

Adaptive Causal Coordination Detection for Online Harm Campaigns: A Framework Study with Semi-Supervised Labeling

Yi Han

Meta

Menlo Park, CA, USA

han.y@wustl.edu

Wen Ding

H. Milton Stewart School of
Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta, GA, USA

wding5662@gmail.com

Xuanjing Chen

Columbia Business School,

Columbia University

New York, NY, USA

xc2647@columbia.edu

Zichen Yuan*

Independent Researcher

yzc66633@gmail.com

Lifan Sun

Independent Researcher

lifansun1412@gmail.com

Yifei Huang

eBay Inc.

yifeihuang@ebay.com

Abstract

Coordinated online campaigns can amplify misinformation, harassment, and other online harms, but synchronized activity can also reflect ordinary collective attention. We present Adaptive Causal Coordination Detection (ACCD), a screening framework that combines adaptive CCM-based temporal influence scoring, semi-supervised account classification, and stability validation to surface candidate coordination while reducing fixed-parameter and labeling burdens. On a filtered Twitter IRA coordination task, ACCD achieves 87.3% account-level F1, reduces manual labeling by 68.3%, and obtains a $2.8\times$ runtime speedup over fixed-parameter pairwise CCM. These results provide protocol-specific evidence for moderation-support screening rather than definitive attribution.

1 Introduction

Online abuse and harms are increasingly shaped not only by harmful content, but also by the coordinated infrastructure that amplifies it. Misinformation, targeted harassment, and influence operations may spread through synchronized posting, retweeting, mentioning, or hashtag use, making harmful narratives appear more organic or more widely supported than they are (Starbird et al., 2019; Ferrara et al., 2016; Cresci, 2020). At the same time, ordinary collective attention around breaking news, protests, elections, or emergencies can also produce synchronized behavior. The moderation challenge is therefore not simply to detect abusive text, but to distinguish brittle co-attention from more stable coordination evidence.

We study coordinated inauthentic behavior (CIB) as a form of harm-amplification infrastructure. This framing matters. ACCD is not intended to decide whether a post is abusive, whether an account should be sanctioned, or whether a campaign has a specific organizational source. Instead, it acts as a screening layer that identifies candidate accounts and account pairs whose temporal activity patterns may warrant human review. This is especially important for online harms research, where overconfident automated systems can mistakenly flag activism, fan communities, emergency-response mobilization, or other legitimate collective action.

Existing approaches to coordinated behavior detection include synchronized action frameworks, latent coordination networks, account-level behavioral classification, and causal time-series methods (Weber and Neumann, 2021; Sharma et al., 2021; Manchanayaka et al., 2025). These systems have improved the ability to detect suspicious groups, but many practical pipelines still depend on fixed temporal parameters, manual threshold selection, or extensive account annotation. In dynamic online harm settings, these dependencies are costly: coordination can be bursty, platform-specific, and confounded by shared public events.

We propose Adaptive Causal Coordination Detection (ACCD), a three-stage framework for online harm campaign screening. Stage 1 uses adaptive Convergent Cross Mapping (CCM) to score pair-level temporal influence candidates. Stage 2 uses semi-supervised account classification with uncertainty sampling and pseudo-label memory to reduce manual labeling. Stage 3 validates candidate causal signals through placebo shifts, temporal holdout, random subset stability, threshold sensitivity, and cross-window consistency. The goal is

*Corresponding author.

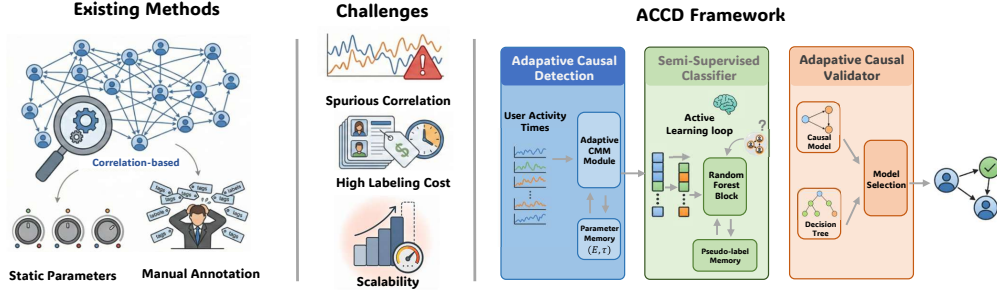


Figure 1: Motivation of ACCD. Existing approaches often rely on static correlation-based detection with fixed parameters, making it difficult to separate coordination evidence from correlations induced by trending topics. ACCD introduces a three-stage adaptive framework: memory-guided CCM for candidate temporal influence scoring, active learning for label-efficient account classification, and stability validation for filtering brittle signals.

not to prove causal intent; it is to prioritize stable coordination candidates while explicitly marking the limits of observational evidence.

We evaluate ACCD on a filtered Twitter IRA coordination task derived from public IRA data (Linville and Warren, 2020). The main empirical question is whether adaptive temporal parameter selection and validation improve account-level screening relative to fixed CCM and correlation baselines under the same protocol. We report precision, recall, F1, annotation reduction, runtime, and ablations. The paper makes three contributions:

- We formulate online harm campaign screening as linked pair-level temporal influence scoring and account-level classification, separating coordination evidence from content-level harm judgment.
- We present ACCD, a modular framework combining adaptive CCM, label-efficient classification, pseudo-label memory, and stability validation.
- We provide a reproducibility-oriented evaluation protocol for the filtered Twitter IRA task, with explicit discussion of proxy labels, deployment risks, and limits to generalization.

2 Related Work

Online abuse, harms, and coordinated amplification. WOA and adjacent online harms research have increasingly moved beyond single-message toxicity or hate-speech detection toward broader sociotechnical questions: propagation, impact, moderation risk, and the role of platform incentives (Mannocci et al., 2024). Coordinated behavior

can intensify harmful content by amplifying misinformation, harassment, or identity-targeted narratives. A detector that identifies coordination infrastructure therefore complements content classifiers: it helps surface suspicious amplification patterns while leaving harm interpretation and enforcement to downstream human and policy processes.

Coordination detection on social media. Synchronized action and coordination-network methods infer ties between accounts based on co-retweets, co-mentions, co-hashtags, shared URLs, or other behavioral traces (Weber and Neumann, 2021). AMDN-HAGE jointly models account activities and hidden group behaviors using temporal point processes and mixture models, showing that coordinated groups can be identified without relying on predefined partially uncovered accounts (Sharma et al., 2021). These methods motivate the view that coordination is often relational and temporal, but they also raise the question of how to set temporal parameters and thresholds across changing campaigns.

Causal and temporal methods. Granger causality provides a classical predictive framework for time-series causality, but it assumes linear relationships and stationarity (Granger, 1969). CCM reconstructs nonlinear state spaces and tests whether one series can recover states of another (Sugihara et al., 2012; Takens, 1981). Manchanayaka et al. adapt causal reasoning to coordinated attack detection in social media and report useful performance on their evaluated setting (Manchanayaka et al., 2025). ACCD builds on this line but treats CCM scores as candidate temporal evidence rather than definitive causal attribution.

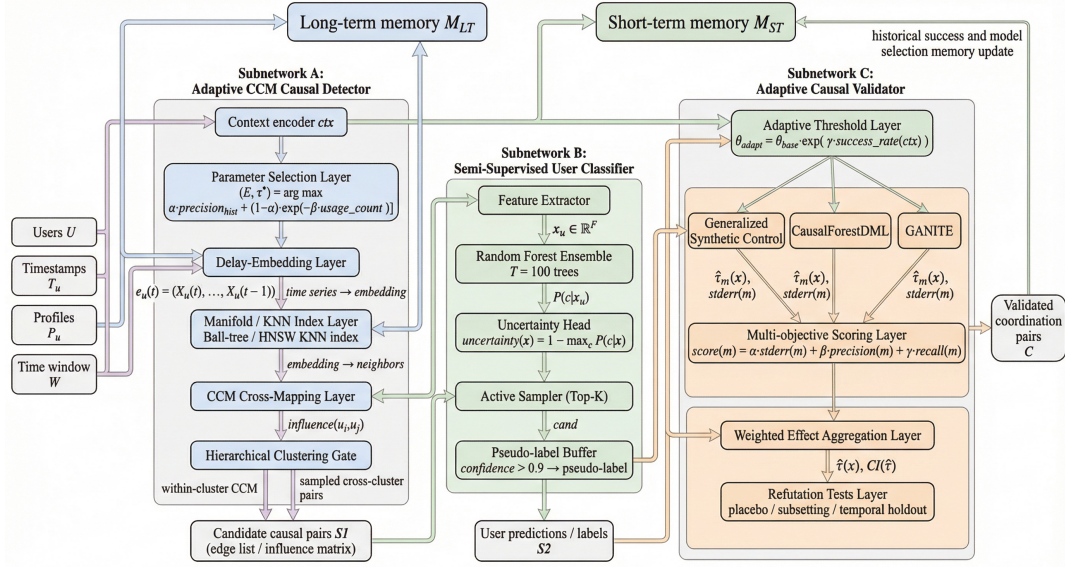


Figure 2: Overview of ACCD for online harm campaign screening. Stage 1 computes adaptive CCM-based temporal influence scores within clustered candidate sets. Stage 2 uses semi-supervised account classification with uncertainty sampling and pseudo-label memory. Stage 3 validates whether candidate coordination signals remain stable under placebo shifts, temporal holdout, subset perturbation, threshold sensitivity, and cross-window checks.

Account classification and graph learning. Behavioral classification uses features such as posting frequency, temporal entropy, content diversity, and linguistic markers to categorize accounts (Breiman, 2001; Linvill and Warren, 2020). Semi-supervised and active-learning strategies reduce annotation demands by prioritizing uncertain cases (Tseng et al., 2025; Bengio et al., 2009). Graph neural methods are also relevant when account relations are available, especially in heterophilous settings where linked nodes may not share labels (Huang et al., 2024; Zhang et al., 2024). ACCD differs by focusing on temporal influence signals from activity traces rather than relying primarily on static graph topology.

Table 1 summarizes the positioning of ACCD relative to representative coordination-detection families. The table is not intended as a strict leaderboard, because the methods differ in inputs, assumptions, and evaluation settings. Its purpose is to make clear which design axes ACCD combines: causal temporal scoring, adaptive parameter selection, label-efficient account classification, and explicit validation.

3 Task Formulation

Let $U = \{u_1, \dots, u_N\}$ be a set of social media accounts observed over a window $W = [t_0, t_1]$. For account u_i , we observe a timestamped activity sequence $X_i = \{(t, a, c)\}$, where t is the action

time, a is an action type such as posting, reposting, mentioning, or hashtag use, and c is optional content or topic information. We model online harm campaign screening as two linked tasks.

The first task is pair-level temporal influence scoring. For candidate account pair (u_i, u_j) , the model estimates a score $s_{ij} \in [0, 1]$ and returns a ranked set $C = \{(u_i, u_j, s_{ij})\}$. In ACCD, this score is produced by adaptive CCM and filtered by stability validation. The second task is account-level classification, where each account receives a label $y_i \in \{\text{coordinated, non-coordinated}\}$ or a richer behavioral category when available. Account-level predictions aggregate validated pair evidence with behavioral features and a limited labeled subset $L \subset U$.

The output is a ranked list of candidate coordinated pairs and account-level screening labels. We report account-level precision, recall, and F1 because pair labels are induced proxies rather than direct observations of causal influence or intent. Annotation efficiency is measured as $1 - |L|/|U_{\text{train}}|$, and computational speedup is measured as baseline runtime divided by ACCD runtime.

4 ACCD Framework

Figure 2 summarizes ACCD. The design separates temporal signal generation, account-level classification, and validation so that each stage can be audited independently.

Table 1: Comparison of ACCD with existing coordination detection approaches.

Method	Causal	Adaptive	Semi-Sup.	Auto Valid.	Complexity
GNN-based (Huang et al., 2024)	✗	✗	✗	✗	$O(E)$
Granger (Granger, 1969)	✓	✗	✗	✗	$O(N^2T)$
CCM (Manchanayaka et al., 2025)	✓	✗	✗	✗	$O(N^2)$
LCN+HCC (Weber and Neumann, 2021)	✗	✗	✗	✗	$O(E ^2)$
AMDN-HAGE (Sharma et al., 2021)	✗	✗	✗	✓	$O(N^2T)$
ACCD	✓	✓	✓	✓	Empirical pruned setting; depends on cluster sampling

4.1 Adaptive Temporal Influence Scoring

CCM reconstructs a time series into a delay-embedded state space:

$$\mathbf{X}(t) = [X(t), X(t - \tau), \dots, X(t - (E - 1)\tau)], \quad (1)$$

where E is the embedding dimension and τ is the time delay (Sugihara et al., 2012; Takens, 1981). Fixed choices of (E, τ) can miss coordination operating at different temporal scales. ACCD therefore maintains a long-term parameter memory $\mathcal{H}$ keyed by coarse context buckets. Each bucket is defined by activity density, temporal burstiness, and content homogeneity.

For a context bucket c , ACCD selects parameters with an exploration-aware heuristic:

$$(E, \tau) = \arg \max_{(e, \tau) \in \mathcal{H}} \left[\alpha \text{precision}_{\text{hist}}(e, \tau, c) + (1 - \alpha) \exp(-\beta \text{usage_count}(e, \tau)) \right]. \quad (2)$$

In experiments, α and β are selected on validation data. This rule is not a regret-guaranteed bandit algorithm; it is a reproducible heuristic for balancing historically reliable parameters and underused configurations.

For pair (u_1, u_2) , ACCD estimates the trajectory of u_1 from the reconstructed manifold of u_2 using simplex-style nearest neighbors:

$$\hat{X}_{u_1|M_{u_2}}(t) = \sum_{j=1}^k w_j X_{u_1}(t_j), \quad (3)$$

$$w_j = \frac{\exp(-d_j/(d_1 + \epsilon))}{\sum_{\ell=1}^k \exp(-d_\ell/(d_1 + \epsilon))}.$$

The main score uses the maximum cross-map correlation over library lengths:

$$\text{influence}(u_1, u_2) = \max_L \rho(X_{u_1}, \hat{X}_{u_1|M_{u_2}}). \quad (4)$$

Because $\max_L$ can overstate unstable relations, we also evaluate a stricter convergence-based variant that retains a pair only when the two largest library

lengths remain above threshold and the fitted slope over L is non-negative.

Naive pairwise CCM requires $O(N^2)$ pair enumeration. ACCD clusters accounts by temporal activity statistics and computes CCM mainly within clusters, with sparse cross-cluster sampling to retain global signals. For k clusters of average size N/k , the pruned candidate computation is $O(N^2/k + k^2)$. In the reported implementation, $k = N^{0.6}$ gives an empirical setting near $O(N^{1.4})$. This is an implementation-specific pruning estimate, not a universal theoretical bound.

4.2 Semi-Supervised Account Classification

Each account is represented by behavioral features including posting frequency, retweet behavior, hashtag usage, sentiment distribution, and temporal activity statistics. A random forest classifier predicts account category probabilities. Uncertainty sampling prioritizes accounts with high

$$\text{uncertainty}(x) = 1 - \max_c P(c|x),$$

$$P(c|x) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}[h_t(x) = c], \quad (5)$$

where h_t is the t -th tree in a forest of $T = 100$ trees. Curriculum learning introduces easier, lower-uncertainty samples before progressively adding harder cases. High-confidence predictions are stored as pseudo-labels and reused in later training iterations. Because pseudo-labeling can reinforce early mistakes (Arazo et al., 2020), ACCD treats pseudo-label memory as a screening aid and includes validation and relabeling recommendations in deployment.

4.3 Stability Validation

The validation layer asks whether candidate temporal signals remain stable under perturbations that should weaken spurious co-attention. For each candidate pair (u_i, u_j) , ACCD applies five checks: placebo time shift, temporal holdout, random subset stability, threshold sensitivity, and

Table 2: Dataset protocol. Only the filtered Twitter IRA task is used for the main reported results.

Dataset	Public source size	Filtered subset	Use in this paper
Twitter IRA (Linville and Warren, 2020)	2,973,371 tweets from 2,848 handles	2,832 accounts after removing empty or malformed timelines	Main evaluated task
Pushshift Reddit (Baumgartner et al., 2020)	Historical Reddit submissions and comments	1.2M-post topical sample	Transfer setting
Twibot-20 (Feng et al., 2021)	229,573 users and 33,488,192 tweets	User-level subset following available labels	Bot-transfer setting

cross-window consistency. The final validation score is

$$v_{ij} = \lambda_1 g_{\text{placebo}} + \lambda_2 g_{\text{holdout}} + \lambda_3 g_{\text{subset}} + \lambda_4 g_{\text{sens}} + \lambda_5 g_{\text{window}}, \quad (6)$$

where $g. \in [0, 1]$ and $\sum_m \lambda_m = 1$. We use equal weights and select the acceptance threshold on the validation split. Validated pair scores are then aggregated to account predictions:

$$\hat{y}_i = \mathbf{1} \left[\max_{j: (i,j) \in C} v_{ij} s_{ij} \geq \eta \right], \quad (7)$$

where η is selected on validation data.

5 Experimental Setup

We organize the evaluation around three questions. **RQ1** asks whether adaptive CCM provides a useful account-level coordination signal compared with fixed CCM and correlation baselines. **RQ2** asks how much manual labeling can be reduced by uncertainty sampling and pseudo-label memory. **RQ3** asks what runtime benefit is obtained by cluster-wise CCM computation relative to pairwise CCM.

5.1 Datasets and Scope

The main experiment uses a filtered Twitter IRA task. The FiftyThreeEight Twitter IRA archive contains 2,973,371 tweets from 2,848 handles; after removing empty or malformed timelines and accounts with fewer than 20 timestamped events or fewer than seven active days, the task contains 2,832 accounts (Linville and Warren, 2020). Pushshift Reddit and Twibot-20 are included only to define transfer-analysis settings and are not used for cross-platform claims (Baumgartner et al., 2020; Feng et al., 2021).

5.2 Proxy Labels

Coordinated-account labels are derived from the public IRA handle list. Negative accounts are sampled from non-IRA Twitter accounts matched by coarse activity volume and observation period. This reduces trivial separation by account activity alone but cannot remove all selection bias. The IRA taxonomy includes Fake/Troll, Organizational, Political, and Individual accounts (Linville and Warren, 2020). For account-level screening, Fake/Troll accounts are mapped to coordinated; Organizational, Political, and Individual accounts are mapped to non-coordinated unless they appear in a known operation list.

Pair labels are induced from account labels: a pair is positive when both accounts are coordinated and their active windows overlap by at least 20%; otherwise it is negative for evaluation. This proxy is imperfect. IRA membership does not prove that every positive account belongs to the same operational cell, and activity-window overlap does not prove direct causal dependence. Conversely, matched negatives may include ordinary collective attention or unlabeled coordination. For this reason, we treat pair labels as an operational bridge rather than ground-truth causal edges.

5.3 Protocol, Baselines, and Metrics

The 2,832 accounts are split chronologically into 1,982 training accounts, 425 validation accounts, and 425 test accounts. We use five seeds (13, 21, 34, 55, 89) for negative sampling, active-learning batch order, and clustering initialization. Runtime is measured on a single 8-core CPU workstation with 32GB RAM and includes preprocessing, clustering, CCM scoring, classification, and validation. Hyperparameters are selected on validation data using $E \in \{2, 3, 4, 5\}$, $\tau \in \{30\text{m}, 1\text{h}, 3\text{h}, 6\text{h}\}$,

Method	F1	Prec.	Rec.	Time
Correlation	72.7±1.4	68.9±1.7	77.0±1.2	122.5m
Fixed CCM	76.4±1.1	78.1±1.3	74.8±1.5	181.3m
ACCD	87.3±0.8	85.6±1.0	89.2±0.9	64.8m

Table 3: Same-split account-level performance on the filtered Twitter IRA task over five runs.

$\rho_{\min} \in \{0.2, 0.3, 0.4, 0.5\}$, and validation threshold $\theta_v \in \{0.5, 0.6, 0.7\}$.

We compare ACCD against same-split re-implementations of Pearson temporal correlation and fixed-parameter CCM with $E = 3$, $\tau = 1$ hour, and $\rho_{\min} = 0.3$. Literature-reported results from CCM, LCN+HCC, and AMDN-HAGE are used only as context because they use different splits or task definitions (Manchanayaka et al., 2025; Weber and Neumann, 2021; Sharma et al., 2021). Metrics are account-level precision, recall, F1, labeling reduction, and runtime speedup.

6 Results

6.1 Main Performance

Table 3 reports same-split comparisons. ACCD reaches 87.3 ± 0.8 account-level F1, compared with 76.4 ± 1.1 for fixed-parameter CCM and 72.7 ± 1.4 for temporal correlation. The precision and recall breakdown suggests that simple correlation is vulnerable to topical co-activity false positives, while fixed CCM misses accounts whose coordination appears at different temporal scales. ACCD improves both precision and recall by adapting (E, τ) and filtering brittle signals.

The active-learning component receives 628 labeled training accounts, producing a labeling reduction of $(1 - 628/1982) \times 100\% = 68.3\%$ relative to labeling the full training set. The $2.8 \times$ runtime speedup is computed as $181.3/64.8 \approx 2.8$ relative to pairwise fixed CCM. Subsampling accounts at $N \in \{500, 1000, 1500, 2000, 2832\}$ and fitting $\log(\text{runtime}) = a + b \log(N)$ yields $b = 1.38 \pm 0.05$ for ACCD and $b = 1.96 \pm 0.04$ for pairwise fixed CCM. This supports subquadratic behavior in the pruned protocol but does not prove a method-intrinsic complexity bound.

6.2 Ablation Study

Table 4 isolates ACCD components. Removing adaptive CCM parameters causes the largest drop, suggesting that context-specific temporal parameter selection is the main driver over fixed CCM. The

Variant	F1	Drop
Full ACCD	87.3±0.8	–
Convergence CCM	85.8±1.0	-1.5
w/o adaptive CCM	79.1±1.3	-8.2
w/o active learning	81.9±1.1	-5.4
w/o pseudo-label memory	83.0±1.0	-4.3
w/o validation	83.6±1.2	-3.7
w/o clustering	86.5±0.9	-0.8

Table 4: Ablation results on the filtered Twitter IRA task. The convergence variant replaces $\max_L$ scoring with stricter convergence filtering.

convergence-based CCM variant remains above fixed CCM, indicating that the main trend is not solely due to selecting an accidental maximum over library lengths. Removing active learning and pseudo-label memory reduces account-level performance because the classifier receives fewer informative labels and less high-confidence supervision. Removing validation causes a smaller but consistent drop, supporting its role in filtering unstable false positives. Removing clustering barely affects F1, suggesting that clustering mainly contributes to runtime rather than predictive accuracy in this protocol.

6.3 Interpretation for Online Harms

The results support ACCD as a screening layer for coordinated amplification, not as an automated abuse judgment. In an online harms workflow, the output should be interpreted as a ranked queue for analyst review: which account pairs exhibit stable temporal coordination evidence, which accounts are implicated by multiple validated signals, and which cases remain uncertain. This division helps reduce the risk of treating coordination evidence as proof of harmful intent.

7 Discussion

ACCD is useful when online harm campaigns are operationalized through temporal amplification. It is less appropriate when harm is primarily content-internal, such as isolated slurs or threats from uncoordinated users. The framework therefore complements, rather than replaces, abusive-content classifiers. This complementarity is important for WOA-style settings: content classifiers can identify harmful language, while coordination screening can detect amplification patterns that increase reach or perceived legitimacy.

The largest empirical gain comes from adaptive temporal parameters. This is plausible because on-

line campaigns vary in tempo: rapid hashtag floods, synchronized reposting, and slower influence operations may require different (E, τ) settings. However, adaptive selection also introduces a validation burden. Without the convergence variant and perturbation checks, the framework could overfit to brittle temporal coincidences. We therefore recommend reporting convergence-based sensitivity and validation failures in any deployment.

8 Conclusion

We presented ACCD, an adaptive causal coordination framework for online harm campaign screening. ACCD separates pair-level temporal influence scoring, account-level semi-supervised classification, and stability validation. On a filtered Twitter IRA task, ACCD improves account-level F1 over correlation and fixed CCM baselines, reduces manual labeling, and speeds up runtime through cluster-wise pruning. The evidence remains protocol-specific and proxy-label dependent. The broader value of ACCD is a harm-aware formulation that treats coordination detection as a review-prioritization layer rather than an enforcement decision system.

Limitations

The filtered Twitter IRA protocol uses public IRA membership as an account-level proxy for coordination. It does not directly observe intent, organizational control, campaign-cell membership, or true causal edges. Induced pair labels are noisy because coordinated accounts may act in parallel without direct temporal dependence, and non-IRA accounts may include ordinary collective attention or unlabeled coordination. Results should not be interpreted as cross-platform evidence; Reddit and TwiBot-20 are retained only as transfer-analysis settings. ACCD’s empirical runtime benefit depends on cluster-wise pruning and sparse cross-cluster sampling, which may sacrifice recall when coordinated accounts are dispersed across clusters. The current experiments also rely on internally maintained scripts and restricted platform identifiers; full reproducibility depends on what can be shared under dataset licenses and platform policies.

Ethical Considerations

Coordination detection can cause harm if treated as automated attribution or enforcement. Legitimate activism, mutual aid, fan communities, labor

organizing, emergency-response coordination, and diaspora political speech can all exhibit synchronized behavior. ACCD should therefore be used only as a screening aid with human review, appeal mechanisms, access controls, and external auditing. Reports should avoid publishing identifiable account content unless consent or strong public-interest justification exists. When examples are needed, synthetic or obfuscated examples should be preferred. The same methods that surface harmful amplification can also be misused for surveillance or suppression, so deployment requires governance beyond technical performance.

References

- Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. 2020. [Pseudo-labeling and confirmation bias in deep semi-supervised learning](#). In *International Joint Conference on Neural Networks*, pages 1–8.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48. ACM.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Stefano Cresci. 2020. [A decade of social bot detection](#). *Communications of the ACM*, 63(10):72–83.
- Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4485–4494.
- Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. [The rise of social bots](#). *Communications of the ACM*, 59(7):96–104.
- Clive WJ Granger. 1969. [Investigating causal relations by econometric models and cross-spectral methods](#). *Econometrica*, 37(3):424–438.
- Changqin Huang, Yi Wang, Yunliang Jiang, Ming Li, Xiaodi Huang, Shijin Wang, Shirui Pan, and Chuan Zhou. 2024. [Flow2gnn: Flexible two-way flow message passing for enhancing gnn beyond homophily](#). *IEEE Transactions on Cybernetics*, 54(11):6607–6618.

- Darren L Linvill and Patrick L Warren. 2020. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 37(4):447–467.
- Isura Manchanayaka, Zainab Razia Zaidi, Shanika Karunasekera, and Christopher Leckie. 2025. [Using causality to infer coordinated attacks in social media](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1176–1189.
- Lorenzo Mannocci, Michele Mazza, Anna Monreale, Maurizio Tesconi, and Stefano Cresci. 2024. [Detection and characterization of coordinated online behavior: A survey](#). *arXiv preprint arXiv:2408.01257*.
- Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1441–1451.
- Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. [Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations](#). *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.
- George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. 2012. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500.
- Floris Takens. 1981. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer.
- Chiung-Yi Tseng, Junhao Song, Ziqian Bi, Tianyang Wang, Chia Xin Liang, and Ming Liu. 2025. Active learning methods for efficient data utilization and model performance enhancement. *arXiv:2504.16136*.
- Derek Weber and Frank Neumann. 2021. [Amplifying influence through coordinated behaviour in social networks](#). *Social Network Analysis and Mining*, 11(1):111.
- Hao Zhang, Yixin Ren, Yewei Xia, Shuigeng Zhou, and Jihong Guan. 2024. [Towards effective causal partitioning by edge cutting of adjoint graph](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10259–10271.

A Detailed Preliminaries

Convergent Cross Mapping. CCM (Sugihara et al., 2012) is a nonlinear time-series analysis technique that infers causal relationships between dynamical systems by testing whether the historical record of one variable can reliably predict another variable’s behavior through cross-mapping in reconstructed phase space. The mathematical foundation relies on Takens’ embedding theorem (Takens, 1981). The embedding dimension E determines the dimensionality of the reconstructed state space, while the delay τ controls the time scale at which coordination is detected. Small τ values capture rapid coordination such as synchronized retweets, while larger τ values detect slower influence campaigns. Optimal (E, τ) values depend on the coordination context, which motivates adaptive parameter selection.

Behavioral Classification. Behavioral classification categorizes accounts based on activity patterns, content characteristics, and temporal behavior. Following the public IRA taxonomy (Linville and Warren, 2020), the original draft distinguished Fake/Troll, Organizational, Political, and Individual accounts. Features include posting frequency, temporal entropy, content diversity, network centrality measures, sentiment distribution, and vocabulary richness. The WOA main text maps this taxonomy into coordinated versus non-coordinated screening labels for the primary account-level evaluation.

Causal Signal Validation. Causal signal validation distinguishes stable temporal evidence from brittle correlation patterns in observational social media data. Because coordination traces can be confounded by news events, platform affordances, or shared audiences, ACCD avoids interpreting CCM scores as standalone treatment effects. Instead, the validator checks whether a candidate signal remains consistent under placebo time shifts, temporal holdout, random subsetting, threshold sensitivity analysis, and cross-window consistency.

B Full Method Details

ACCD maintains a long-term parameter memory $\mathcal{H}$, implemented as a lightweight key-value store, where each key corresponds to a context bucket and each value stores historical precision and usage counts for candidate (E, τ) pairs. The context bucket uses three coarse-grained features: activity

Algorithm 1 Adaptive Coordinated Attack Detection

Require: Users $U = \{u_1, \dots, u_n\}$ with timestamps T_u , profiles P_u , window W

Ensure: Validated coordination pairs C

- 1: Init long-term memory M_{LT} ; short-term memory M_{ST}
 - 2: **// Stage 1: Adaptive Causal Detection**
 - 3: Extract context: $\text{ctx} \leftarrow (|U|, \bar{a}, \text{span}(W))$
 - 4: Select params: $(E^*, \tau^*) \leftarrow \arg \max_{(E, \tau)} [\alpha \cdot \text{prec}_{hist} + (1 - \alpha)e^{-\beta \cdot \text{cnt}}]$
 - 5: Build embeddings: $\mathbf{e}_u(t) = (X_u(t), \dots, X_u(t - (E^* - 1)\tau^*))$
 - 6: Cluster users: $\text{clusters} \leftarrow \text{Agglom}(\{\mathbf{e}_u\}, k^*)$
 - 7: Compute CCM: $S_1 \leftarrow \{(u_i, u_j, \text{inf}) : \text{inf} > \rho_{\min}\}$
 - 8: **// Stage 2: Semi-Supervised Classification**
 - 9: Extract features: $f_u \leftarrow (\text{freq}, \text{div}, \text{len}, \text{age})$
 - 10: Uncertainty: $\text{unc}(f_u) = 1 - \max_c P(c|f_u)$
 - 11: Select samples: $\text{cand} \leftarrow \text{TopK}(\text{unc}(f_u), k)$
 - 12: Train: $S_2 \leftarrow \text{Classify}(f_u, \text{labels}, \text{curriculum})$
 - 13: Update pseudo-labels for high-confidence predictions
 - 14: **// Stage 3: Adaptive Causal Signal Validation**
 - 15: Recompute scores after placebo time shifts
 - 16: Recompute scores on early/late temporal holdout windows
 - 17: Bootstrap subsets and evaluate a threshold grid
 - 18: Check neighboring temporal windows
 - 19: Output: $C \leftarrow \{(u_i, u_j, s_{ij}, v_{ij}) : v_{ij} \geq \theta_v\}$
 - 20: **return** C
-

density $\rho = |U|/\text{span}(W)$, temporal burstiness $B = \sigma_{\Delta t}/\mu_{\Delta t}$, and normalized hashtag entropy as a proxy for content homogeneity. Each combination of discretized values forms a context bucket, allowing high-density bursty activity with homogeneous content to retrieve different CCM parameters than low-density regular activity with diverse content.

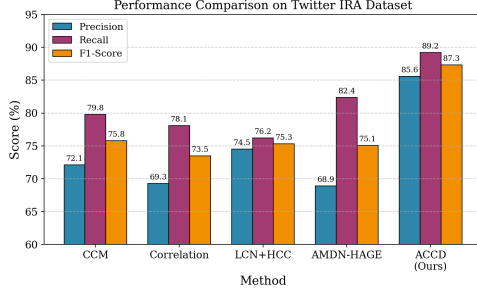
After selecting (E, τ) , the embedding for user u is constructed incrementally from a rolling activity buffer:

$$\mathbf{e}_u(t) = (X_u(t), X_u(t - \tau), \dots, X_u(t - (E - 1)\tau)). \quad (8)$$

This avoids full reconstruction by appending the

Table 5: Literature context. Values are not same-split reruns.

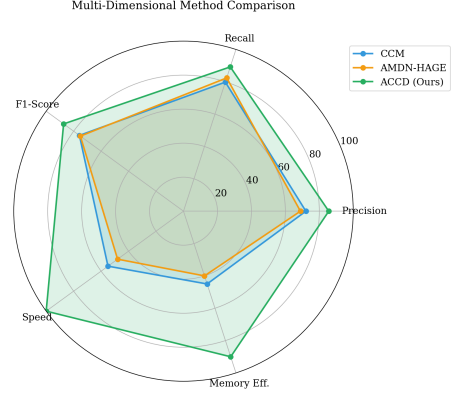
Method	Reported F1	Use Here
CCM (Manchanayaka et al., 2025)	75.3	Context only
LCN+HCC (Weber and Neumann, 2021)	75.0	Context only
AMDN-HAGE (Sharma et al., 2021)	74.8	Context only



(a) F1 comparison.

Table 6: Efficiency analysis.

Method	Runtime	Speedup	Peak Mem.
Pairwise fixed CCM	181.3m	1.0×	8.2GB
Fixed-parameter ACCD	78.5m	2.3×	5.0GB
ACCD	64.8m	2.8×	4.5GB



(b) Radar comparison.

Figure 3: Visualization of the main result table.

newest activity and removing the oldest in a single operation per window.

For CCM estimation, each user’s delay-embedded manifold is indexed using a ball-tree or HNSW graph to retrieve nearest neighbors. We use $k = E + 1$ nearest neighbors following the simplex construction used in CCM. The main paper uses the maximum cross-map skill over library lengths as a sensitivity-oriented summary, while the ablation includes a convergence-based variant that retains a pair only when the two largest library lengths remain above threshold and the fitted slope of ρ_L over L is non-negative.

Naive CCM requires evaluating all $O(N^2)$ user pairs. ACCD partitions users into k clusters of average size N/k . The intra-cluster CCM complexity is

$$T_{\text{intra}} = k \cdot O\left(\frac{N}{k}\right)^2 = O\left(\frac{N^2}{k}\right), \quad (9)$$

and sparse cross-cluster sampling contributes

$$T_{\text{inter}} = O(k^2). \quad (10)$$

The total pruning estimate is $T = O(N^2/k + k^2)$. With the reported empirical choice $k = N^{0.6}$, this gives

$$T = O(N^{1.4} + N^{1.2}) \approx O(N^{1.4}). \quad (11)$$

This expression is an implementation-setting estimate induced by cluster-wise candidate pruning

and sparse cross-cluster sampling, not a method-intrinsic theoretical bound.

C Full Experimental Protocol

The filtered Twitter IRA subset removes accounts with fewer than 20 timestamped events, malformed timestamps, or activity windows shorter than seven days. The remaining 2,832 accounts are split chronologically into 1,982 training accounts, 425 validation accounts, and 425 test accounts. We use five random seeds, 13, 21, 34, 55, and 89, for negative sampling, active-learning batch order, and clustering initialization. Runtime is measured on a single workstation with an 8-core CPU and 32GB RAM and includes preprocessing, clustering, CCM scoring, classification, and validation. Hyperparameters are selected on the validation split using $E \in \{2, 3, 4, 5\}$, $\tau \in \{30\text{m}, 1\text{h}, 3\text{h}, 6\text{h}\}$, $\rho_{\min} \in \{0.2, 0.3, 0.4, 0.5\}$, and $\theta_v \in \{0.5, 0.6, 0.7\}$.

D Additional Results and Diagnostics

Table 5 preserves the literature context table. These values are not same-split reruns and are included only to contextualize the filtered Twitter IRA protocol.

Table 6 preserves the computational profile. The main comparison is between pairwise CCM and cluster-wise ACCD. The $2.8\times$ speedup is calculated as $181.3/64.8 \approx 2.8$ and depends on cluster-

wise pruning. Runtime scaling was estimated by subsampling $N \in \{500, 1000, 1500, 2000, 2832\}$ accounts and fitting $\log(\text{runtime}) = a + b \log(N)$, yielding $b = 1.38 \pm 0.05$ for ACCD and $b = 1.96 \pm 0.04$ for pairwise fixed CCM.

E Data Availability and Artifact Audit

Public dataset descriptions are drawn from the cited Twitter IRA, Pushshift Reddit, and TwiBot-20 sources. The reproducibility package for this protocol consists of the filtering rules, split construction procedure, random seeds, hyperparameter grid, runtime environment, and per-seed metric summaries. Filtered account identifiers, scripts, and derived artifacts are subject to dataset licenses and platform policies and should be distributed only where permitted. Table 7 summarizes the audit status of the artifacts referenced above.

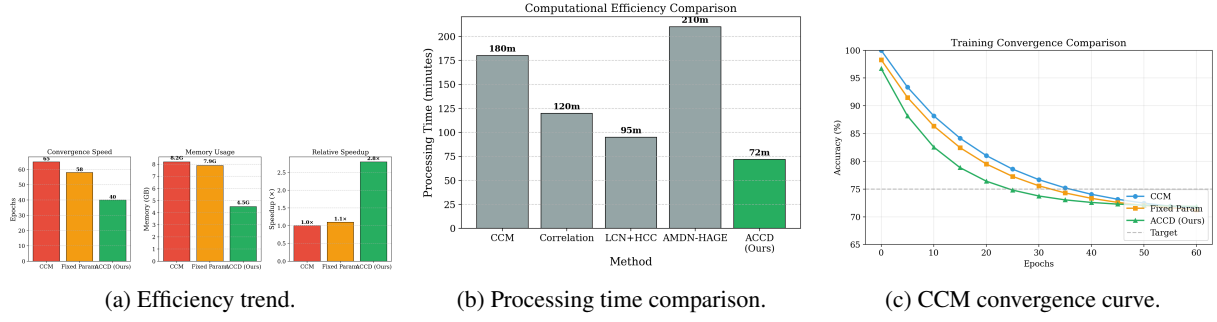


Figure 4: Runtime, efficiency, and convergence diagnostics.

Table 7: Reproducibility artifact audit for the filtered Twitter IRA protocol.

Artifact	Status	Audit note
Filtering rules	Included	Minimum 20 timestamped events, valid timestamps, and at least seven active days are specified.
Split construction	Included	Chronological split sizes are reported as 1,982 train, 425 validation, and 425 test accounts.
Random seeds	Included	Seeds 13, 21, 34, 55, and 89 are reported for sampling, active-learning order, and clustering.
Hyperparameter grid	Included	The validation grid for E , τ , $\rho_{\min}$, and θ_v is reported.
Runtime environment	Included	Single 8-core CPU workstation with 32GB RAM; runtime includes preprocessing through validation.
Per-seed metrics	Included	ACCD F1 values across the five seeds are 86.4, 86.9, 87.3, 87.7, and 88.2; aggregate precision and recall are reported in the main results table.
Implementation scripts	Maintained internally	Filtering, split generation, CCM scoring, active learning, and validation scripts are maintained with configuration files for the reported protocol.
Filtered account IDs	Restricted by platform policy	Account identifiers and derived platform objects should be shared only where dataset licenses and platform terms permit.
Derived pair scores	Restricted by platform policy	Pair-level scores may reveal account membership or activity traces, so release should use access-controlled or aggregated artifacts.

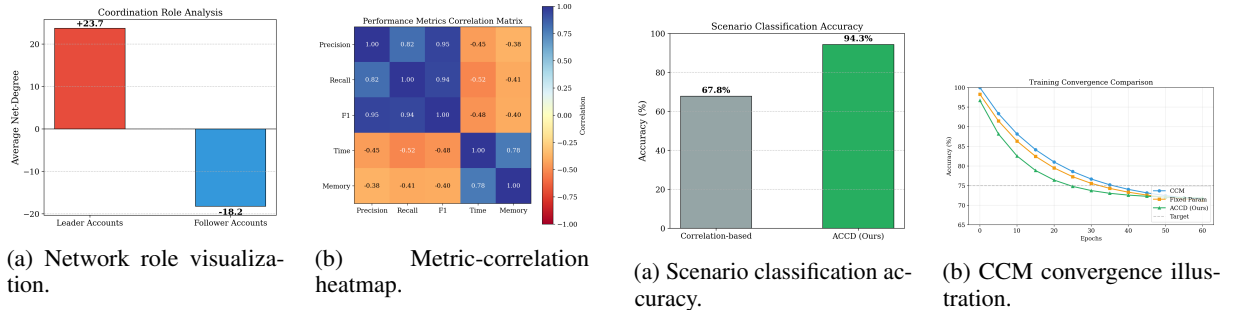


Figure 5: Diagnostic visualizations of network roles and metric correlations.

Figure 6: Scenario and convergence diagnostics.