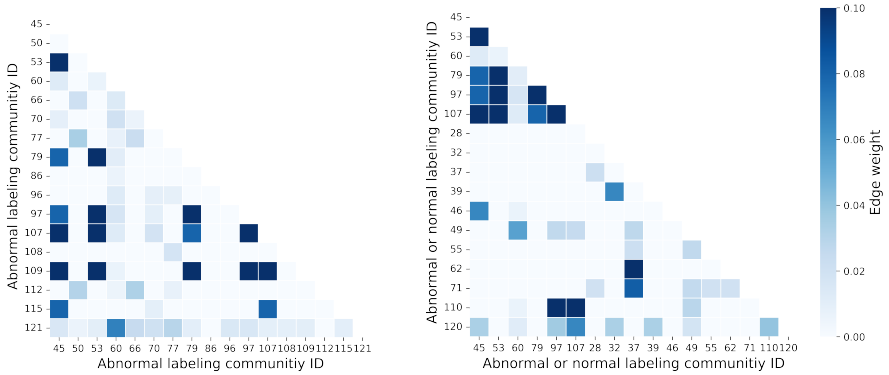
This study presents a GNN-based model trained on a graph constructed with custom-designed node and edge features to classify suspicious communities. Node features include account creation concentration, propaganda scores across nine techniques, coordination degree, and account freshness. Edge features represent similarity in shared URLs and repost sources. Experimental results (Tables 2, 3) show that combinations involving propaganda scores consistently yielded high F1 scores. This indicates that propaganda content is a strong indicator of IO. Interestingly, the best performance was observed when the coordination feature was excluded (Case No.13), while the full-feature model (Case No.15) slightly underperformed in recall. Although prior studies (e.g., Tardelli et al.[13]) highlight coordination as a key signal, our results suggest its impact may depend on dataset scale, warranting further investigation with larger samples. Edge features also contributed significantly to classification. Both combined and individual edge features retained high prediction accuracy. Heatmaps (Fig. 3a, Fig. 3b) show that abnormal communities exhibit strong internal ties and weak external connections to normal communities, consistent with expected structural separation. Notably, some anomalous communities with weak connectivity were still accurately classified, demonstrating the strength of node-based features alone. In summary, both node and edge features were effective in detecting abnormal communities, with propaganda content and inter-community structure emerging as particularly informative.

Table 2: Performance comparison of node features

No.	Node features	Precision	Recall	F1
1	ACD Concentration (A)	0.43	0.18	0.25
2	Propagnada (H)	0.94	0.88	0.91
3	Coordination (O)	0.57	0.47	0.52
4	ACD Freshness (T)	0.75	0.53	0.62
5	A,H	0.93	0.83	0.87
6	A,O	0.57	0.47	0.52
7	A,T	0.71	0.59	0.65
8	H,O	0.85	1.0	0.92
9	H,T	0.94	1.0	0.97
10	O,T	0.85	0.65	0.73
11	H,O,T	0.94	0.88	0.91
12	A,O,T	0.86	0.71	0.77
13	A,H,T	1.0	1.0	1.0
14	A,H,O	0.94	1.0	0.97
15	A,H,O,T	1.0	0.94	0.97

Table 3: Performance comparison of edge features

No.	Edge features	Precision	Recall	F1
15	2 types similarity	1.0	0.94	0.97
16	Shared link similarity	0.94	0.94	0.94
17	Reposted user similarity	0.78	0.82	0.80



(a) Abnormal labeling communities edge weight

(b) Abnormal and normal labeling communities edge weight

Fig. 3: Edge weight heatmap between community.

5 Discussion

Detecting coordinated malicious behavior online has long posed challenges in interpreting intent and community traits based on user, content, and network data. Existing methods, such as those by Tardelli[13] and Edoardo[12], have improved community-level understanding through multi-dimensional characterization. While effective, these approaches often rely on manual analyst judgments to identify abnormal communities, limiting their capacity for early, automated detection.

This study addressed this limitation by proposing a GNN-based method to classify communities as anomalous or normal, based on coordination-specific features. Our model not only learns characteristics at the community level but also captures inter-community relationships via a graph structure. LOO-CV demonstrated high performance, achieving a precision of 1.0, recall of 0.94, and F1 score of 0.97. These results extend the work of Tardelli[13], validating the effectiveness of learning structural representations for detecting coordinated abnormality. Importantly, our approach requires no manual intervention from data collection to classification, making it suitable for real-time application.

In classification accuracy, the GNN model outperformed traditional machine learning baselines trained on non-graph features. Notably, our model improved the F1 score by a factor of 7.8 compared to logistic regression, the simplest baseline. This highlights that while handcrafted feature characterization is informative, it alone is insufficient for capturing the complex, often non-linear coordination signals that arise in real-world IO. In coordinated IO, communities often exhibit non-local relational patterns such as mirroring narratives, oppositional discourse, and coordinated propagation. These patterns manifest not only in the attributes of a single community but also in how communities relate to one another—information that traditional flat-feature classifiers inherently ignore. By leveraging neighborhood aggregation, the GNN learns subtle inter-community similarities and antagonisms, allowing it to detect coordination more effectively, especially when labeled data is scarce. Our model benefits from semi-supervised learning capability and the ability to leverage graph structure, which can be advantageous in settings where relational patterns play a key role.

In this study, we adopted a transductive learning setting using GCN, assuming that all communities (nodes) and their inter-community relationships (edges) are known at training time. This setting reflects practical scenarios where analysts have already constructed the full repost network and aim to rapidly detect abnormal communities within the known graph. Under such conditions, where the entire network structure is observable, this framework can be both effective and meaningful. However, it is important to note that in many real-world applications, inductive inference is required for unseen communities or subgraphs. Future work should explore inductive GNN models that exclude test nodes and their associated edges during training, as this will be essential for properly evaluating generalization capability.

Furthermore, our approach surpasses prior methods by holistically capturing the behavioral profiles of coordinated communities. The integration of dimen-

sions such as authenticity, harmfulness, orchestration, and time-variance with graph-based linkages enables the discovery of faint yet meaningful behavioral signals that would otherwise be lost in non-graph methods.

This research focused on state-linked coordination surrounding the ALPS-treated water discourse on X. Our method enables early detection of such IO-related coordinated communities, supporting rapid response strategies such as counter-messaging or mitigation actions. However, the evaluation was conducted using a single case study (the ALPS-treated water debate) based on X data, involving 89 manually labeled communities. This raises limitations regarding generalizability. IO occur in diverse geopolitical and sociocultural contexts, such as COVID-19 disinformation or Russian narratives about Ukrainian biolabs. Future work should assess whether the framework retains its effectiveness across such varied domains and datasets. Moreover, our method has so far only been applied to repost networks on X. To test the robustness and platform-independence of our framework, future evaluations should consider applying it to other social media platforms such as Reddit or Facebook, which involve different interaction patterns. Additionally, our current comparison baselines primarily involve standard classifiers. While sufficient to demonstrate improvements over basic methods, a more comprehensive evaluation should incorporate recent community-level GNN models or hierarchical detection frameworks. The current framework is limited to binary classification of IO-related coordinated communities. Other coordinated communities may exist with distinct motivations, for instance, grassroots activists, health advocates, or conspiracy theorists. Future research should expand classification granularity to account for varied intent. Furthermore, attribution remains an open challenge; identifying responsible regions, organizations, or actors behind anomalous behavior was beyond this study's scope. Advancing attribution capabilities will be key to moving from detection to actionable insights.

6 Conclusion

This study addressed the automatic detection of communities engaged in coordinated online behavior (COB) for influence operations (IO) on social media. By characterizing communities detected through network science and training a graph-based GNN model on these features, we constructed a binary classifier to distinguish abnormal from normal coordination. Experimental results showed that even with limited labeled data, representing community features in a graph structure enabled highly accurate classification (F1 = 0.97), thus advancing the capabilities of conventional COB detection methods. Our approach enables rapid identification of high-risk communities from a large pool of candidates, supporting timely and informed responses. Moreover, by incorporating multi-dimensional features—authenticity, harmfulness, orchestration, time-variance—and behavioral similarities between communities, the model captures complex and diverse traits associated with IO. While these results are promising, the current evaluation is limited to a single-topic dataset on X. To validate the

generalizability of our approach, future work should expand to other sociopolitical topics and platforms. Additionally, incorporating more advanced baselines and enabling multi-class classification and attribution will further enhance the practical utility of this framework in real-world IO detection and response. These findings highlight the utility of graph-based models in operational contexts and suggest directions for future work, including expanding the detection scope beyond binary classification and developing attribution methods to identify the actors or entities behind coordinated communities.

References

1. Linhares, R.S., Rosa, J.M., Ferreira, C.H.G., Murai, F., Nobre, G., Almeida, J.: Uncovering coordinated communities on Twitter during the 2020 US election. In: IEEE/ACM ASONAM, pp. 80–87 (2022).
2. X. Zheng, H. Tian, Z. Wan, X. Wang, D. D. Zeng and F. -Y. Wang, Game Starts at GameStop: Characterizing the Collective Behaviors and Social Dynamics in the Short Squeeze Episode, in IEEE Transactions on Computational Social Systems, vol. 9, no. 1, pp. 45-58, Feb. (2022). doi:10.1109/TCSS.2021.3122260.
3. McClain, A.: Social media and the rise of the #MeToo movement. Medium (2020). https://medium.com/analyzing-media-bias-case-studies-2020/social-media-and-the-rise-of-the-metoo-movement-b68963f39ecf
4. Gruzd, A., Mai, P., Soares, F.B.: How coordinated link sharing behavior and partisans' narrative framing fan the spread of COVID-19 misinformation and conspiracy theories. Soc. Netw. Anal. Min. 12, 118 (2022).
5. Ng, L.H.X., Carley, K.M.: A combined synchronization index for grassroots activism on social media. arXiv (2022). https://doi.org/10.48550/arXiv.2212.13221
6. Lucchini, L., Aiello, L.M., Alessandretti, L., De Francisci Morales, G., Starnini, M., Baronchelli, A.: From Reddit to Wall Street: The role of committed minorities in financial collective action. R. Soc. Open Sci. 9(4), 211488 (2022).
7. Steinert-Threlkeld, Z.C., Mocanu, D., Vespignani, A., Fowler, J.H.: Online social networks and offline protest. EPJ Data Sci. 4, 19 (2015). France's Yellow
8. Associated Press: France's yellow vests: Who they are, what they want, and why (2019). https://apnews.com/article/49c77d0552ce4e7abfe65d5f85ecf8d9
9. Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., Blackburn, J.: Who let the trolls out? Towards understanding state-sponsored trolls. In: Proc. 10th ACM Conf. Web Sci., pp. 353–362 (2019). https://doi.org/10.1145/3292522.3326016
10. Vasconcelos, V.V., Constantino, S.M., Dannenberg, A., Lumkowsky, M., Weber, E., Levin, S.: Segregation and clustering of preferences erode socially beneficial coordination. Proc. Natl. Acad. Sci. USA 118(50), e2102153118 (2021). https://doi.org/10.1073/pnas.2102153118
11. Facebook: Removing bad actors from Facebook (2018). https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/
12. Loru, E., Cinelli, M., Tesconi, M., Quattrociocchi, W.: The influence of coordinated behavior on toxicity. Online Soc. Netw. Media 43–44, 100289 (2024). https://doi.org/10.1016/j.osnem.2024.100289
13. Tardelli, S., Nizzoli, L., Avvenuti, M. et al.: Multifaceted online coordinated behavior in the 2020 US presidential election. EPJ Data Sci. 13, 33 (2024). https://doi.org/10.1140/epjds/s13688-024-00467-0

14. Graham, T., Hames, S., Alpert, E.: The coordination network toolkit: A framework for detecting and analysing coordinated behaviour on social media. J. Comput. Soc. Sci. 7, 1139–1160 (2024). https://doi.org/10.1007/s42001-024-00260-z
15. Ng, L.H.X., Carley, K.M.: Online coordination: Methods and comparative case studies of coordinated groups across four events in the United States. In: Proc. 14th ACM Web Sci. Conf., pp. 12–21 (2022). https://doi.org/10.1145/3501247.3531542
16. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M.: Coordinated behavior on social media in the 2019 UK general election. In: Proc. Int. AAAI Conf. Web Soc. Media 15(1), 443–454 (2021). https://doi.org/10.1609/icwsm.v15i1.18074
17. Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., Cresci, S.: Detection and characterization of coordinated online behavior: A survey (2024). https://doi.org/10.48550/arXiv.2408.01257
18. Assenmacher, D., Clever, L., Pohl, J.S., Trautmann, H., Grimme, C.: A two-phase framework for detecting manipulation campaigns in social media. In: HCII, pp. 201–214 (2020).
19. Carnein, M., Assenmacher, D., Trautmann, H.: Stream clustering of chat messages with applications to Twitch streams. In: ER, pp. 79–88 (2017).
20. Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J.L., Stringhini, G.: "You know what to do": Proactive detection of YouTube videos targeted by coordinated hate attacks. In: ACM CSCW, pp. 1–21 (2019).
21. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv (2016). https://doi.org/10.48550/arXiv.1609.02907
22. Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., Leskovec, J.: Graph convolutional neural networks for web-scale recommender systems. In: Proc. KDD, pp. 974–983 (2018). https://doi.org/10.1145/3219819.3219890
23. Zhao, J., Zhao, J., Feng, J.: Information diffusion prediction based on cascade sequences and social topology. Comput. Electr. Eng. 109(B), 108782 (2023). https://doi.org/10.1016/j.compeleceng.2023.108782
24. Government of Japan: The discharge of ALPS treated water: Ensuring safety and paving the way toward decommissioning (2023). https://www.japan.go.jp/kizuna/2023/11/discharge_of_alps_treated_water.html
25. BBC News: Fukushima: China's anger at Japan is fuelled by disinformation (2023). https://www.bbc.com/news/world-asia-66667291
26. IAEA: IAEA comprehensive report on the safety review of the ALPS-treated water at the Fukushima Daiichi Nuclear Power Station (2023). https://www.iaea.org/sites/default/files/iaea_comprehensive_alps_report.pdf
27. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008(10), P10008 (2008). https://doi.org/10.48550/arXiv.0803.0476
28. Schliebs, M., Bailey, H., Bright, J., Howard, P.N.: China's inauthentic UK Twitter diplomacy: A coordinated network amplifying PRC diplomats. Programme on Democracy and Technology, Oxford University (2021).
29. Morio, G., Morishita, T., Ozaki, H., Miyoshi, T.: Hitachi at SemEval-2020 Task 11: An empirical study of pre-trained transformer family for propaganda detection. In: Proc. Fourteenth Workshop on Semantic Evaluation, pp. 1739–1748 (2020).