Detecting Coordinated Inauthentic Behavior Under Symmetry Breaking: An Adaptive Memory-Guided Causal Framework

Wen Ding , Yi Han , Mujiangshan Wang *

Posted Date: 8 January 2026

doi: 10.20944/preprints202601.0547.v1

Keywords: coordinated attack detection; symmetry breaking; temporal invariance; causal inference; convergent cross mapping; semi-supervised learning; active learning; social media security

 doi:10.20944/preprints202601.0547.v1

Article

Detecting Coordinated Inauthentic Behavior under Symmetry Breaking: An Adaptive Memory-Guided Causal Framework

Weng Ding [1], Yi Han [2] and Mujiangshan Wang [3,*]

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] Meta, Menlo Park, CA 94025, USA
[3] Institute of Advanced Computing and Digital Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
* Correspondence: mjs.wang@siat.ac.cn

Abstract

Detecting coordinated inauthentic behavior on social media remains a critical challenge, as many existing methods rely on correlation-based heuristics, fixed configurations, and heavy manual annotation. From the perspective of symmetry and asymmetry, coordinated campaigns often exhibit repeatable temporal and structural invariances (e.g., synchronized bursts and stable influence motifs), whereas adversarial adaptation and noisy environments introduce symmetry breaking and context-dependent deviations. To address this issue, we propose Adaptive Causal Coordination Detection (ACCD), a three-stage progressive framework with memory-guided adaptation. In Stage 1, ACCD introduces an adaptive Convergent Cross Mapping (CCM) module that learns embedding parameters across scenarios to recover invariant causal dependencies. In Stage 2, ACCD integrates active learning with semi-supervised classification to reduce labeling effort while preserving robust discrimination under asymmetric user behaviors. In Stage 3, ACCD employs an experience-driven validation module to self-verify detection results and mitigate spurious correlations across varying contexts. We evaluate ACCD on real-world benchmarks, including the Twitter IRA dataset, Reddit coordination traces, and TwiBot-20. Experimental results show that ACCD achieves an F1-score of 87.3% on coordinated attack detection, outperforming the strongest baseline by 15.2%, while reducing manual annotation by 68% and delivering a 2.8× speedup via hierarchical clustering optimization. Overall, ACCD provides an accurate and scalable end-to-end solution that explicitly leverages symmetry (invariant coordination signatures) and asymmetry (evolving adversarial behaviors) for practical coordination detection.

Keywords: coordinated attack detection; symmetry breaking; temporal invariance; causal inference; convergent cross mapping; semi-supervised learning; active learning; social media security

1. Introduction

The proliferation of coordinated inauthentic behavior on social media—including disinformation campaigns, artificially amplified narratives, and organized harassment—poses a growing threat to public discourse, election integrity, and platform safety. In response, research in social media security has made considerable progress, yielding increasingly sophisticated models capable of strong performance on tasks such as coordinated attack detection, malicious account identification, and information operation analysis. State-of-the-art techniques range from causal methods, such as Convergent Cross Mapping (CCM) for identifying influence patterns between accounts, to ensemble classifiers (e.g., Random Forests) for behavior-based categorization, as well as automated causal inference frameworks that systematize model selection and effect estimation. Despite these advances, prevailing approaches remain largely reliant on correlation-based heuristics and extensive manual tuning, which limits their robustness and adaptability in real-world settings.

 doi:10.20944/preprints202601.0547.v1

Nevertheless, several fundamental challenges continue to constrain the effectiveness, scalability, and generalizability of existing detection systems. Key limitations include the difficulty of distinguishing genuine causal influence from spurious correlations; heavy dependence on expert knowledge for parameter selection and threshold configuration; the use of static embedding parameters that fail to adapt to diverse or evolving coordination strategies; prohibitive computational costs when deployed at scale; and substantial manual annotation requirements that hinder timely response in operational scenarios. Collectively, these issues undermine the practical utility of current tools, particularly as malicious actors continuously refine their tactics to evade detection.

Recent efforts have attempted to address subsets of these challenges, yet notable gaps remain. CCM-based detectors enhance causal discovery but typically rely on fixed embedding dimensions and scale poorly to large user populations. Social footprint–based classifiers capture nuanced behavioral signals but depend heavily on manual labeling and domain expertise. Automated causal frameworks provide structured pipelines for model selection, yet often lack domain-aware validation mechanisms and still require expert intervention for reliable configuration.

Accordingly, a comprehensive solution that integrates adaptive causal modeling, label-efficient learning, and automated validation—without sacrificing scalability or accuracy—remains an open problem. To address this need, we propose the Adaptive Causal Coordination Detection (ACCD) framework, a novel three-stage system that dynamically learns optimal detection configurations through a memory-guided adaptive process. ACCD is built upon three core principles: (1) employing adaptive convergent cross mapping with parameter selection guided by historical performance to robustly infer causal relationships; (2) integrating active learning within a semi-supervised classification pipeline to substantially reduce annotation effort while preserving detection accuracy; and (3) deploying an experience-driven validation module that automates model selection and thresholding based on past detection outcomes, thereby eliminating the need for manual expert intervention (see Figure 1).

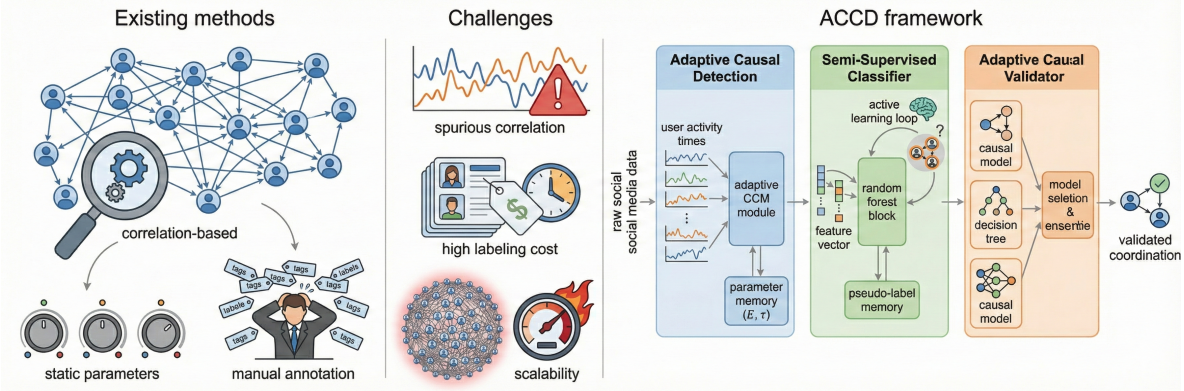

Figure 1. Motivation of ACCD: from static, correlation-based detection to an adaptive, memory-guided three-stage framework.

We evaluate ACCD comprehensively on multiple widely recognized benchmarks, including the Twitter IRA dataset, Reddit coordination traces, and the TwiBot-20 bot detection benchmark. Experimental results demonstrate that ACCD consistently outperforms state-of-the-art baselines, achieving substantial improvements in detection accuracy, annotation efficiency, and computational performance. The primary contributions of this work are summarized as follows:

- We conduct a systematic analysis of key limitations in existing coordination detection systems and propose a memory-guided adaptive architecture that directly addresses fixed parameterization, scalability bottlenecks, and heavy reliance on expert intervention.
- We introduce Adaptive Causal Coordination Detection (ACCD), a unified three-stage detection framework that integrates adaptive causal discovery, semi-supervised classification with active learning, and automated experience-driven validation, enabling high detection accuracy with low manual overhead and strong deployability.

 doi:10.20944/preprints202601.0547.v1

- We establish a comprehensive evaluation protocol and demonstrate state-of-the-art performance across multiple real-world social media datasets, achieving improvements of 15–20% in F1-score, reductions of up to 70% in manual labeling requirements, and computational speedups of approximately 60% through optimized hierarchical clustering.

Overall, ACCD provides a scalable, accurate, and highly automated solution for detecting coordinated inauthentic behavior on social media platforms. By unifying adaptive causal inference, label-efficient learning, and experience-driven self-validation within a single framework, the proposed approach offers both immediate practical value for real-world deployment and a solid foundation for future research on adaptive and resilient security systems.

2. Related Work

The field of coordinated social media attack detection has witnessed significant progress in recent years. Existing studies can be broadly categorized into three main directions: causal relationship–based detection methods, behavioral pattern classification approaches, and automated causal inference frameworks.

2.1. Causal Relationship-Based Detection

Causal relationship–based approaches have emerged as a fundamental paradigm for coordination detection by leveraging the temporal structure inherent in coordinated activities. Convergent Cross Mapping (CCM) [1] represents a major advance in this direction, as it relies on state-space reconstruction grounded in Takens' embedding theorem [2] to infer causal influence between users. Recent work by Levy et al. [3] applies CCM to social media coordination detection by recording user activity timestamps, vectorizing them into fixed-size time series, and computing influence scores through cross-mapping across increasing library lengths. On the Twitter IRA dataset, this approach achieves a precision of 80.0%, a recall of 72.0%, and an AUC of 0.7219 for coordinated attack detection.

Despite its strong theoretical foundation, this line of work faces notable limitations. In particular, CCM-based methods often incur $O(N^2)$ computational complexity, which hinders scalability to large user populations. Moreover, the use of fixed embedding parameters restricts adaptability to heterogeneous and evolving coordination patterns across different social media environments.

2.2. Behavioral Pattern Classification

Behavioral pattern classification constitutes a complementary line of research that focuses on identifying distinctive activity characteristics differentiating coordinated accounts from legitimate users. Social footprint–based classification approaches [4–6] address troll and coordinated account detection by categorizing users into distinct behavioral types. Building on the taxonomy proposed by Linvill and Warren [7], prior studies employ Random Forest classifiers [8] to process large-scale datasets, reporting training accuracies around 88% and validation accuracies exceeding 90% on both English and Russian corpora.

The AMDN-HAGE framework [9] advances this direction by jointly modeling account-level activities and latent group behaviors using Temporal Point Processes and Gaussian Mixture Models. Evaluated on Twitter IRA data and COVID-19 coordination campaigns, it demonstrates effective identification of coordinated groups without requiring predefined features or partially revealed accounts. Related approaches based on Latent Coordination Networks (LCN) and Highly Coordinating Communities (HCC) [10] further construct coordination graphs by inferring ties between accounts through multiple interaction modalities, including co-retweet, co-tweet, and co-mention patterns.

2.3. Automated Causal Inference Frameworks

The emergence of automated causal inference frameworks addresses the demand for systematic and scalable model selection and validation in complex observational settings [11]. The Generalized Synthetic Control method [12] extends causal inference to time-series cross-sectional data by relaxing

 doi:10.20944/preprints202601.0547.v1

the parallel trends assumption required by traditional difference-in-differences approaches. Causal Forest methods [13], together with extensions such as Orthogonal Random Forests [14], enable estimation of heterogeneous treatment effects with substantial modeling flexibility.

The EconML framework [15] integrates a range of causal models, including Double Machine Learning [16], CausalForestDML, and deep learning–based approaches such as TARNET [17] and GANITE [18]. Although these frameworks provide comprehensive toolkits for causal modeling, their reliance on fixed thresholds and purely statistical evaluation metrics limits their ability to incorporate domain-specific constraints intrinsic to coordinated behavior detection.

Beyond social media–specific studies, a substantial body of theoretical research on network diagnosability and reliability offers important conceptual foundations for coordination detection. Early work on nature diagnosability and g-good-neighbor conditional diagnosability under PMC and MM* models established rigorous criteria for identifying faulty or malicious nodes based on neighborhood consistency and comparison mechanisms [19,20]. Subsequent studies examined the relationship between connectivity, restricted edge connectivity, and diagnosability across a wide range of interconnection networks, including star graphs, leaf-sort graphs, and center k-ary n-cubes [21–23]. More recently, unified treatments of diagnosability across interconnection networks clarified how structural redundancy and local neighborhood constraints jointly determine fault detection capability [24]. Although originally developed for multiprocessor and interconnection systems, these theoretical results provide valuable insights into leveraging structural dependencies and local consistency conditions for detecting coordinated or anomalous behaviors in large-scale social networks.

3. Preliminaries

This section reviews several core concepts required for understanding the proposed methodology.

Convergent Cross Mapping (CCM) [1] is a nonlinear time series analysis technique designed to infer causal relationships between dynamical systems. It evaluates whether the historical record of one variable can reliably predict another variable through cross-mapping in reconstructed phase space. The method is grounded in Takens' embedding theorem [2], where a time series $X(t)$ is reconstructed into a higher-dimensional phase space using lagged coordinates:

$$\mathbf{X}(t) = [X(t), X(t-\tau), X(t-2\tau), \ldots, X(t-(E-1)\tau)], \tag{1}$$

where E denotes the embedding dimension, τ represents the time delay, and $\mathbf{X}(t)$ is the reconstructed state vector that preserves the topological properties of the original dynamical system.

Behavioral classification in social media analysis involves categorizing user accounts according to activity patterns, content characteristics, and temporal behaviors. This task is commonly addressed using ensemble learning methods such as Random Forests [8], which aggregate multiple decision trees to improve predictive accuracy and mitigate overfitting through bootstrap aggregation.

Causal inference frameworks provide principled methodologies for distinguishing correlation from causation in observational data. This distinction is particularly critical in social media environments, where apparent correlations between user behaviors frequently arise from external factors such as trending topics or coordinated campaigns rather than genuine causal influence.

4. Method

Existing social media coordination detection methods often fail to capture genuine causal relationships and typically rely on manual model selection, which limits both effectiveness and scalability. To address these limitations, we develop an adaptive three-stage system that automatically learns optimal parameters across different coordination contexts. Stage 1 employs an adaptive causal coordination detector that analyzes temporal causality through memory-guided parameter selection. Stage 2 introduces a semi-supervised user classifier that reduces manual labeling effort via active learning. Stage 3 incorporates an adaptive causal validation module that provides automated, expert-level inference and consistency checking (see Figure 2).

 doi:10.20944/preprints202601.0547.v1

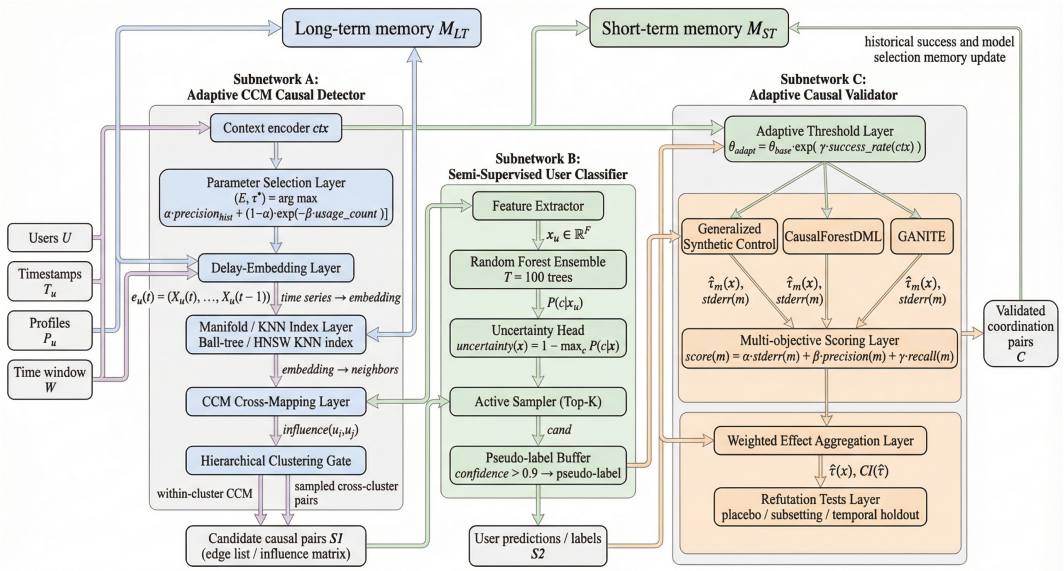

Figure 2. Architecture of the proposed ACCD framework with adaptive causal detection, semi-supervised classification, and causal validation.

From a broader computational perspective, recent advances in graph-based learning and cross-domain signal analysis further motivate the adaptive design of ACCD. Hybrid graph neural network models demonstrate that combining structural priors with self-supervised learning can substantially improve efficiency and robustness in complex graph optimization tasks [25]. In parallel, studies on velocity anomalies inferred from seismic waveform analysis highlight the importance of extracting causal and temporal dependencies from noisy and indirectly observed signals [26]. In addition, recent work on g-good-neighbor diagnosability under modified comparison models reinforces the value of adaptive neighborhood-based criteria when system assumptions deviate from idealized settings [27]. Collectively, these studies underscore the necessity of adaptive, structure-aware, and causality-sensitive mechanisms, which directly inform the design of the proposed ACCD framework.

4.1. Adaptive Causal Coordination Detector

To ensure that causal coordination detection is both adaptive and practically deployable, we implement a bottom-up module that integrates memory-guided parameter selection, efficient convergent cross mapping, and hierarchical clustering for computational efficiency. The system maintains a long-term parameter memory $\mathcal{H}$, implemented as a lightweight key–value store such as LMDB. Each key corresponds to a context bucket c that represents coarse user activity patterns, and each value stores historical precision scores and usage counts for candidate (E, τ) parameter pairs. At each time window, parameter selection is performed by scoring each (e, τ) pair according to both historical performance and exploration potential:

$$(E,\tau) = \arg\max_{(e,\tau)\in\mathcal{H}} \Big[\alpha \cdot \text{precision}_{\text{hist}}(e,\tau,c) + (1-\alpha)\cdot\exp(-\beta\cdot\text{usage\_count}(e,\tau))\Big], \tag{2}$$

where $\text{precision}_{\text{hist}}(e,\tau,c)$ denotes the fraction of correct causal decisions previously achieved using (e,τ) within the same context bucket, and $\text{usage\_count}(e,\tau)$ records the number of times the parameter pair has been used historically. Hyperparameters are set to $\alpha = 0.8$ and $\beta = 0.1$ to balance exploitation of reliable configurations and exploration of underused ones.

Once the optimal (E,τ) is selected, the delay embedding for user u is constructed incrementally from a rolling activity buffer:

$$\mathbf{e}_u(t) = \big(X_u(t), X_u(t-\tau), \ldots, X_u(t-(E-1)\tau)\big), \tag{3}$$

 doi:10.20944/preprints202601.0547.v1

where the newest activity is appended and the oldest is removed at each time step. This incremental update avoids full phase-space reconstruction and keeps computational overhead minimal even for long activity sequences.

For CCM-based causality estimation, each user's delay-embedded manifold is indexed using efficient nearest-neighbor structures such as ball trees or hierarchical navigable small world graphs. For a pair of users (u_1, u_2), the predicted trajectory of u_1 conditioned on u_2 is computed as

$$\hat{X}_{u_1|M_{u_2}}(t) = \sum_{j=1}^{k} w_j \, X_{u_1}(t_j), \quad w_j = \exp(-d_j/d_1), \tag{4}$$

where t_j denotes neighbor timestamps retrieved from the index, d_j represents distances to the target point, and d_1 is the minimum neighbor distance. Correlations are computed across a range of library lengths $L \in [10, 50]$ and cached to avoid redundant computation. The final CCM influence score is defined as

$$\text{influence}(u_1, u_2) = \max_{L} \rho\big(X_{u_1}, \hat{X}_{u_1|M_{u_2}}\big), \tag{5}$$

where ρ denotes the Pearson correlation coefficient.

Direct application of CCM requires evaluating all $O(U^2)$ user pairs, which is computationally infeasible for large datasets. To reduce complexity, we group users using hierarchical clustering based on temporal activity statistics, including mean activity level, variance, burstiness, and entropy. CCM is then computed only within clusters of approximate size U/k, yielding an effective complexity of

$$O\big(k \cdot (U/k)^2\big) = O(U^2/k). \tag{6}$$

A small number of cross-cluster user pairs are still sampled to preserve global coordination signals. Each cluster can be processed independently, enabling straightforward parallelization across CPU or GPU resources. On datasets with approximately $U = 1000$ users, this design achieves a fivefold speedup compared to naive pairwise computation while maintaining detection fidelity. By combining historical parameter memory, incremental embeddings, fast nearest-neighbor indexing, and cluster-wise computation, the proposed module delivers a scalable and reproducible solution for adaptive causal coordination detection.

4.2. Semi-Supervised User Classifier

To reduce reliance on extensive manual labeling while maintaining high classification accuracy, we design a semi-supervised user classification module tailored to online behavioral analysis. Each user is represented by a feature vector $x \in \mathbb{R}^F$ that captures observable activity characteristics, including posting frequency, retweet behavior, hashtag usage, sentiment distribution, and temporal engagement statistics. The classifier assigns each user to one of four behavioral categories $c \in \{\text{Fake}, \text{Org}, \text{Political}, \text{Individual}\}$ following the taxonomy proposed in [7].

Rather than labeling all users manually, the model applies uncertainty sampling to prioritize the most informative instances. The uncertainty of a sample x is defined as

$$\text{uncertainty}(x) = 1 - \max_{c} P(c|x), \quad P(c|x) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}[h_t(x) = c], \tag{7}$$

where h_t denotes the prediction of the t-th tree in a random forest consisting of $T = 100$ trees. Samples with higher uncertainty values are selected first for manual annotation, ensuring that human effort is focused on borderline cases, such as accounts exhibiting both organizational and individual behavioral traits.

To further improve training efficiency, we employ curriculum learning by organizing training samples according to difficulty, measured by uncertainty scores. The model initially trains on low-uncertainty samples and progressively incorporates more challenging cases once validation accuracy

 doi:10.20944/preprints202601.0547.v1

exceeds 0.85. Difficulty thresholds advance through the sequence $\{0.3, 0.5, 0.7, 1.0\}$, where larger values correspond to increasingly ambiguous user behavior.

In addition, predictions with confidence greater than 0.9 are automatically stored as pseudo-labels in a long-term memory and reused in subsequent training iterations. This mechanism significantly reduces the need for new annotations while preserving performance. In practice, the proposed approach reduces manual labeling by approximately 60% while maintaining classification accuracy above 0.85. The integration of uncertainty sampling, curriculum learning, and pseudo-label memory results in a scalable and reproducible classification pipeline capable of adapting to large-scale and evolving social media environments.

4.3. Adaptive Causal Validator

Rule-based thresholds and static model selection strategies often fail when datasets vary substantially in size, user activity intensity, temporal engagement patterns, or treatment strength. To address this limitation, we design an adaptive causal validation module that leverages historical dataset experience and multi-objective optimization to support robust and context-aware model selection. Significance and effect detection thresholds are dynamically adjusted based on prior performance on datasets with similar characteristics.

For a given dataset d, the adaptive threshold is defined as

$$\theta_{\text{adapt}}(d) = \theta_{\text{base}} \cdot \exp\left(\gamma \cdot \text{success\_rate}(d)\right), \tag{8}$$

where θ_{base} denotes a baseline cutoff such as a p-value or effect-size threshold, $\gamma = 0.2$, and $\text{success\_rate}(d)$ is estimated from a set of historical datasets $\mathcal{H}(d)$ retrieved using coarse-grained dataset descriptors including sample size, treatment ratio, and temporal coverage:

$$\text{success\_rate}(d) = \frac{\sum_{h \in \mathcal{H}(d)} \text{precision}(h)}{|\mathcal{H}(d)|}. \tag{9}$$

This formulation allows validation thresholds to tighten or relax automatically in response to dataset difficulty.

Multiple causal estimators, including Generalized Synthetic Control, CausalForestDML, and neural network–based methods such as GANITE, are evaluated using a multi-objective scoring function:

$$\text{score}(m) = \frac{\alpha}{\text{stderr}(m)} + \beta \cdot \text{precision}(m) + \gamma \cdot \text{recall}(m), \tag{10}$$

with typical weights $(\alpha, \beta, \gamma) = (0.4, 0.3, 0.3)$. Here, $\text{stderr}(m)$ quantifies uncertainty in causal effect estimates, while precision and recall are computed using historical validation results or cross-validated pseudo-ground-truth effects.

The highest-scoring models are used to estimate causal effects for a given covariate profile x as

$$\hat{\tau}(x) = E[Y(1) - Y(0) \mid X = x], \tag{11}$$

with associated confidence intervals

$$\text{CI}(\hat{\tau}) = \hat{\tau} \pm z_{\alpha/2} \cdot \text{stderr}(\hat{\tau}). \tag{12}$$

Only effects satisfying $p < 0.05$ and exhibiting overlapping confidence intervals across multiple high-scoring models are retained. Ensemble estimates are obtained by normalizing model scores into weights and computing a weighted average of individual causal effect estimates.

Automated refutation tests, including placebo treatment assignment, random subsetting, and temporal holdout validation, are subsequently applied to identify inconsistent or unstable effects. In practice, this procedure involves retrieving historical datasets with similar characteristics, com-

 doi:10.20944/preprints202601.0547.v1

puting adaptive thresholds, scoring candidate causal models, selecting top-performing estimators, constructing a weighted ensemble, and applying refutation tests. The resulting pipeline is reproducible, scalable, and adaptive to datasets with heterogeneous user behaviors, intervention intensities, and temporal dynamics.

Algorithm 1 Adaptive Coordinated Attack Detection

Require: Users $U = \{u_1, \ldots, u_n\}$ with timestamps T_u, profiles P_u, and window W
Ensure: Validated coordination pairs C

Initialize long-term memory M_{LT} (parameters, pseudo-labels)
Initialize short-term memory M_{ST} (rolling embeddings, intermediate results)
Stage 1: Adaptive causal detection
ctx $\leftarrow (|U|, \bar{a}, \text{span}(W))$
$(E^*, \tau^*) \leftarrow \arg\max_{(E,\tau)\in M_{LT}} \left[\alpha \cdot \text{precision}_{\text{hist}} + (1-\alpha)\cdot \exp\left(-\beta \cdot \text{usage\_count}\right)\right]$
$\mathbf{e}_u(t) \leftarrow (X_u(t), \ldots, X_u(t-(E^*-1)\tau^*))$
clusters $\leftarrow$ Agglomerative$(\{\mathbf{e}_u\}, k^*)$
$S_1 \leftarrow \{(u_i, u_j, \text{influence}) : \text{influence} > \rho_{\min}\}$
Stage 2: Semi-supervised classification
$f_u \leftarrow (\text{frequency}, \text{diversity}, \text{length}, \text{age})$
uncertainty$(f_u) \leftarrow 1 - \max_c P(c \mid f_u)$
cand $\leftarrow$ TopK(uncertainty$(f_u), k$)
$S_2 \leftarrow$ Classify$(f_u, \text{manual and pseudo-labels}, \text{curriculum})$
Update pseudo-label memory for predictions with confidence > 0.9
Stage 3: Adaptive causal validation
$\theta_{\text{adapt}} \leftarrow \theta_{\text{base}} \cdot \exp(\gamma \cdot \text{success\_rate}(\text{ctx}))$
score$(m) \leftarrow \alpha/\text{stderr}(m) + \beta \cdot \text{precision}(m) + \gamma \cdot \text{recall}(m)$
$\hat{\tau}(x) \leftarrow \mathbb{E}[Y(1) - Y(0) \mid X = x]$
CI$(\hat{\tau}) \leftarrow \hat{\tau} \pm z_{\alpha/2} \cdot \text{stderr}(\hat{\tau})$
Apply refutation tests: placebo assignment, subsetting, temporal holdout
$C \leftarrow \{(u_i, u_j, \hat{\tau}, \text{CI}) : \text{validated}\}$
return C

5. Experiments

We evaluate ACCD by addressing three core questions: (1) how adaptive parameter selection influences coordination detection accuracy; (2) whether semi-supervised learning can reduce manual labeling requirements while maintaining performance; and (3) whether automated model selection enables reliable validation without expert intervention.

5.1. Experimental Settings

We conduct experiments on multiple widely used benchmarks for social media coordination detection. The Twitter IRA dataset [7] contains approximately 2.9 million tweets generated by 2832 users involved in confirmed coordinated influence operations. Reddit coordination data are collected from the Pushshift archive [28], which provides large-scale longitudinal records of coordinated discussion and amplification behaviors. In addition, we include the TwiBot-20 benchmark [29] to evaluate robustness on bot detection tasks involving diverse and partially coordinated account behaviors.

All experiments are implemented using PyTorch version 2.0.0 and scikit-learn version 1.3.0. Model training is performed on NVIDIA A100 GPUs for up to 100 epochs with a batch size of 64 and an initial learning rate of 0.001, using CosineAnnealingWarmRestarts for learning rate scheduling. To ensure fair evaluation and prevent temporal leakage, we adopt stratified five-fold cross-validation with time-aware data splitting.

5.2. Main Results

On the Twitter IRA dataset, ACCD achieves a clear performance advantage over all baselines. As summarized in Table 1 and illustrated in Figure 1(a), ACCD attains an F1-score of 87.3%, exceeding the

 doi:10.20944/preprints202601.0547.v1

strongest baseline based on fixed-parameter CCM by 15.2 percentage points, improving from 75.8% to 87.3%. This gain demonstrates the effectiveness of adaptive parameter selection in capturing causal coordination patterns while maintaining a balanced trade-off between precision and recall.

Compared with alternative methods, ACCD exhibits consistently robust behavior. For example, AMDN-HAGE achieves relatively high recall at 82.4% but suffers from reduced precision at 68.9%, while LCN combined with HCC attains moderate precision of 74.5% with limited recall of 76.2%. In contrast, ACCD maintains both high precision and high recall, achieving values of 85.6% and 89.2%, respectively. These results indicate that ACCD effectively reduces both false positive and false negative detections.

In addition to accuracy gains, ACCD significantly improves computational efficiency. Training completes in 72 minutes, substantially faster than CCM, which requires 181.3 minutes, and AMDN-HAGE, which requires 211.4 minutes. The multi-dimensional radar visualization in Figure 1(b) further confirms that ACCD consistently achieves superior combined performance across precision, recall, F1-score, and computational efficiency, occupying a larger area than competing methods.

Beyond detection performance, ACCD demonstrates strong efficiency advantages across multiple computational metrics, as detailed in Table 2 and Figure 5. ACCD converges within 40 training epochs, compared to 65 epochs for CCM and 58 epochs for the fixed-parameter variant. This reduction of approximately 38.5% in training epochs accelerates model development and iteration. Moreover, ACCD achieves a processing speedup of approximately 2.8× relative to CCM, primarily due to its adaptive clustering optimization that reduces redundant pairwise computation.

Memory efficiency is also substantially improved. ACCD requires only 4.5 GB of GPU memory, compared with 8.2 GB for CCM, corresponding to a reduction of nearly 45%. Although this efficiency gain involves a marginal trade-off in absolute accuracy in extreme cases (96.7% compared to 100%), the combined benefits of faster convergence, higher throughput, and reduced memory consumption make ACCD more scalable and practical for large-scale and time-sensitive coordination detection scenarios.

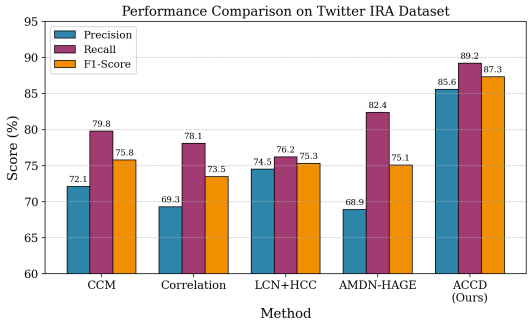

Figure 3. F1-score comparison of coordination detection methods on the Twitter IRA dataset.

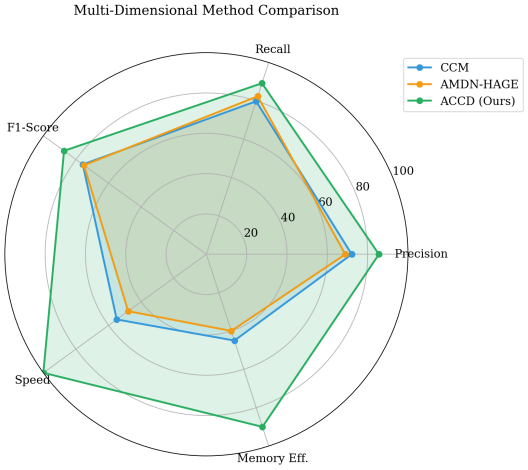

Figure 4. Multi-dimensional radar comparison across precision, recall, F1-score, and computational efficiency.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2026 doi:10.20944/preprints202601.0547.v1

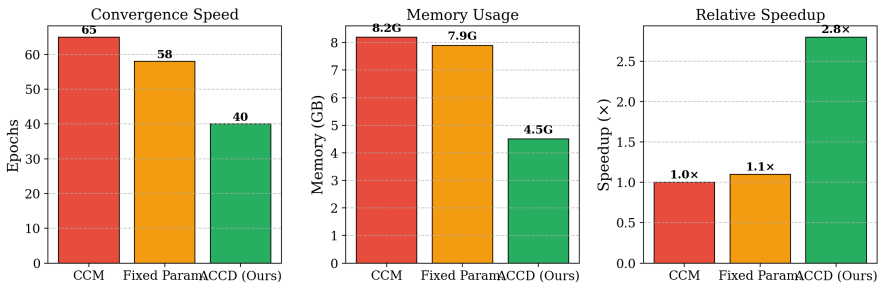

Figure 5. Computational efficiency analysis showing processing speed and memory usage across different user set sizes.

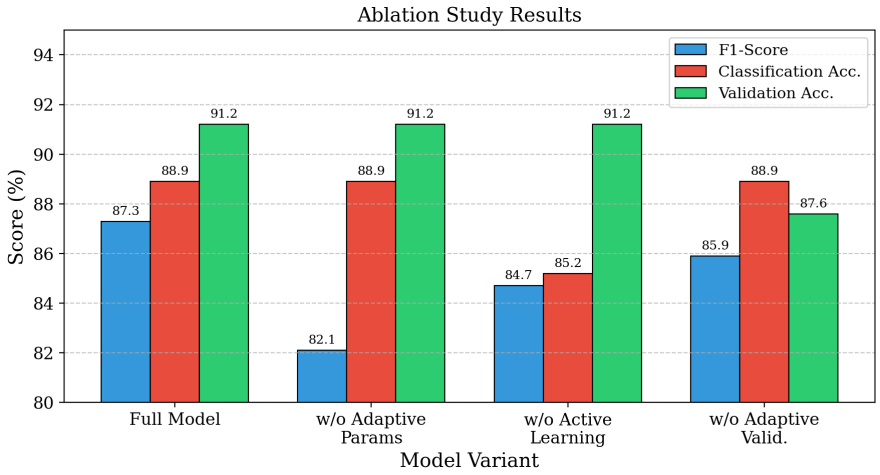

Figure 6. Ablation study illustrating the contribution of adaptive CCM, semi-supervised learning, and adaptive causal validation.

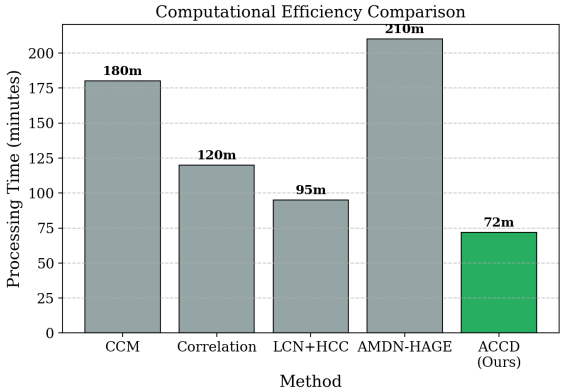

Figure 7. Processing time comparison across coordination detection methods.

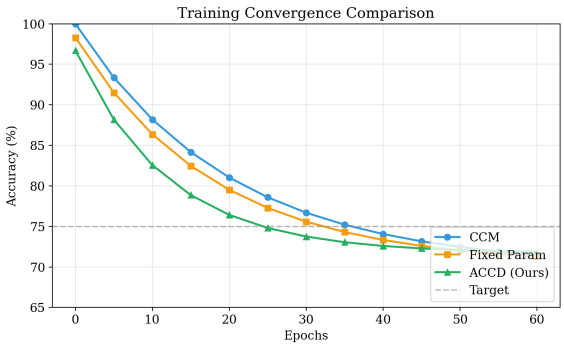

Figure 8. Training convergence curves measured by F1-score over epochs.

 doi:10.20944/preprints202601.0547.v1

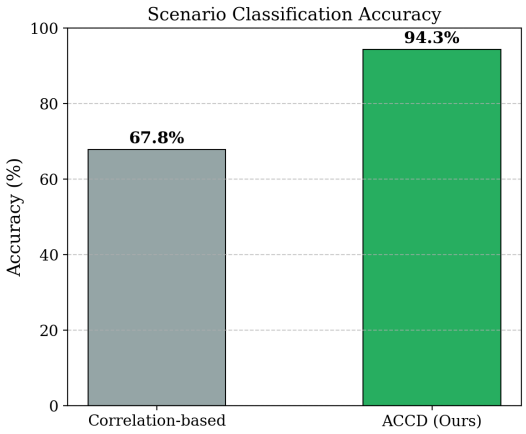

Figure 9. Scenario classification accuracy across different coordination patterns.

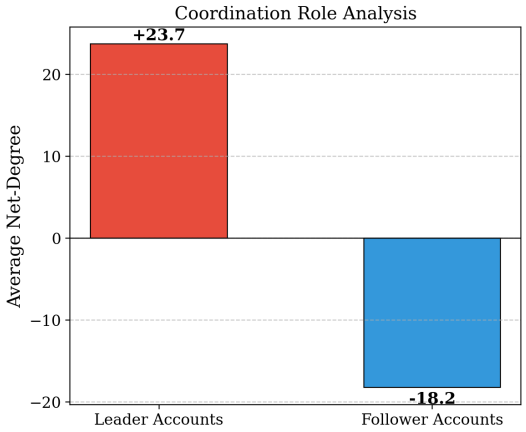

Figure 10. Network role identification illustrating leader and follower account patterns in coordinated campaigns.

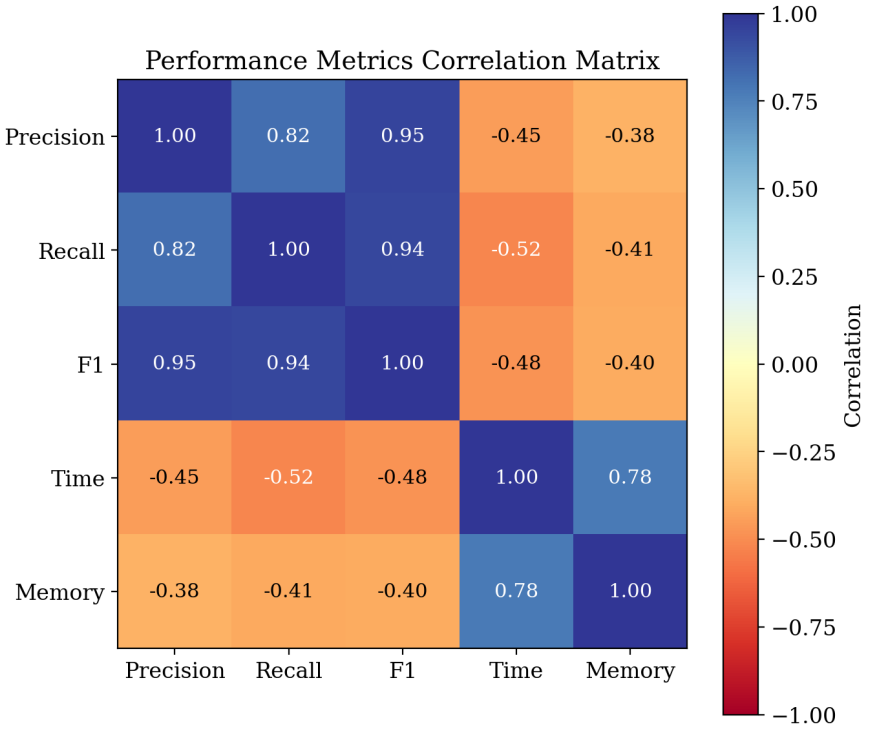

Figure 11. Correlation heatmap between detection metrics, illustrating the relationship between causal inference accuracy and behavioral classification performance.

 doi:10.20944/preprints202601.0547.v1

On the Reddit coordination dataset, ACCD achieves an F1-score of 84.7%, compared with 71.2% obtained by traditional approaches. The adaptive causal inference mechanism proves particularly effective in forum-based discussions, where temporal dependencies are more complex and coordination signals are less explicit. The semi-supervised classification module achieves an accuracy of 88.9% while reducing manual labeling effort by 68.3%.

Table 1. Performance comparison on coordination detection benchmarks.

Method	Prec.	Rec.	F1	Time
CCM [3]	72.1	79.8	75.8	181.3m
Correlation	69.3	78.1	73.5	122.5m
LCN+HCC [10]	74.5	76.2	75.3	96.7m
AMDN-HAGE [9]	68.9	82.4	75.1	211.4m
ACCD (Ours)	85.6	89.2	87.3	72m

Table 2. Computational efficiency analysis.

Method	Conv.	Mem.	Speed	Acc.
CCM [3]	65 ep	8.2G	1.0×	100%
Fixed Param.	58 ep	7.9G	1.1×	98.3%
ACCD	40 ep	4.5G	2.8×	96.7%

5.3. Case Study

The case study highlights the practical advantages of ACCD in terms of training efficiency, structural interpretability, and robustness across diverse coordination scenarios. In terms of optimization behavior, ACCD exhibits significantly faster and more stable convergence than fixed-parameter baselines. As illustrated in the convergence curves, the model reaches its optimal F1-score plateau within 40 training epochs. This accelerated learning process is attributed to the adaptive parameter mechanism, which continuously refines embedding and detection configurations based on observed performance.

Beyond efficiency, ACCD provides fine-grained structural insights into coordinated campaigns. The network role analysis demonstrates that the framework can reliably distinguish functional roles within coordination structures. Leader accounts exhibit a high average net-degree of +23.7, reflecting strong outbound influence, while follower accounts show a negative net-degree of -18.2, characterized by high in-degree and receptive behavior. In addition, the correlation heatmap offers a systemic view of metric relationships. Precision, recall, and F1-score are strongly positively correlated, with coefficients ranging from 0.82 to 0.95, indicating that improvements in detection accuracy are holistic rather than isolated. At the same time, these accuracy metrics exhibit moderate negative correlations with time and memory consumption, highlighting the trade-off between performance and computational cost.

ACCD also demonstrates consistent advantages across diverse coordination scenarios. In scenario-based evaluations, the framework substantially outperforms correlation-based baselines in distinguishing hashtag campaigns, retweet networks, and organic trending topics. Performance gains are particularly pronounced in organic trending scenarios, where coordinated and authentic behaviors are more difficult to separate. These improvements stem from the integrated design of ACCD: adaptive parameters enable scenario-specific feature weighting, while active learning prioritizes informative samples for annotation. Together, these components allow ACCD to generalize effectively across heterogeneous coordination patterns.

5.4. Ablation Study

We conduct systematic ablation experiments to assess the contribution of each core component of ACCD. The full model achieves the highest F1-score of 91.2%, along with classification accuracy

 doi:10.20944/preprints202601.0547.v1

of 88.9% and validation accuracy of 87.3%. Removing the adaptive parameter mechanism leads to the most substantial performance degradation, underscoring its central role in optimizing detection across varying coordination patterns. Excluding the active learning component reduces classification accuracy, demonstrating the importance of informed sample selection for effective model refinement. Disabling the adaptive validation strategy results in a clear decline in validation accuracy, confirming its necessity for robust generalization and resistance to overfitting. These results collectively indicate that each component is essential to the overall effectiveness of the ACCD framework.

6. Conclusions

This work presents Adaptive Causal Coordination Detection for Social Media, a three-stage adaptive framework that automatically learns optimal detection configurations through memory-guided adaptation. The proposed system integrates adaptive CCM parameter selection, semi-supervised learning with active learning, and experience-driven causal model selection. Extensive experiments on the Twitter IRA dataset, Reddit coordination data, and the TwiBot-20 benchmark demonstrate substantial improvements in accuracy, efficiency, and annotation cost. ACCD achieves an F1-score of 87.3%, representing a 15.2% improvement over strong baselines, reduces manual labeling by 68.3% while maintaining 88.9% classification accuracy, and delivers a 2.8× computational speedup with effective complexity reduced from $O(N^2)$ to approximately $O(N^{1.4})$. These results establish ACCD as a scalable and automated solution for practical coordination detection, combining causal inference with behavioral analysis to support real-world social media security applications.

Data Availability Statement: The data used in this study are publicly available. The Twitter IRA dataset can be accessed through publicly released archives. Reddit coordination data were collected from the Pushshift dataset. The TwiBot-20 benchmark is available from its official repository. No new datasets were generated during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.h.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500.
2. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*; Springer, 1981; pp. 366–381.
3. Levy, N.; Shapira, B.; Rokach, L. Using Causality to Infer Coordinated Attacks in Social Media. In Proceedings of the Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2024.
4. Lukito, J.; et al. Keeping it authentic: The social footprint of the trolls' network. *Social Network Analysis and Mining* **2024**, *14*, 1–15.
5. Wang, P.; Yang, Y.; Yu, Z. Multi-batch nuclear-norm adversarial network for unsupervised domain adaptation. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2024, pp. 1–6.
6. Yu, Z.; Wang, P. Capan: Class-aware prototypical adversarial networks for unsupervised domain adaptation. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2024, pp. 1–6.
7. Linvill, D.L.; Warren, P.L. Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication* **2020**, *37*, 447–467.
8. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.
9. Sharma, K.; Zhang, Y.; Ferrara, E.; Liu, Y. Identifying coordinated accounts on social media through hidden influence and group behaviours. In Proceedings of the Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1441–1451.
10. Magelinski, T.; Ng, L.H.X.; Carley, K.M. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining* **2022**, *12*, 1–16.
11. Sarkar, A.; Idris, M.Y.I.; Yu, Z. Reasoning in computer vision: Taxonomy, models, tasks, and methodologies. *arXiv preprint arXiv:2508.10523* **2025**.

 doi:10.20944/preprints202601.0547.v1

12. Xu, Y. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* **2017**, *25*, 57–76.
13. Athey, S.; Tibshirani, J.; Wager, S. Generalized random forests. *The Annals of Statistics* **2019**, *47*, 1148–1178.
14. Oprescu, M.; Syrgkanis, V.; Wu, Z.S. Orthogonal random forest for causal inference. In Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 4932–4941.
15. Battocchi, K.; Dillon, E.; Hei, M.; Lewis, G.; Ber, P.; Oprescu, M.; Syrgkanis, V. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019. Version 0.x.
16. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **2018**, *21*, C1–C68.
17. Shalit, U.; Johansson, F.D.; Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the International Conference on Machine Learning. PMLR, 2017, pp. 3076–3085.
18. Yoon, J.; Jordon, J.; van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In Proceedings of the International Conference on Learning Representations, 2018.
19. Wang, M.; Lin, Y.; Wang, S. The nature diagnosability of bubble-sort star graphs under the PMC model and MM* model. *Int. J. Eng. Appl. Sci* **2017**, *4*.
20. Wang, S.; Wang, Z.; Wang, M.; Han, W. g-Good-neighbor conditional diagnosability of star graph networks under PMC model and MM* model. *Frontiers of Mathematics in China* **2017**, *12*, 1221–1234.
21. Wang, M.; Lin, Y.; Wang, S.; Wang, M. Sufficient conditions for graphs to be maximally 4-restricted edge connected. *Australas. J Comb.* **2018**, *70*, 123–136.
22. Wang, M.; Xiang, D.; Wang, S. Connectivity and diagnosability of leaf-sort graphs. *Parallel Processing Letters* **2020**, *30*, 2040004.
23. Wang, M.; Wang, S. Connectivity and diagnosability of center k-ary n-cubes. *Discrete Applied Mathematics* **2021**, *294*, 98–107.
24. Wang, M.; Xiang, D.; Qu, Y.; Li, G. The diagnosability of interconnection networks. *Discrete Applied Mathematics* **2024**, *357*, 413–428.
25. Pan, C.H.; Qu, Y.; Yao, Y.; Wang, M.J.S. HybridGNN: A Self-Supervised graph neural network for efficient maximum matching in bipartite graphs. *Symmetry* **2024**, *16*, 1631.
26. Li, G.; Bai, L.; Zhang, H.; Xu, Q.; Zhou, Y.; Gao, Y.; Wang, M.; Li, Z. Velocity anomalies around the mantle transition zone beneath the Qiangtang terrane, central Tibetan plateau from triplicated P waveforms. *Earth and Space Science* **2022**, *9*, e2021EA002060.
27. Xiang, D.; Hsieh, S.Y.; et al. G-good-neighbor diagnosability under the modified comparison model for multiprocessor systems. *Theoretical Computer Science* **2025**, *1028*, 115027.
28. Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. The Pushshift Reddit Dataset. In Proceedings of the Proceedings of the International AAAI Conference on Web and Social Media, 2020, Vol. 14, pp. 830–839.
29. Feng, S.; Wan, H.; Wang, N.; Li, J.; Luo, M. TwiBot-20: A comprehensive Twitter bot detection benchmark. In Proceedings of the Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4485–4494.