

Detecting Coordinated Inauthentic Behavior on Social Media with Memory-Guided Causal Inference

Yi Han¹, Wen Ding², Xuanjing Chen³, Zhiguo Tao^{4,*}

¹Meta, Menlo Park, CA 94025, USA

²H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA 30332, USA

³Columbia Business School, Columbia University, New York, NY 10027, USA

⁴University of Nottingham, Nottingham, UK

han.y@wustl.edu wding5662@gmail.com

xc2647@columbia.edu ssyzt3@outlook.com

Abstract

Coordinated inauthentic behavior (CIB) threatens the integrity of public conversation, yet existing detectors remain unreliable in practice. Most systems measure correlation between account activities, which conflates genuine influence with shared reactions to trending topics. They also apply fixed detection parameters and require extensive manual labeling, so configurations tuned on one campaign rarely transfer to another. We present ACCD (Adaptive Causal Coordination Detection), a framework that treats this problem as memory-guided causal inference. ACCD screens account pairs in three stages that each narrow the candidate set: it retrieves convergent-cross-mapping parameters from a context-indexed memory of past performance to estimate causal influence, concentrates annotation on uncertain accounts through active learning in a semi-supervised classifier whose labels filter the surviving pairs, and validates the remainder with estimators and thresholds chosen from prior experience. On the Twitter Internet Research Agency (IRA) dataset, ACCD improves pair-level F1 from 75.8% to 87.3% over a fixed-parameter causal baseline, reduces manual labeling by 68.3%, and cuts end-to-end detection time by 2.5 $\times$. On Reddit coordination traces it reaches 84.7% F1. On synthetic data with known causal structure, it recovers true influence edges at 0.934 area under the ROC curve (AUC) against 0.708 for correlation-based detection, and the gap widens to 0.861 against 0.586 once strong confounding is introduced. These results suggest that adapting causal configurations to detection context provides a practical and efficient screening tool for protecting online public discourse.

1 Introduction

Coordinated inauthentic behavior (CIB) uses multiple accounts to manipulate public discourse

* Corresponding author.

Existing Approaches



ACCD (Ours)

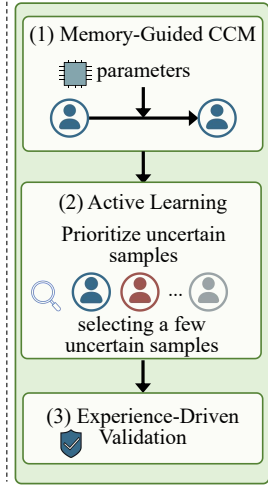


Figure 1: Motivation of ACCD. Existing approaches (left) rely on static, correlation-based parameters and cannot separate genuine causal influence from spurious correlation induced by confounders. ACCD (right) adapts detection configurations from accumulated experience rather than fixing them by hand.

through synchronized posting, amplification, or harassment (Linville and Warren, 2020). Recurring signals include short-interval co-retweet/co-mention bursts (Pacheco et al., 2021; Weber and Neumann, 2020), reused hashtags or near-duplicate templates (Alizadeh et al., 2020; George et al., 2025), and division of labor between seeding and amplifying accounts (Chen, 2026; Thiele et al., 2025). These patterns span co-activity, content similarity, and temporal synchrony (Mannocci et al., 2024); ACCD targets the temporal axis by asking whether one account’s activity drives another’s. Reliable detection supports election integrity, public-health communication, and online safety (Lazer et al., 2018; Varol et al., 2017; Cresci et al., 2017), aligning with NLP for positive social impact (D. Lewis et al., 2025).

Prior approaches use CCM for temporal influ-

ence (Sugihara et al., 2012), behavior classifiers for account screening (Breiman, 2001; Cresci et al., 2018), or causal machine learning for estimator selection (Chen and Jing, 2025). They remain vulnerable to confounding and often rely on fixed embeddings, heavy labeling, or fixed statistical thresholds that transfer poorly across campaigns. We propose Adaptive Causal Coordination Detection (ACCD), a memory-guided three-stage pipeline that retrieves context-specific CCM parameters, filters candidate pairs using an active-learning semi-supervised classifier, and validates survivors with experience-selected causal estimators and thresholds (Figure 1).

We evaluate ACCD on Twitter IRA, Reddit coordination traces, and synthetic data. The contributions are:

1. A context-indexed memory that adapts causal configurations to coordination context, helping separate influence from confounded correlation.
2. A sequential three-stage design combining adaptive CCM, semi-supervised role filtering, and experience-driven validation; hierarchical clustering reduces pairwise evaluation from $O(N^2)$ to $O(N^{1.4})$.
3. On Twitter IRA, ACCD improves pair-level F1 by 11.5 points over CCM, reduces labeling by 68.3%, and cuts end-to-end time by 2.5 $\times$; it reaches 84.7% F1 on Reddit and 0.934/0.861 AUC on synthetic data under low-noise/strong-confounding conditions.

2 Related Work

2.1 Causal and Time-Series Detection

CCM (Sugihara et al., 2012) reconstructs time-series state spaces via Takens’ theorem (Takens, 1981) and has been used for cross-mapped temporal influence (Isufaj et al., 2025). Granger causality (Granger, 1969) is simpler but assumes linear, stationary dynamics. Standard CCM avoids those assumptions yet fixes (E, τ) and a threshold and costs $O(N^2)$ pairwise. Graph neural networks capture relational structure (Wu et al., 2026; Guyan et al., 2025) but require explicit graphs and model temporal coordination indirectly. ACCD instead adapts temporal causal parameters from prior detection experience.

2.2 Behavioral and Network Classification

Behavioral methods classify accounts from activity features. Social-footprint methods (Cresci et al., 2018; Wang et al., 2025; Zhou et al., 2025) and Random Forest classifiers based on the troll taxonomy of Linvill and Warren (2020) can identify influence-operation accounts, while AMDN-HAGE (Sharma et al., 2021) models temporal group behavior and LCN+HCC (Pacheco et al., 2021) builds coordination graphs from co-retweet, co-tweet, and co-mention signals. Their labeling or graph-construction burden can hinder rapid transfer to new campaigns.

2.3 Automated Causal Inference and Positioning

Automated causal learning supports estimator selection and treatment-effect estimation. Double machine learning debiases flexible nuisance models (Chernozhukov et al., 2018; Colangelo and Lee, 2026); related work studies heterogeneous effects and policy optimization (Chen and Jing, 2025; Nizam et al., 2025; Imai and Li, 2025), and surveys cover instrumental-variable and representation-based estimators (Wu et al., 2025; Abécassis et al., 2025). These pipelines remain largely domain agnostic. ACCD combines causal inference with context adaptation, semi-supervised filtering, and experience-driven validation at $O(N^{1.4})$ cost.

3 Problem Formulation and Preliminaries

3.1 Problem Formulation

Observational unit. The observation period is divided into non-overlapping detection windows of fixed length $|W|$. An account is *co-active* in a window if it produces at least $n_{\min}$ posts inside it, and the observational unit is an unordered pair (u, v) of accounts co-active in the same window. Accounts active in different windows are never compared. Within a window, candidate pairs are restricted to the hierarchical clusters described in Section 4.1, supplemented by a small random sample of cross-cluster pairs, so the evaluated candidate set is a strict subset of all $\binom{N}{2}$ pairs.

Target variable. Each candidate pair carries a binary label $y(u, v) \in \{0, 1\}$. We set $y = 1$ when both accounts are attributed to a campaign confirmed as coordinated and are co-active in the same window, and $y = 0$ when at least one member is an authentic account. The system predicts $\hat{y}(u, v)$ and, for predicted-positive pairs, additionally reports a

causal effect estimate with a confidence interval. Detection quality is thus binary classification over pairs. The four-way behavioral labels produced in Stage 2 are an intermediate account-level representation, not the evaluation target; they are reported separately as an auxiliary metric.

Decision rule. A candidate pair is returned as coordinated only when all three stages agree:

$$\hat{y}(u, v) = \mathbf{1}[\text{influence}(u, v) > \theta_{\text{adapt}}] \cdot \mathbf{1}[r(u) \in \mathcal{R} \vee r(v) \in \mathcal{R}] \cdot \mathbf{1}[\text{valid}(u, v)] \quad (1)$$

The first indicator uses the Stage 1 influence score of (5) against the Stage 3 adapted threshold of (8). The second requires that at least one member receives a campaign-plausible Stage 2 label, where $\mathcal{R}$ contains the fake-or-troll and organizational categories. The third, $\text{valid}(u, v)$, holds when the Stage 3 causal effect is significant and survives every refutation test.

Causal quantities. Stage 3 evaluates the ordered direction $u \rightarrow v$. The treatment $T_u(t) \in \{0, 1\}$ equals one when u 's posting rate in time slice t exceeds α times its own historical baseline; the outcome $Y_v(t) \in \mathbb{Z}_{\geq 0}$ is v 's activity count over the following Δ hours; and the covariate vector $\mathbf{x}_{\text{cov}}$ collects window-level confounders such as topic popularity, overall window volume, and each account's baseline activity. The estimand is the conditional average treatment effect of (10). Identification relies on the usual assumptions of unconfoundedness given $\mathbf{x}_{\text{cov}}$, overlap, and the stable unit treatment value assumption. The refutation tests of Stage 3 probe violations of the first.

3.2 Technical Background

We briefly review the three technical foundations on which ACCD is built.

Convergent Cross Mapping (CCM) (Sugihara et al., 2012) infers causal influence between dynamical systems by testing whether the history of one variable can predict another's through cross-mapping in a reconstructed phase space. Let $X_u(t) \in \mathbb{Z}_{\geq 0}$ denote the activity count of account u in discrete time step t , with embedding dimension $E \in \mathbb{Z}^+$ and time delay $\tau \in \mathbb{Z}^+$ measured in time steps of one hour. Following Takens' embedding theorem (Takens, 1981), $X_u(t)$ is lifted into

the delay-embedding vector $\mathbf{X}_u(t) \in \mathbb{R}^E$:

$$\mathbf{X}_u(t) = [X_u(t), X_u(t-\tau), \dots, X_u(t-(E-1)\tau)]. \quad (2)$$

Small τ captures rapid synchronization such as coordinated retweets, while larger values expose slower influence campaigns. Because the best (E, τ) depends on the coordination context, we select them adaptively; Section 5 evaluates this choice empirically.

Behavioral classification assigns accounts to activity-based categories from observable features such as posting frequency, temporal regularity, content diversity, and linguistic markers. Following Linvill and Warren (2020) we distinguish four types, fake or troll, organizational, political, and individual, which are the labels predicted in Stage 2.

Causal inference separates correlation from causation in observational data, a distinction that matters here because shared trending topics routinely induce spurious correlations among unrelated accounts. Standard pipelines fix significance thresholds such as $p < 0.05$ and generalize poorly across datasets with different activity dynamics; ACCD's Stage 3 validator replaces them with dataset-specific settings learned from past detection experience.

4 Method

ACCD addresses coordination detection through three sequential stages (Figure 2), summarized in Algorithm 1. Stage 1, the Adaptive Causal Coordination Detector, infers temporal causal influence through memory-guided parameter selection. Stage 2, the Active-Learning Semi-Supervised Classifier, reduces annotation cost by directing human effort toward uncertain accounts. Stage 3, the Adaptive Causal Validator, automates estimator selection and threshold determination from accumulated detection experience.

The three stages are strictly sequential, and each one narrows the candidate set. Stage 1 converts raw activity streams into a candidate set S_1 of account pairs whose CCM influence exceeds a loose screening threshold $\rho_{\min} = 0.25$. Stage 2 labels the individual accounts appearing in S_1 and discards every pair in which neither member carries a campaign-plausible role, producing $S_{12} \subseteq S_1$. Stage 3 estimates a causal effect for each surviving pair and keeps only those that clear the adapted

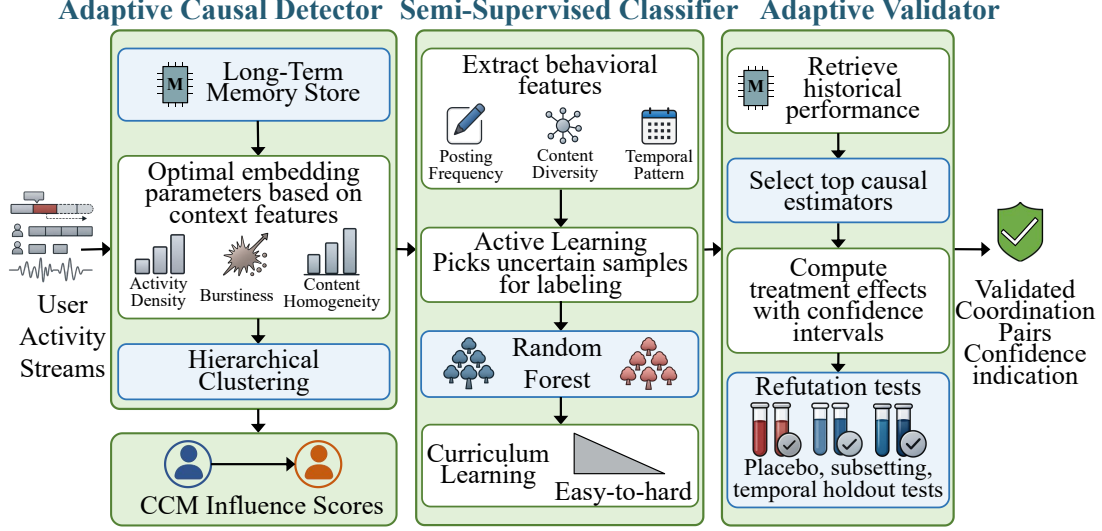


Figure 2: Architecture of ACCD. Stage 1 retrieves context-appropriate (E, τ) from memory and computes CCM influence within hierarchical clusters; Stage 2 classifies accounts in a semi-supervised manner, using active learning to label only uncertain cases; Stage 3 selects causal estimators and thresholds from past experience and validates candidate links with refutation tests (placebo, subsetting, temporal holdout).

significance threshold and pass all refutation tests. The final output is therefore the conjunction of (1): no pair enters the returned set $\mathcal{P}$ without passing all three stages, and each stage can only remove pairs, never add them.

4.1 Stage 1: Adaptive Causal Coordination Detector

The core of Stage 1 is a parameter memory $\mathcal{M}$, a lightweight key-value store whose key is a context bucket b encoding coarse characteristics of the current detection window, and whose value records historical precision and usage counts for candidate parameter pairs (E, τ) drawn from a finite set $\mathcal{C}$. At each time window, the system selects with (3) the pair that balances past reliability against under-exploration, using exploitation weight $\lambda \in [0, 1]$ and decay rate $\mu \in \mathbb{R}^+$:

$$(E^*, \tau^*) = \arg \max_{(E, \tau) \in \mathcal{C}} \left[\lambda \cdot \text{prec}_{\text{hist}}(E, \tau, b) + (1 - \lambda) \cdot \exp(-\mu \cdot n_{\text{use}}(E, \tau)) \right], \quad (3)$$

where $\text{prec}_{\text{hist}}(E, \tau, b)$ is the historical precision of parameter pair (E, τ) in context b , and $n_{\text{use}}(E, \tau) \in \mathbb{Z}_{\geq 0}$ counts prior uses of that pair, so that a pair never tried before receives the maximal exploration bonus. We set $\lambda = 0.8$ and $\mu = 0.1$, selected by grid search over $\lambda \in \{0.6, 0.7, 0.8, 0.9\}$ and $\mu \in \{0.05, 0.1, 0.2\}$ on a held-out validation set.

Each context bucket b is defined by three coarse window-level features. Activity density $\eta = |U|/\text{span}(W)$ measures active users per unit time, discretized into low, medium, and high terciles. Temporal burstiness $\mathcal{B} = \sigma_{\Delta t}/\mu_{\Delta t}$ is the coefficient of variation of inter-event times. Content homogeneity $H = 1 - \text{entropy}(\text{hashtag distribution})$ captures topic concentration. Each combination of discretized values forms a distinct bucket from which the memory retrieves relevant historical configurations.

Given the selected (E^*, τ^*) , each account u is represented by the delay embedding of (2) instantiated with those values, written $\mathbf{e}_u(t) \in \mathbb{R}^{E^*}$. It is maintained incrementally from a rolling activity buffer that appends the newest observation and discards the oldest, avoiding full reconstruction across long sequences.

For each account pair (u, v) , the shadow manifold $\mathcal{S}_v$ reconstructed from v 's embedding is indexed with a ball-tree or an approximate nearest-neighbor graph for efficient retrieval. Let t_j be the timestamp of the j -th nearest neighbor on $\mathcal{S}_v$, $d_j \in \mathbb{R}^+$ the distance to that neighbor, and $d_1 = \min_j d_j$ the smallest such distance. The trajectory of u cross-mapped through $\mathcal{S}_v$ is the distance-weighted convex combination of (4),

$$\hat{X}_{u|\mathcal{S}_v}(t) = \frac{\sum_{j=1}^{k_{\text{nn}}} w_j X_u(t_j)}{\sum_{j=1}^{k_{\text{nn}}} w_j}, \quad w_j = e^{-d_j/d_1}, \quad (4)$$

using $k_{\text{nn}} = E + 1$ neighbors, the minimum for a valid simplex in E -dimensional space (Sugihara et al., 2012). Scaling the exponent by d_1 makes the weights scale-invariant, and normalizing by their sum keeps $\hat{X}$ on the same scale as X . Predictions are computed across library lengths $L \in [10, 50]$ and cached to avoid redundant computation. The final CCM influence score, (5), takes the maximum correlation over library lengths:

$$\text{influence}(u, v) = \max_L \hat{\rho}_{\text{CCM}}(X_u, \hat{X}_{u|S_v}), \quad (5)$$

where $\hat{\rho}_{\text{CCM}}$ denotes the Pearson correlation between the observed and cross-mapped trajectories.

Evaluating all $O(N^2)$ pairs is infeasible at scale. To reduce this cost, accounts are first grouped by hierarchical clustering over temporal activity statistics (mean, variance, burstiness, and entropy). CCM is then applied within K clusters of approximate size N/K , yielding intra-cluster cost $O(N^2/K)$, supplemented by a small sample of cross-cluster pairs. Setting $K = N^{0.6}$ minimizes the combined cost and yields overall $O(N^{1.4})$ complexity, down from the naive $O(N^2)$. Clusters are evaluated independently, enabling parallel execution.

4.2 Stage 2: Active-Learning Semi-Supervised Classifier

Each account u is represented by a feature vector $\mathbf{x}_u \in \mathbb{R}^F$ with $F = 24$, capturing posting frequency, retweet behavior, hashtag usage, sentiment distribution, and temporal engagement patterns. The classifier assigns one of four behavioral labels from the troll taxonomy of Linvill and Warren (2020): fake or troll, organizational, political, and individual.

Rather than labeling all accounts, the system prioritizes annotation using uncertainty sampling (Wen et al., 2025). Let $N_{\text{tree}} = 100$ denote the number of trees in the random forest and h_i the prediction of the i -th base learner. The uncertainty of account u in (6) follows from the vote distribution over the four behavioral classes c given by (7):

$$\text{unc}(\mathbf{x}_u) = 1 - \max_c P(c | \mathbf{x}_u), \quad (6)$$

where

$$P(c | \mathbf{x}_u) = \frac{1}{N_{\text{tree}}} \sum_{i=1}^{N_{\text{tree}}} \mathbf{1}[h_i(\mathbf{x}_u) = c]. \quad (7)$$

Each active-learning round sends the $q = 64$ most uncertain accounts for human review, concentrating effort on borderline cases such as accounts that mix political messaging with individual-like engagement.

Training follows a curriculum in which difficulty, measured by uncertainty, increases progressively. The model begins on accounts with uncertainty below 0.3 and advances to higher thresholds (0.5, 0.7, 1.0) as validation accuracy exceeds 0.85 at each stage. High-confidence predictions above 0.9 are stored as pseudo-labels and incorporated into subsequent training rounds, further reducing annotation requirements while maintaining accuracy above 0.85.

The resulting account labels re-enter the pipeline as the role filter a_{role} of Eq. (1). A candidate pair from Stage 1 is forwarded to Stage 3 only if at least one of its two accounts is labeled fake or troll or organizational; pairs whose members are both labeled political or individual are discarded here. This filter removes pairs of ordinary users whose activity happens to be temporally aligned, and it is the only channel through which account-level classification affects the final coordination decision.

4.3 Stage 3: Adaptive Causal Validator

Stage 3 receives the filtered candidate set S_{12} and decides which of its pairs survive as validated coordination links. Fixed significance thresholds and universal estimator choices perform poorly when datasets differ in size, treatment prevalence, or temporal dynamics, so Stage 3 adapts both from historical detection experience.

Let $\mathcal{H}_{\mathcal{D}}$ denote the set of past detection cases retrieved for dataset $\mathcal{D}$ using coarse dataset-level features (sample size, treatment ratio, temporal coverage). Given baseline cutoff $\theta_{\text{base}} \in \mathbb{R}^+$ and sensitivity parameter $\kappa \in \mathbb{R}^+$, the dataset-specific threshold of (8) is

$$\theta_{\text{adapt}}(\mathcal{D}) = \theta_{\text{base}} \cdot \exp(\kappa \cdot \text{sr}(\mathcal{D})), \quad (8)$$

where $\kappa = 0.2$ and $\text{sr}(\mathcal{D}) = \sum_{h \in \mathcal{H}_{\mathcal{D}}} \text{prec}(h) / |\mathcal{H}_{\mathcal{D}}|$ is the mean historical precision over retrieved cases. We set $\theta_{\text{base}} = 0.40$, stricter than the Stage 1 cutoff $\rho_{\text{min}} = 0.25$, so $\theta_{\text{adapt}} \in [0.40, 0.49]$ and Stage 3 can only remove pairs that Stage 1 admitted. Contexts with reliable detection history thus receive a stricter final cutoff, while difficult

contexts keep the threshold near its baseline and send more candidates to the refutation tests.

Candidate estimators, including Generalized Synthetic Control, CausalForestDML, and GAN-ITE, are ranked by the weighted objective of (9), with $\beta_1, \beta_2, \beta_3 \in \mathbb{R}^+$:

$$\text{score}(m) = \beta_1 \frac{\text{se}_{\min}}{\text{se}(m)} + \beta_2 \text{prec}(m) + \beta_3 \text{rec}(m), \quad (9)$$

with $(\beta_1, \beta_2, \beta_3) = (0.4, 0.3, 0.3)$, where $\text{se}(m)$ is the standard error of the causal effect estimates from estimator m and $\text{se}_{\min} = \min_{m'} \text{se}(m')$. The ratio places all three terms on the same $(0, 1]$ scale. Top-scoring estimators produce conditional average treatment effects for covariate vector $\mathbf{x}_{\text{cov}}$:

$$\hat{\delta}(\mathbf{x}_{\text{cov}}) = \mathbb{E}[Y_v(1) - Y_v(0) \mid \mathbf{X} = \mathbf{x}_{\text{cov}}], \quad (10)$$

where $Y_v(1)$ and $Y_v(0)$ are the potential outcomes of v 's activity under $T_u = 1$ and $T_u = 0$, with 95% confidence intervals $\hat{\delta} \pm 1.96 \text{se}(\hat{\delta})$. Only effects significant at $p < 0.05$ whose intervals overlap across multiple high-scoring estimators are retained, and automated refutation tests (placebo treatment assignment, random subsetting, temporal holdout) flag any that prove inconsistent.

5 Experiments

We evaluate whether ACCD improves pair-level detection, reduces annotation, recovers causal structure under confounding, and remains useful as coordination tactics drift.

5.1 Experimental Settings

Twitter IRA. The release of [Linville and Warren \(2020\)](#) contains 2.9 million tweets published by accounts attributed to the Internet Research Agency. We sampled 472 of these accounts as the coordinated class. Because the release contains no authentic accounts, we assembled a control set of 2,360 accounts that posted under the same political hashtags over the same period, January 2015 to December 2017, and that appear on no published influence-operation list; this control set contributed 2,184,673 tweets. The evaluation pool thus held 2,832 accounts at a 1:5 coordinated-to-authentic ratio.

Pair construction. Following Section 3.1, we divided the observation period into non-overlapping 24-hour windows ($|W| = 24 \text{ h}$) and retained a window when at least two accounts posted $n_{\min} = 6$ or more times inside it, giving 368 usable windows.

Algorithm 1 Adaptive Coordinated Attack Detection

Require: Accounts $U = \{u_1, \dots, u_N\}$ with timestamp sequences $\mathcal{T}_u$, profiles P_u , window W
Ensure: Validated coordination pairs $\mathcal{P}$

- 1: Init long-term parameter memory $\mathcal{M}_{\text{LT}}$
- 2: **// Stage 1: Adaptive Causal Detection**
- 3: Extract context bucket: $b \leftarrow (\eta, \mathcal{B}, H)$ // density, burstiness, homogeneity
- 4: Select params: $(E^*, \tau^*) \leftarrow \arg \max_{(E, \tau)} [\lambda \text{prec}_{\text{hist}}(E, \tau, b) + (1-\lambda)e^{-\mu n_{\text{use}}}]$
- 5: Cluster accounts by temporal statistics: clusters $\leftarrow \text{Agglom}(\{(\bar{X}_u, \sigma_u^2, \mathcal{B}_u, H_u)\}, K)$
- 6: Build embeddings: $\mathbf{e}_u(t) \in \mathbb{R}^{E^*}$ via rolling buffer with (E^*, τ^*)
- 7: Compute CCM within clusters: $S_1 \leftarrow \{(u, v) : \text{influence}(u, v) > \rho_{\min}\}$ // $\rho_{\min} = 0.25$
- 8: **// Stage 2: Semi-Supervised Classification**
- 9: Extract features: $\mathbf{x}_u \in \mathbb{R}^F$ // frequency, retweets, hashtags, sentiment, timing
- 10: Uncertainty: $\text{unc}(\mathbf{x}_u) = 1 - \max_c P(c \mid \mathbf{x}_u)$
- 11: Query labels: cand $\leftarrow \text{TopK}(\text{unc}(\mathbf{x}_u), q)$
- 12: Train: $S_2 \leftarrow \{(u, r(u))\} \leftarrow \text{Classify}(\mathbf{x}_u, \text{labels}, \text{curriculum})$
- 13: Update pseudo-labels for high-confidence predictions
- 14: Role filter: $S_{12} \leftarrow \{(u, v) \in S_1 : r(u) \in \mathcal{R} \vee r(v) \in \mathcal{R}\}$ // $r(\cdot)$ from S_2
- 15: **// Stage 3: Adaptive Causal Validation**
- 16: Threshold: $\theta_{\text{adapt}} \leftarrow \theta_{\text{base}} \cdot \exp(\kappa \text{sr}(\mathcal{D}))$
- 17: Score: $\text{score}(m) \leftarrow \beta_1 \text{se}_{\min} / \text{se}(m) + \beta_2 \text{prec}(m) + \beta_3 \text{rec}(m)$
- 18: **for** each $(u, v) \in S_{12}$ **do**
- 19: Estimate: $\hat{\delta}(\mathbf{x}_{\text{cov}}) \leftarrow \mathbb{E}[Y_v(1) - Y_v(0) \mid \mathbf{X} = \mathbf{x}_{\text{cov}}]$, $\text{CI} \leftarrow \hat{\delta} \pm z_{0.025} \text{se}$
- 20: Refutation: placebo, subsetting, temporal holdout
- 21: **end for**
- 22: $\mathcal{P} \leftarrow \{(u, v, \hat{\delta}, \text{CI}) \in S_{12} : \text{influence}(u, v) > \theta_{\text{adapt}} \wedge \text{valid}(u, v)\}$
- 23: **return** $\mathcal{P}$

Within each retained window, candidate pairs were drawn from the hierarchical clusters of Section 4.1 plus a small random sample of cross-cluster pairs. This produced 220,800 candidate pairs, of which 9,600 were positive, meaning both members were IRA-attributed and co-active in the same window, and 211,200 were negative, meaning at least one member was authentic. The resulting positive-to-negative ratio is 1:22. Because campaign attribution is available only at the account level, positive pairs represent joint campaign membership rather than verified direct links; we return to this proxy in the Limitations section.

Reddit. Pushshift ([Baumgartner et al., 2020](#)) is a raw archive rather than a labeled coordination benchmark, so we constructed the labels ourselves. We selected 12 subreddits that were targets of publicly documented brigading episodes between June 2016 and June 2017 and collected 1,176,392 comments from 1,648 accounts active in them. An

account was labeled coordinated when it took part in a documented episode and that participation survived manual verification against the episode record, which gave 284 coordinated accounts and 1,364 authentic ones. Windowing, clustering, and pair construction followed the Twitter procedure without modification.

Implementation. We implemented ACCD in PyTorch 2.0.0 and scikit-learn 1.3.0, trained on NVIDIA A100 GPUs for 100 epochs with batch size 64, learning rate 0.001, and CosineAnnealing-WarmRestarts scheduling. In Stage 3, the treatment threshold is $\alpha = 2.0$ times an account’s own baseline posting rate, the outcome lag is $\Delta = 3$ h, and $\mathbf{x}_{\text{cov}}$ has nine dimensions. Evaluation used stratified 5-fold cross-validation; folds are split by window so that no window contributes to both training and test, which prevents temporal leakage between pairs drawn from the same burst of activity.

Baselines. We compared against methods from three categories, all evaluated on the identical candidate pair set and the same folds. For causal detection, we used standard CCM (Sugihara et al., 2012) with fixed parameters ($E = 3$, $\tau = 1$ h, $\rho_{\min} = 0.3$, the recommended configuration from the original paper) and a Pearson correlation baseline to assess the value of causal over correlational detection. For network-based detection, we used LCN+HCC (Pacheco et al., 2021), which builds coordination graphs from co-retweet, co-tweet, and co-mention patterns. For temporal modeling, we included AMDN-HAGE (Sharma et al., 2021), which jointly models individual activities and latent group behaviors using temporal point processes. Methods that natively output account-level scores are converted to pair-level predictions by scoring a pair positive when both members are flagged.

Metrics. Precision, recall, and F1 are computed over candidate pairs with the coordinated class as positive, on held-out test folds. Stage 2 performance is reported separately as four-class account-level accuracy. Labeling reduction measures the percentage decrease in manually annotated accounts relative to a fully supervised baseline: $(1 - n_{\text{labeled}}/N_{\text{train}}) \times 100\%$.

5.2 Main Results

Detection performance. Table 1 shows that ACCD reaches 87.3% F1 on Twitter IRA, 11.5 points above CCM (75.8%), with 85.6% precision and 89.2% recall. On Reddit it reaches 84.7% F1 versus 71.2% for correlation.

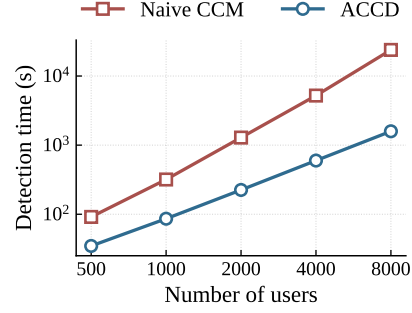


Figure 3: Detection time versus number of accounts. The log-log slope is 1.41 ($R^2 = 0.988$), consistent with $O(N^{1.4})$.

Table 1: Pair-level detection performance on held-out folds. Subscripts are standard deviations over 5 folds; bold marks the best result. IRA = Twitter IRA; Reddit = Pushshift traces.

Method	Prec.	Rec.	F1	Time
<i>Twitter IRA dataset</i>				
CCM (Sugihara et al., 2012)	72.1 $\pm$ 1.4	79.8 $\pm$ 1.6	75.8 $\pm$ 1.5	181.3m
Correlation	69.3 $\pm$ 1.7	78.1 $\pm$ 1.8	73.5 $\pm$ 1.6	122.5m
LCN+HCC (Pacheco et al., 2021)	74.5 $\pm$ 1.2	76.2 $\pm$ 1.4	75.3 $\pm$ 1.3	96.7m
AMDN-HAGE (Sharma et al., 2021)	68.9 $\pm$ 1.0	82.4 $\pm$ 1.2	75.1 $\pm$ 1.1	211.4m
ACCD (Ours)	85.6$\pm$0.6	89.2$\pm$0.8	87.3$\pm$0.7	72m
<i>Reddit coordination traces</i>				
Correlation	68.4 $\pm$ 1.9	74.3 $\pm$ 2.1	71.2 $\pm$ 2.0	98.4m
ACCD (Ours)	82.1$\pm$0.9	87.6$\pm$1.1	84.7$\pm$1.0	61m

Table 2: Computational efficiency. Speed is relative to exhaustive CCM; Pair Ret. is the share of CCM-detected coordinated pairs retained after clustering (coverage, not classification accuracy).

Method	Iter.	Mem.	Speed	Pair Ret.
CCM (Sugihara et al., 2012)	65 ep	8.2G	1.0 $\times$	100%
Fixed Param.	58 ep	7.9G	1.1 $\times$	98.3%
ACCD	40 ep	4.5G	2.8$\times$	96.7%

Labeling efficiency. Annotation is requested at the account level, since Stage 2 labels accounts rather than pairs. With a 70/30 split of the 2,832-account pool, a fully supervised approach would label all 1,982 training accounts; ACCD’s active-learning module reached equivalent accuracy using only 628 labeled accounts, a reduction of 68.3%.

Computational efficiency. Table 2 and Figure 3 show a 2.8 $\times$ average speedup over exhaustive CCM while retaining 96.7% of its detected coordinated pairs. At $N = 2,832$, the end-to-end gain is 2.5 $\times$. ACCD converged in 40 versus 65 iterations and used 4.5 versus 8.2 GB. Runtime over $N \in [500, 8000]$ has slope 1.41 ($R^2 = 0.988$), consistent with $O(N^{1.4})$.

Table 3: Causal-edge recovery AUC on synthetic data; bold marks the best result.

Method	Low noise	High noise + conf.
Pearson Corr.	0.708	0.586
Fixed CCM	0.842	0.732
ACCD (Ours)	0.934	0.861

5.3 Synthetic Causal Recovery

Synthetic series use known directed links, a latent leader for coordinated accounts, and a shared trending confounder. Table 3 shows ACCD at 0.934 AUC under low noise and 0.861 under strong confounding, while Pearson correlation falls from 0.708 to 0.586.

5.4 Ablation Study

The full model achieved 87.3% pair-level F1 and 88.9% Stage-2 account accuracy. Removing adaptive parameters reduced pair-level F1 by 8.2 points; removing active learning reduced Stage-2 accuracy by 5.4 points; removing adaptive validation reduced out-of-distribution generalization by 3.7 points. These quantities are not additive.

Adaptive parameters matter because bursty campaigns favor short embeddings ($E = 2$), whereas slower operations favor longer ones ($E = 5$). Under the IRA pool’s 5:1 account imbalance, uncertainty sampling focuses labels on minority coordinated accounts. Adaptive validation mainly improves transfer beyond the training distribution.

5.5 Concept Drift and Retraining Cadence

With memory and Stage 2 frozen after training on the first six IRA months, temporal extrapolation reduced pair-level F1 by 1.6, 5.1, and 11.2 points after one, three, and six months. Updating the memory and classifier on newly labeled uncertain accounts took 9.4 minutes versus a 72-minute detection run. A six-week cadence therefore keeps degradation near two F1 points; faster-changing deployments can update more often.

6 Conclusion

ACCD combines context-adaptive CCM, active-learning role filtering, and experience-driven causal validation in a sequential pipeline. It reaches 87.3% pair-level F1 on Twitter IRA and 84.7% on Reddit, reduces labeling by 68.3%, and gives a $2.5\times$ end-to-end speedup with empirical $O(N^{1.4})$ scaling. Synthetic recovery reaches 0.934 AUC and

remains 0.861 under strong confounding, supporting context-adaptive causal inference for practical coordination screening.

7 Limitations

Computational cost. Despite the $O(N^{1.4})$ reduction achieved through hierarchical clustering, CCM-based causal analysis remains more expensive than simple correlation heuristics. Deployments at the scale of millions of accounts may require additional approximations to maintain practical response times.

Cold start. The memory-guided mechanism requires historical detection experience to function well. For entirely novel coordination patterns without analogous precedents, the system falls back to default parameters and may underperform until sufficient experience accumulates. The exploration term in the parameter selection objective partially mitigates this, but initial detection of unprecedented attack patterns remains difficult.

Pseudo-label bias. High-confidence predictions above 0.9 are incorporated as pseudo-labels in subsequent training rounds. If the model initially misclassifies a coordinated account with high confidence, this error can propagate through self-training. Periodic re-evaluation of pseudo-labels and confidence calibration reduce but do not eliminate this risk.

Pair-level label proxy. Positive pairs are defined as two accounts attributed to the same confirmed campaign and co-active in the same window. Campaign membership does not establish that two specific accounts influenced each other directly, so some positive pairs may reflect parallel rather than mutual coordination. This proxy is the strongest label available for public influence-operation releases, but it means reported precision is measured against campaign attribution rather than against verified pairwise links. The Reddit labels, built from documented brigading episodes with manual verification, carry the same caveat at a smaller scale.

Role filter dependence. The final decision requires at least one member of a pair to receive a campaign-plausible Stage 2 label. Coordination conducted entirely through accounts that appear political or individual would be filtered out before causal validation, trading recall for fewer false positives among ordinary users.

Adversarial adaptation. Sophisticated actors who become aware of the detection mechanism

may adapt their coordination patterns to evade it. While the memory-guided system can update from new patterns over time, deliberate evasion targeting known detection signatures remains an open research problem.

8 Ethical and societal implications

Positive impact. By reducing manual annotation by 68.3% and cutting end-to-end detection time by 2.5 $\times$, ACCD lowers the barrier for civil-society organizations, academic researchers, and platform trust-and-safety teams to identify influence operations at scale, directly supporting the protection of democratic discourse and election integrity. More broadly, ACCD's outputs, which include flagged coordination pairs with causal effect estimates and confidence intervals, can serve as quantitative instruments for measuring the scale and intensity of coordinated inauthentic behavior, contributing to the broader goal of evaluating AI's social impact.

False positive risks. Automated coordination detection carries inherent risks of falsely flagging legitimate collective action, including grassroots movements, fan communities, or activist networks that exhibit coordination-like patterns for authentic purposes. We recommend that ACCD be used as a screening tool, with human review of flagged accounts before any enforcement action, particularly in politically sensitive contexts.

Dual-use concerns. Coordination detection technology could be misused to suppress legitimate dissent or target specific communities. We advocate for transparency in deployment, clear appeal mechanisms for affected users, and independent auditing of detection systems to prevent misuse.

References

- J Abécassis, É Dumas, J Alberge, and G Varoquaux. 2025. [From prediction to prescription: Machine learning and causal inference for the heterogeneous treatment effect](#). *Annual Review of Biomedical Data Science*, 8:381–404.
- Meysam Alizadeh, Jacob N. Shapiro, Cody Buntain, and Joshua A. Tucker. 2020. [Content-based features predict social media influence operations](#). *Science Advances*, 6(30):eabb5824.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.
- Leo Breiman. 2001. [Random forests](#). *Machine Learning*, 45(1):5–32.
- JE Chen and A Jing. 2025. [Recent advances in causal machine learning and dynamic policy learning](#). *Wiley Interdisciplinary Reviews: Computational Statistics*, 17(4).
- QY Chen. 2026. [Networked manipulation: multi-actor coordinated communities and formation mechanisms in disinformation campaigns](#). *Online Information Review*, 50(3):485–504.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. [Double/debiased machine learning for treatment and structural parameters](#). *The Econometrics Journal*, 21(1):C1–C68.
- K Colangelo and YY Lee. 2026. [Double debiased machine learning nonparametric inference with continuous treatments](#). *Journal of Business & Economic Statistics*, 44(1):67–79.
- Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. [The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race](#). In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 963–972.
- Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. [Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling](#). *IEEE Transactions on Dependable and Secure Computing*, 15(4):561–576.
- William D. Lewis, Haotian Zhu, Keaton Strawn, and Fei Xia. 2025. [Tapping into social media in crisis: A survey](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 306–331, Vienna, Austria. Association for Computational Linguistics.
- N George, A Sham, T Ajith, and M Bastos. 2025. [Forty thousand fake twitter profiles: A computational framework for the visual analysis of social media propaganda](#). *Social Science Computer Review*, 43(3):451–474.
- C. W. J. Granger. 1969. [Investigating causal relations by econometric models and cross-spectral methods](#). *Econometrica*, 37(3):424–438.
- Q Guyan, YW Liu, J Liu, and P Zhang. 2025. [PEGNN: Peripheral-enhanced graph neural network for social bot detection](#). *Expert Systems with Applications*, 278.
- K Imai and ML Li. 2025. [Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments](#). *Journal of Business & Economic Statistics*, 43(1):256–268.

- A Isufaj, CD Martins, M Cavazza, and H Prendinger. 2025. [Applying time delay convergent cross mapping to bitcoin time series](#). *Expert Systems with Applications*, 277.
- David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. [The science of fake news](#). *Science*, 359(6380):1094–1096.
- Darren L. Linvill and Patrick L. Warren. 2020. [Troll factories: Manufacturing specialized disinformation on twitter](#). *Political Communication*, 37(4):447–467.
- 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*.
- S Nizam, A Codi, ER McQuade, and D Benkeser. 2025. [Highly adaptive lasso for estimation of heterogeneous treatment effects and treatment recommendation](#). *Journal of Causal Inference*, 13(1).
- Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. [Uncovering coordinated networks on social media: Methods and case studies](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 455–466.
- 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.
- 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 David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, Berlin, Heidelberg.
- D Thiele, M Milzner, A Heft, BN Gong, and B Pfetsch. 2025. [Attributing coordinated social media manipulation: A theoretical model and typology](#). *New Media & Society*.
- Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. [Online human-bot interactions: Detection, estimation, and characterization](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 280–289.
- KK Wang, XJ Wang, KM Chen, ZX Wang, and KF Zheng. 2025. [Multi-stage self-training social bot detection based on graph neural network](#). *Engineering Applications of Artificial Intelligence*, 152.
- Derek Weber and Frank Neumann. 2020. [Who’s in the gang? revealing coordinating communities in social media](#). In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 89–93.
- ZT Wen, O Pizarro, and S Williams. 2025. [Active self-semi-supervised learning for few labeled samples](#). *Neurocomputing*, 614.
- AP Wu, K Kuang, RX Xiong, and F Wu. 2025. [Instrumental variables in causal inference and machine learning: A survey](#). *ACM Computing Surveys*, 57(11).
- K Wu, YG Lin, XM Wang, and L Wang. 2026. [Multimodal learning-based relational graph neural networks for social bot detection](#). *IEEE Transactions on Network and Service Management*, 23:2694–2711.
- M Zhou, D Zhang, YD Wang, YA Geng, YX Dong, and J Tang. 2025. [LGB: Language model and graph neural network-driven social bot detection](#). *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4728–4742.