

A Multi-Agent Ai Framework for Detecting Coordinated Fraud Campaigns In Digital Media

Dr Priyanka Verma ^{1*} Dr Jitendra Sharma ² Mohammed Firdos Alam Sheikh ³ Dr Aastha Saxena ⁴ Ranu Raj ⁵ Pankaj Pareek ⁶

¹ Professor, Department of Computer Science and Engineering Poornima University Jaipur
priyankag.gianchandani@gmail.com

² Professor, Faculty of Design and Arts Poornima University Jaipur
jitendra.sharma@poornima.edu.in

³ Assistant Professor, Department of Computer Science and Engineering Poornima University Jaipur
firdos.sheikh@gmail.com

⁴ Head, Department of Journalism and Mass Communication Poornima University Jaipur
aastha.saxena@poornima.edu.in

⁵ Assistant Professor, Faculty of Design and Arts, Poornima University Jaipur
ranu.raj@poornima.edu.in

⁶ Assistant Professor, Faculty of Design and Arts, Poornima University Jaipur
pankaj.pareek@poornima.edu.in

Abstract

Coordinated fraud campaigns in digital media spanning ad-fraud (clicks, impressions, installs), coordinated inauthentic behavior (CIB), astroturfing, and manipulation networks—have evolved into adaptive, multi-stage operations that exploit platform affordances, automation, and cross-channel amplification. Single-model detectors (e.g., isolated bot classifiers or per-event anomaly scoring) are often brittle against adversarial drift, data sparsity, and the inherently relational nature of coordination. This review synthesizes the state of the art and proposes a multi-agent AI framework that operationalizes coordination detection as a pipeline of collaborating agents: (i) ingestion and provenance agents, (ii) feature and representation agents, (iii) coordination-graph construction agents, (iv) anomaly and campaign discovery agents, (v) evidence fusion and adjudication agents, and (vi) monitoring and governance agents. We ground the design in established findings on coordinated information operations, coordination archetypes and temporal stability, and practical coordination toolkits, while incorporating advances in dynamic-graph anomaly detection, graph neural approaches to fraud, and evaluation pitfalls in bot detection. We provide a “camera-ready” research agenda covering robust temporal/multimodal fusion, agent-to-agent calibration, explainable coordination evidence, and responsible deployment under uncertainty.

Keywords: coordinated fraud, digital media, multi-agent systems, coordination graphs, dynamic graph anomaly detection, ad fraud, CIB, bot detection, evidence fusion, explainable AI

1. Introduction

Digital media fraud has diversified beyond isolated spam accounts and one-off click bots into coordinated campaigns that distribute tasks across accounts, devices, IPs, pages, apps, and content modalities. Adversaries increasingly optimize *group-level outcomes* (reach, ranking manipulation, install attribution, narrative dominance) rather than individual account stealth. The result is an arms race where detectors must identify *collective signatures*: synchronized actions, shared resources, and structurally improbable patterns across time.

Empirically, coordination manifests as suspicious similarity in timing, content, URLs, hashtags, audience targeting, or referral paths. In social platforms, coordinated inauthentic behavior can exploit feed algorithms by orchestrating engagement bursts and “signal boosting” [1]. In political and issue campaigns, coordination often takes the form of coordinated link sharing, where multiple accounts rapidly distribute the same URLs to shape media narratives [7]. Recent work shows that coordination has distinct temporal archetypes and that campaign influence relates not only to volume but to the stability and timing of coordinated clusters [28]. Practical frameworks have emerged to build coordination networks and detect such clusters at scale [29].

Meanwhile, advertising ecosystems face well-documented threats including click fraud, impression fraud, install fraud, domain spoofing, and invalid traffic routed through botnets or incentivized human farms. Surveys highlight the breadth of ad-fraud mechanisms and defenses [3,16], and systematic reviews have cataloged AI-driven click-fraud approaches and gaps in evaluation realism [15]. Dataset skew, attacker adaptation, and multi-entity collusion all complicate detection.

Across these domains, coordination is relational. That fact has accelerated the adoption of graph-based anomaly detection methods. Systematic reviews show that graph-based anomaly detection (GBAD) is central to modern fraud detection because it captures interactions and shared entities (e.g., devices, IPs,

publishers) [18]. Newer surveys cover anomaly detection in dynamic graphs [10] and GNN-based graph anomaly detection [23]. At the same time, applied GNN work and reviews argue that graph neural models are especially suitable for fraud because they capture higher-order relational patterns [27]. Yet purely “end-to-end” GNN systems can be opaque and fragile, motivating attention to explainability for graph models [26] and to evaluation pitfalls (e.g., false positives and language sensitivity in bot scoring tools) [35].

Why multi-agent AI?

Coordinated fraud detection is not a single classification task—it is a systems problem: ingest heterogeneous signals, align identities across spaces, build evolving graphs, detect group-level anomalies, explain and triage results, and monitor drift. Surveys on anomaly detection emphasize that robust systems require careful modeling assumptions, evaluation, and thresholding under uncertainty [33]. Multi-agent systems (MAS) offer an architectural pattern where specialized agents collaborate, negotiate uncertainty, and provide modularity for evolving threats [32]. With the rise of LLM-based and tool-using agents, MAS is increasingly used to orchestrate complex workflows; a structured survey notes common components and workflows for LLM-based MAS [31]. This review therefore frames coordinated fraud detection as a multi-agent pipeline: not merely multiple models, but multiple *roles* with explicit interfaces, accountability, and governance.

2. Threat landscape: what “coordinated fraud campaigns” look like

2.1 Coordination in social and information ecosystems

Coordination may involve synchronized posting, shared URL bursts, or distributed amplification across accounts and communities. Coordinated inauthentic behavior can manipulate algorithmic ranking by producing an engineered pattern of engagements that appear organic at the individual level but are statistically implausible at group level [1]. Coordinated link sharing has been documented during elections, where clusters of accounts push the same URLs in tight time windows [7]. Review work on coordinated social media manipulation (CSMM) highlights how campaigns differ in actor types, platforms, content forms, and attribution challenges [30]. Specific studies also document novel manipulation tactics during COVID-19, emphasizing cross-platform and narrative amplification patterns [9]. Coordination can also target follower/following structures and engagement metrics, motivating unsupervised approaches to detect coordinated fake-follower behavior [38].

Automation further complicates detection. Foundational work characterizes the role and evolution of social bots [4], while later overviews describe a decade of bot detection methods and how bots adapt [5]. Research on “evolved” spambots shows that adversaries can optimize bot behavior to evade detectors through algorithmic evolution [13], and group-behavior modeling can help detect spambot clusters by comparing behavioral “fingerprints” rather than profiles alone [36].

2.2 Ad-fraud and digital advertising campaigns

Advertising fraud is economically motivated and operationally sophisticated. Surveys of online advertising fraud catalog attacks across the ad supply chain, including invalid traffic, click farms, impression inflation, and attribution manipulation [16]. Click fraud surveys describe both attacker techniques and defenses, emphasizing the difficulty of distinguishing legitimate from fraudulent engagement in noisy environments [3]. Dedicated reviews of AI-based click-fraud detection outline features (timing, navigation, dwell, conversion) and stress evaluation gaps: non-representative datasets, missing ground truth, and limited adversarial testing [15].

Empirical studies also target specific ad-fraud modalities. For example, impression fraud in mobile advertising has been addressed via ensemble learning using behavioral features [17]. Click-fraud detection in online advertising has been studied using machine learning on real browsing behavior logs [11]. Other work uses graph models to detect fraudulent traffic patterns in online advertising and demonstrates improvements when relational structures are modeled explicitly [22]. These studies collectively motivate coordination-aware detection rather than per-event scoring.

2.3 Fraud as a graph anomaly problem

Fraud campaigns are often multi-entity phenomena: accounts share devices, IPs, payment instruments, publishers, and content templates. A systematic literature review of graph-based anomaly detection in fraud detection shows how GBAD methods model these relationships to identify suspicious subgraphs and communities [18]. Dynamic graphs matter because campaigns evolve: adversaries rotate resources and adapt tactics. Surveys of anomaly detection in dynamic graphs summarize temporal graph representations and detection taxonomies [10]. Surveys of GNN-based graph anomaly detection detail model families (autoencoders, contrastive approaches, supervised/semi-supervised) and open challenges (dynamic graphs, explainability, adversarial robustness) [23]. Recent surveys also broaden anomaly detection to distributed systems, reinforcing that graph-based approaches must handle heterogeneity and evolving structure [24].

3. Design principles for a multi-agent coordination-fraud framework

A multi-agent architecture should follow five principles:

1. **Relational first:** treat coordination as a property of *sets of entities* and their interactions, aligning with findings on coordinated user similarity [1] and coordination network methodologies [29].
2. **Temporal native:** preserve time explicitly, since coordination archetypes and influence depend on temporal dynamics [28].
3. **Evidence separation and fusion:** maintain interpretable evidence objects (e.g., “shared URL burst,” “synchronized action,” “resource reuse”) and fuse them later, enabling explainability [26].
4. **Continuous evaluation and drift monitoring:** integrate monitoring agents to detect shifting baselines and false positive drift—especially important given known bot detection pitfalls [35].
5. **Governance and responsible use:** embed privacy, bias, and due-process constraints at the architectural level, reflecting the risks of mislabeling and misuse [35].

4. Proposed Multi-Agent AI Framework

We propose a modular set of cooperating agents. The goal is to transform heterogeneous digital traces into *campaign hypotheses* supported by a structured evidence graph and an audit trail.

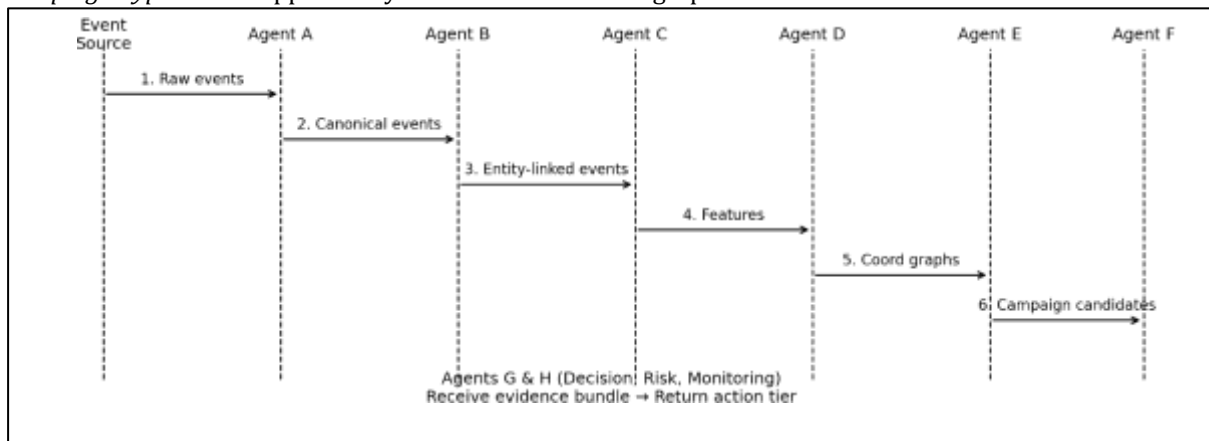


Figure 1: UML Sequence Diagram

4.1 Agent roles and responsibilities

Agent A: Ingestion & Provenance Agent

Collects logs (events, clicks, impressions, posts), resolves schemas, and attaches provenance metadata. In ad ecosystems, this agent enforces source integrity (server logs vs. client beacons) and tracks attribution paths. In social media, it handles API sampling constraints and data quality issues, a known factor in bot/coordination measurement error [35].

Agent B: Identity Resolution & Entity Linking Agent

Maps raw identifiers into a multi-entity model: user/account, device, IP/ASN, cookie, payment token, publisher/app/site, content/URL template, referral path. This agent is crucial because coordinated fraud frequently uses partial reuse (e.g., rotating accounts but reusing devices).

Agent C: Feature & Representation Agent

Produces multi-modal features:

- **Behavioral:** inter-event times, burstiness, session depth (often used in click-fraud detection [11,15]).
- **Content/URL:** URL canonicalization, near-duplicate text, template similarity (used in coordinated link sharing studies [7]).
- **Graph features:** node centrality, motif counts, co-occurrence measures (central in GBAD fraud detection [18]).
- **Temporal embeddings:** time-bucketed interaction vectors.

Agent D: Coordination Graph Construction Agent

Builds one or more coordination graphs, such as:

- **Shared-artifact graphs:** accounts ↔ URLs, devices ↔ publishers, IPs ↔ apps.
- **Synchronized action graphs:** edges weighted by temporal co-activity, aligning with synchronized action approaches to coordination detection [39].
- **Multiplex temporal graphs:** separate layers for different coordination channels, consistent with coordination archetype analysis [28].
- Tooling can reuse coordination network frameworks described in practice-oriented toolkits [29].

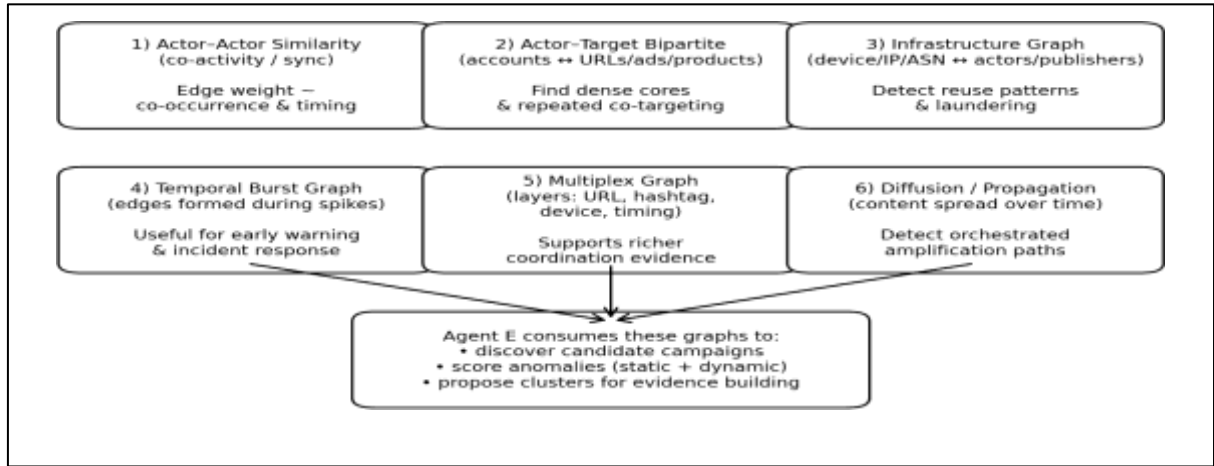


Figure 2: Coordination Graph Types

Agent E: Campaign Discovery Agent (Dynamic Graph Anomaly + Community)
Runs detection on evolving graphs using a portfolio:

- GBAD baselines (odd subgraphs, dense blocks, anomalous bipartite cores) grounded in fraud-graph reviews [18].
 - Dynamic graph anomaly methods informed by surveys of temporal graph AD [10].
 - GNN-based anomaly scoring for complex relational signals [23], guided by GNN-fraud review insights [27].
- Outputs are *candidate campaigns* = sets of entities + time intervals + preliminary anomaly rationale.

Agent F: Evidence Builder & Explainer Agent

Constructs a human-readable evidence bundle per candidate:

- Top shared artifacts (URLs, devices, IP blocks), burst windows, synchronized action summaries [39].
- Counterfactual checks and explanation artifacts for graph models where possible, leveraging taxonomies in GNN explainability research [26].
- This agent focuses on “why this looks coordinated,” not just “score > threshold.”

Agent G: Adjudication & Risk Agent

Combines evidence using calibrated fusion: weighted rules + probabilistic models + uncertainty bounds. It explicitly tracks *confidence* and *expected false positive risk*, a best practice given measurement pitfalls and domain shifts [35]. It can incorporate ensemble strategies common in ad-fraud contexts [17] and reconcile conflicting signals.

Agent H: Monitoring, Drift & Governance Agent

Monitors feature drift, alert rates, and post-hoc validation outcomes. It runs continuous checks for spurious correlations and language/region biases (relevant to bot-detection validity) [35]. It also enforces governance: data minimization, retention, and audit logs.

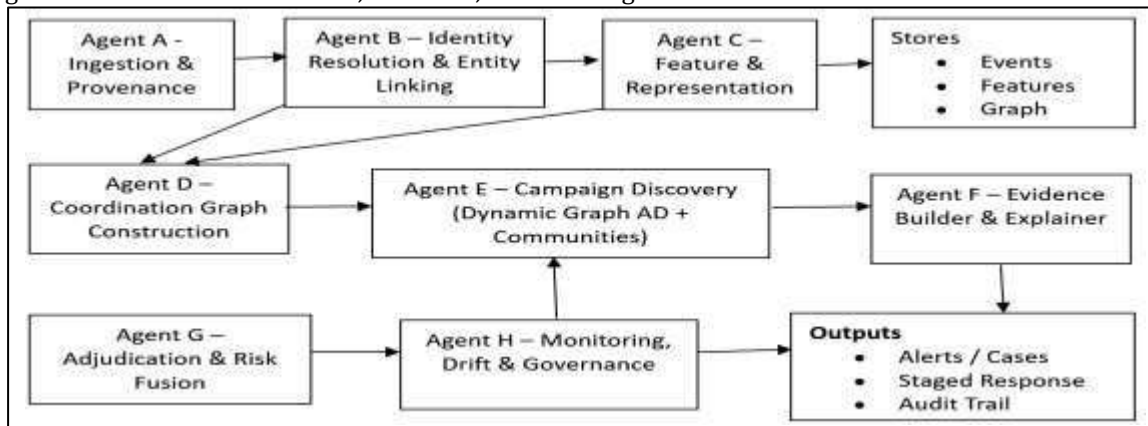


Figure 3: Multi-Agent Framework

4.2 Information flow: “evidence objects” instead of raw scores

A key architectural choice is to pass **evidence objects** between agents: structured claims like

- “Accounts {u1...uK} posted URL X within Δt windows repeatedly” (link-sharing coordination) [7]
- “Entities show multiplex synchronized actions across layers” [28,39]

- “Bipartite core (devices $\leftrightarrow$ publishers) significantly denser than baseline” (GBAD) [18]
This improves interpretability and enables later fusion to be auditable.

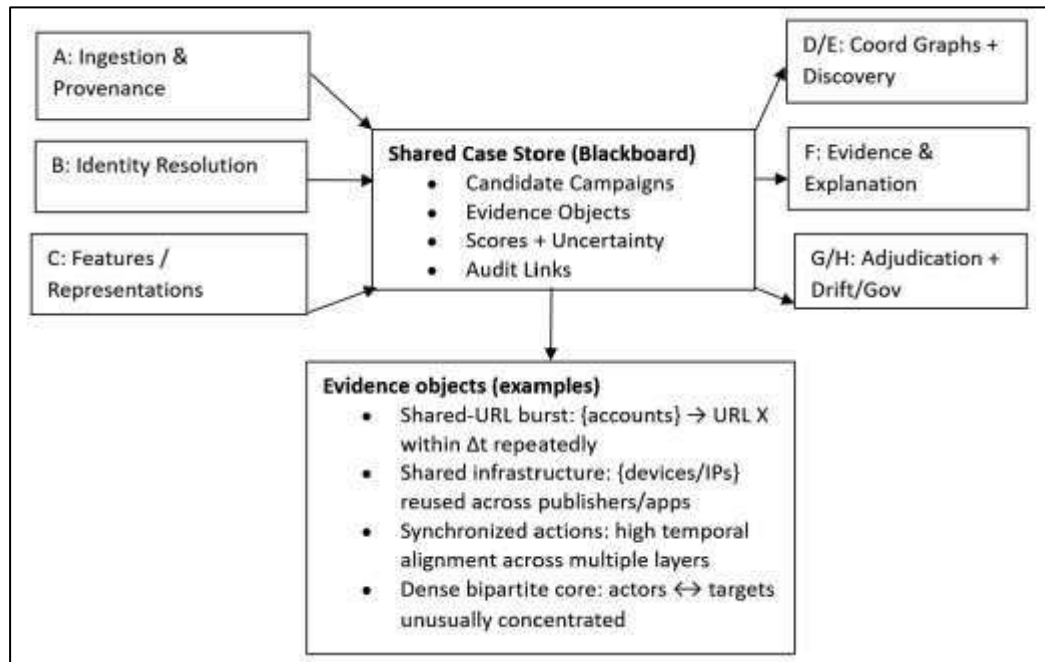


Figure 4: Evidence-Based Flow(Blackboard/Case store)

5. Algorithmic Toolkit Mapped to Agent Roles

The CAMPAIGN-MA framework assigns each major analytical task to a specialized agent, but the effectiveness of the architecture ultimately depends on the underlying algorithmic toolkit. Coordinated fraud requires algorithms capable of capturing relational structures, temporal evolution, and multi-modal signals. This section expands on the three core algorithmic layers: coordination-graph formation, dynamic graph anomaly detection, and GNN-based relational modelling [40].

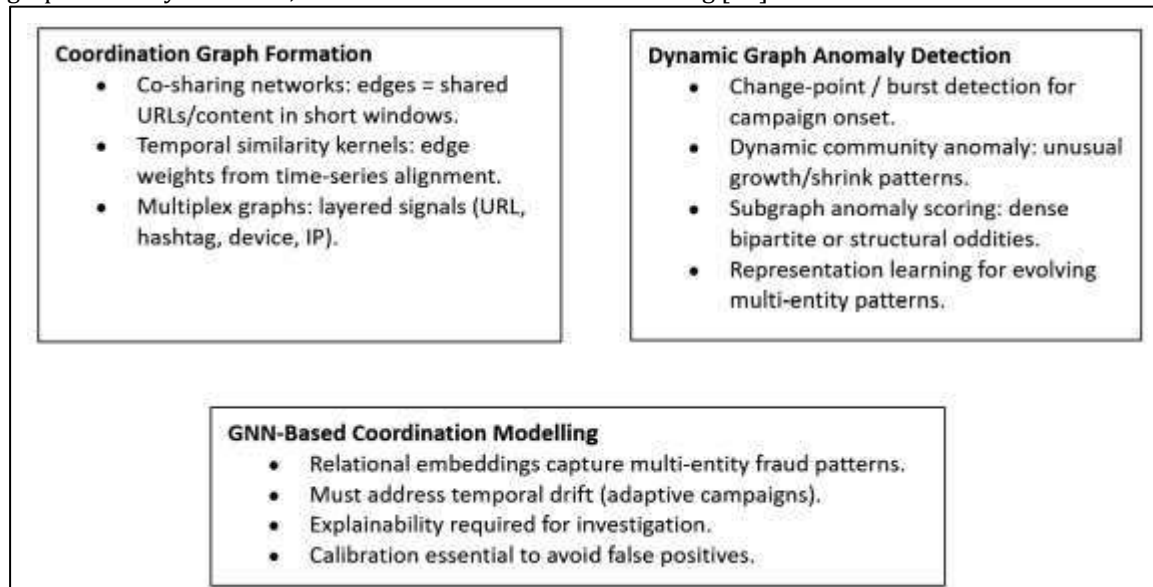


Figure 5: Coordination Analysis Toolkit

5.1 Coordination Graph Formation Strategies

A coordination graph is a structured representation that captures how actors, artifacts, and behaviors relate to one another over time. The construction of these graphs is central to detecting coordinated campaigns because coordination is fundamentally a relational property—one that emerges not from isolated actions but from patterns of similarity, repetition, and synchrony across multiple actors.

Co-sharing networks.

One of the most intuitive and empirically validated ways to represent coordination is through co-sharing networks. These graphs create edges between accounts when they share URLs, hashtags, content templates,

or targets within a short defined time window. Research on coordinated link sharing behavior (e.g., Italian election studies [7]) shows that malicious campaigns often rely on repeated bursts of identical or near-identical URLs distributed across multiple accounts in a way that is statistically implausible for organic behavior. These co-sharing edges allow downstream detection agents to surface clusters of actors who repeatedly engage with the same content in highly compressed temporal intervals.

Temporal similarity kernels.

While simple co-sharing edges capture simultaneous activity, many campaigns operate with subtle timing patterns rather than perfect synchrony. Temporal similarity kernels address this by computing edge weights based on time-series similarity—including dynamic time warping, lag alignment, or cross-correlation measures. The result is a weighted coordination graph in which two actors may not act at the exact same moment but still exhibit highly aligned temporal rhythms, indicative of centrally orchestrated behavior. These kernels support the identification of coordination archetypes, which recent work (e.g., [28]) highlights as essential for distinguishing short, intense bursts from long-running, slow-burn campaigns [41].

Multiplex graphs.

Coordinated fraud does not rely on a single communication channel or a single artifact type. Actors may coordinate via shared URLs, but also via shared device infrastructure, repeated IP block reuse, similar text templates, synchronized posting, or common audiences. A multiplex graph explicitly separates these heterogeneous signals into layers—for example:

- URL-sharing layer
- hashtag/content layer
- device/IP infrastructure layer
- timing layer
- diffusion layer

By modeling each layer independently, the system preserves the integrity of each signal while enabling cross-layer reasoning. Studies on temporal multiplex modeling demonstrate that coordination often becomes visible only when signals are combined, as some campaigns purposely mask individual-layer signals while unintentionally revealing cross-layer consistency [28].

Together, these three graph formation strategies enable Agent D (Coordination Graph Construction) to create a rich, structured, and interpretable representation of potential campaign behavior, setting the stage for anomaly detection and deeper analysis.

5.2 Dynamic Anomaly Detection Families

Once coordination graphs are built, the next task is to detect anomalous patterns that indicate the presence of a coordinated fraud campaign. Dynamic graph anomaly detection (DGAD) provides the technical foundation for this step. Recent surveys (e.g., [10]) categorize dynamic graph AD into several methodological families, each targeting a different kind of underlying irregularity.

1. Change-point and burst detection: A coordinated campaign often begins with a measurable temporal disturbance—an abrupt increase in traffic, synchronized engagement, or a sudden emergence of shared artifacts. Change-point detection algorithms identify points in time where statistical properties shift, while burst detection highlights rapid, concentrated activity increases. These tools provide the earliest warning signals of campaign onset and can trigger immediate construction of short-window coordination graphs for rapid investigation.

2. Dynamic community anomaly detection: Communities are natural structures in social and fraud graphs. Legitimate communities tend to evolve gradually and exhibit stable patterns of interaction. Coordinated campaigns, however, often produce communities that form too quickly, persist too rigidly, or dissolve too abruptly. Dynamic community anomaly detection methods track how clusters evolve across time slices, flagging those with unusual formation trajectories, dense intra-group activity, or unexpectedly consistent behavior across windows.

3. Subgraph anomaly scoring: Many fraud patterns manifest as unusual subgraph structures rather than whole anomalous communities. For instance, in ad fraud, a small number of devices connecting to a disproportionately large number of publishers, or vice versa, forms a suspicious bipartite core. Subgraph anomaly detection methods evaluate the local structural irregularity of such cores, assessing whether their density, degree patterns, or motif frequencies significantly deviate from baseline expectations. These methods are particularly useful when campaigns rely on shared infrastructure rather than explicit content coordination [42].

4. Representation learning for multi-entity patterns: Dynamic graph embeddings and graph neural approaches (surveyed in [23] and reinforced by fraud-specific GNN reviews [27]) provide a more flexible way to learn complex relational patterns. These methods can encode nonlinear structural, temporal, and feature-based relationships to surface anomalies that are too subtle for rule-based or traditional statistical techniques. In coordinated fraud, representation learning captures higher-order similarities—such as actors coordinating across multiple content types, devices, and time periods simultaneously. However,

given their complexity and potential opacity, these models must be used with proper calibration and as part of a larger evidence-fusion system.

Together, these families offer a multi-layered strategy: from early burst detection to deep structural anomaly scoring and relational representation learning.

5.3 GNN-Based Fraud and Coordination Modeling

Graph neural networks have become a major focus of recent fraud detection research due to their ability to incorporate both network structure and node/edge features. A growing body of work shows that GNNs outperform traditional ML models when fraud involves multi-entity relationships, cross-layer coordination, or subtle structural cues [27]. GNNs can aggregate signals across neighbors, capture community-level patterns, and integrate heterogeneous information such as content embeddings, device fingerprints, and timing features.

However, deploying GNNs in operational fraud-detection environments introduces several practical challenges:

Temporal drift.

Campaigns adapt quickly, making their signatures time-dependent. Dynamic graph surveys emphasize that models trained on older data degrade when new coordination patterns emerge [10]. This necessitates retraining policies, continual learning modules, or hybrid systems where GNNs provide candidate signals rather than final judgments.

Explainability requirements.

In fraud and CIB contexts, decisions must be explainable for auditing, appeals, and regulatory compliance. GNN explainability surveys (e.g., the taxonomy in [26]) highlight the difficulty of attributing predictions to specific subgraph structures or features. For coordinated fraud, investigators must understand *which actors*, *which shared artifacts*, and *which timing patterns* triggered the alert. Thus, GNNs must be paired with evidence-extraction agents, not relied upon as black-box classifiers.

Calibration and false positives.

Studies on automated bot detection tools reveal significant false positive risks when models are not calibrated to diverse languages, platforms, or activity distributions [35]. In fraud detection, false positives carry economic and reputational impact, making calibrated thresholds, uncertainty estimation, and human-in-the-loop review essential. For this reason, our framework positions GNN outputs as one form of evidence feeding into the fused adjudication process, rather than as an authoritative classifier [43].

Overall, GNN-based approaches are powerful components of the detection pipeline but must be embedded within a structured, multi-agent system where their outputs are contextualized, calibrated, and complemented by more interpretable evidence sources.

6. Evaluation: datasets, baselines

6.1 Ground Truth Realities

Evaluating coordinated fraud detection systems requires confronting the inherent difficulty of obtaining high-quality ground truth. Verified labels for coordinated campaigns are scarce, often incomplete, and sometimes unreliable due to the covert nature of malicious operations. Studies in click-fraud AI have documented that many published evaluations depend heavily on small, synthetic, or artificially curated datasets, which limits the generalizability of their findings and inflates performance estimates [15]. Similarly, work on bot detection demonstrates that automated tools can produce substantial false positives, especially when applied across different linguistic or cultural contexts without proper adaptation [35]. Because of these limitations, robust evaluation cannot rely on a single labeling strategy. Instead, it should combine (a) known-case validation, using confirmed takedown datasets or previously investigated coordinated operations; (b) weak supervision, where partial labels, heuristics, and structural priors help guide the evaluation without claiming perfect certainty; and (c) counterfactual stress tests, which simulate adversarial variations such as shifted timing, rotated infrastructure, or randomized activation bursts to assess model robustness. Together, these strategies acknowledge the realistic uncertainty in fraud environments while ensuring meaningful performance assessment.

6.2 Metrics Beyond Accuracy

Traditional accuracy metrics are insufficient for evaluating systems that detect *campaign-level* phenomena rather than single-event predictions. Because the target of interest is the coordinated group or operation, metrics should emphasize cluster-level performance, including how well the system identifies whole campaigns (cluster precision/recall) and avoids splitting or merging them incorrectly. Another crucial dimension is time-to-detect, capturing how early the system surfaces a campaign relative to its actual onset. Research on temporal coordination patterns [28] highlights that the timing of coordinated behavior is often as important as its structure, making temporal stability across sliding windows an essential metric. Furthermore, because coordinated-fraud decisions frequently require human adjudication, explanation quality becomes a measurable outcome. Explanation-oriented evaluations—such as the extent to which investigators can successfully resolve cases using system-generated evidence summaries—are grounded in

explainability research [26]. Finally, any deployment in real-world platforms must systematically quantify false positive costs and perform fairness assessments, especially in light of bot-detection validity concerns showing how automated systems may mislabel minority language groups or niche communities at higher rates [35]. In combination, these metrics move evaluation beyond simplistic accuracy and toward operational impact.

6.3 Baselines and Comparators

A strong experimental design must be anchored in well-chosen baselines that reflect the diversity of existing methods across fraud detection research. Classical feature-only machine-learning baselines, widely used in click-fraud detection studies, remain important for demonstrating the incremental value of relational and temporal modeling [11]. Similarly, ensemble methods, which are common in impression fraud [17] and click fraud [14], offer competitive and interpretable benchmarks for real-world systems. Graph-based anomaly detection literature recommends GBAD (Graph-Based Anomaly Detection) baselines, as documented in systematic reviews [18], to evaluate structural irregularities relative to graph norms. To assess performance in temporally evolving environments, evaluations should include dynamic graph anomaly detection baselines drawn from survey recommendations [10,23], ensuring proper comparison across both structural and temporal modeling families. Finally, coordination-specific research has produced specialized methods and toolkits designed for detecting orchestrated behavior on social platforms [29,39], and these represent essential comparators for validating both accuracy and operational robustness. Together, these baselines create a comprehensive foundation for evaluating multi-agent, graph-based fraud detection systems.

7. Responsible Deployment Considerations

Detecting coordinated fraud is not purely a technical challenge; it is deeply intertwined with issues of governance, ethics, and potential misuse. Coordination-detection outputs can be misinterpreted as attribution tools, overstating intentionality or identity behind suspicious patterns. Review work on computational social manipulation (CSMM) warns that practical deployments face persistent concerns around transparency, proportionality, and evidentiary rigor [30]. Moreover, bot-detection validity studies clearly illustrate how automated tools can mislabel benign users—especially in multilingual settings—due to distributional shifts and contextual blind spots [35]. Accordingly, any operational system must incorporate safeguards that prevent over-reliance on automated outputs.

Key protections include human-in-the-loop adjudication, especially for high-impact actions such as account takedowns, monetization penalties, or enforcement escalations. Systems should support tiered response mechanisms that allow intermediate actions—rate limiting, friction, verification requests—before irreversible sanctions occur. To maintain transparency and auditability, platforms should store evidence objects associated with detection decisions rather than sensitive raw data, ensuring case reconstructability without unnecessary privacy exposure. Finally, periodic recalibration is essential. As coordination patterns evolve, drift monitoring and scheduled back-testing—consistent with anomaly detection best practices in survey literature [33]—help prevent degradation in accuracy or unintended bias amplification. Taken together, these safeguards ensure that coordinated-fraud detection technologies contribute to platform integrity without compromising fairness or accountability [44].

8. Open Research Directions

Several research challenges remain open, offering opportunities to improve the scientific and operational foundations of coordinated-fraud detection. First, cross-platform and cross-modal coordination graphs represent an important frontier. Modern campaigns integrate URLs, video snippets, images, and referral chains across multiple platforms, but fusing these modalities while preserving privacy remains unsolved. Second, emerging work on temporal archetypes and influence dynamics [28] suggests the need to move beyond static coordination detection toward causal and impact-based analysis—asking not only *who coordinated*, but *how that coordination altered downstream outcomes*. Third, as graph explainability methods expand [26], there is growing need for explainable campaign discovery pipelines that integrate structural explanations directly into investigator workflows, enabling human analysts to understand why a candidate campaign is flagged. Fourth, adversarial robustness remains challenging; spambot evolution research [13] shows how coordinated actors continually adapt to evade detectors, requiring systematic evaluation against realistic evasion strategies. Fifth, as multi-agent systems gain popularity, there is room to formalize agent calibration and disagreement modeling, studying how adjudication agents should aggregate conflicting evidence while maintaining uncertainty bounds—an area informed by anomaly-detection evaluation practice [33]. Finally, the broader ecosystem needs standardized operational toolchains for coordination analysis. Existing toolkits [29] are promising, but require reproducible evaluation environments, bias-sensitivity tests, and benchmark datasets that reflect the diverse contexts in which coordinated fraud occurs, as emphasized by fairness concerns in bot-detection research [35].

9. Conclusion

Coordinated fraud campaigns are best understood as multi-entity, temporal, and adaptive phenomena. The literature increasingly converges on coordination graphs and dynamic anomaly detection as core tools [1,7,10,18,28,29]. Yet practical detection requires more than a single model: it requires a system that ingests heterogeneous traces, constructs evolving coordination representations, discovers candidate campaigns, explains evidence, adjudicates risk, and monitors drift responsibly. A multi-agent AI framework provides a natural, modular way to operationalize this pipeline, blending graph-based detection strengths [18,23,27] with governance and explainability requirements [26,35]. The resulting architecture supports both research progress and safer deployment in real digital media ecosystems.

References

- [1] Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems*, 160, 113819.
- [2] Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management*, 57(4), 102250.
- [3] Sadeghpour, S., & Vlajic, N. (2021). Click fraud in digital advertising: A comprehensive survey. *Computers*, 10(12), 164.
- [4] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.
- [5] Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10), 72–83.
- [6] Latah, M. (2020). Detection of malicious social bots: A survey and a refined taxonomy. *Expert Systems with Applications*, 151, 113383.
- [7] Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6), 867–891.
- [8] Gruz, A., Mai, P., Soares, F. B., et al. (2022). (Coordination study in SNAM; coordinated online behavior). *Social Network Analysis and Mining*.
- [9] Murero, M., et al. (2023). An innovative manipulation tactic to amplify COVID-19 anti-... *Frontiers in Sociology*.
- [10] Ekle, O. A., Eberle, W., et al. (2024). Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, 18(8), Article 192, 1–44.
- [11] Aljabri, M., & Mohammad, R. M. A. (2023). Click fraud detection for online advertising using machine learning. *Egyptian Informatics Journal*, 24(2), 341–350.
- [12] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 193, 116429.
- [13] Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2019). On the capability of evolved spambots to evade detection via genetic engineering. *Online Social Networks and Media*, 9, 1–16.
- [14] Batool, A., & Byun, Y.-C. (2022). An ensemble architecture based on deep learning model for click fraud detection in pay-per-click advertisement campaign. *IEEE Access*, 10, 113410–113426.
- [15] Alzahrani, R. A., & Aljabri, M. (2022). AI-Based Techniques for Ad Click Fraud Detection and Prevention: Review and Research Directions. *Journal of Sensor and Actuator Networks*, 12(1), 4.
- [16] Sadeghpour, S., & Vlajic, N. (2021). Ads and fraud: A comprehensive survey of fraud in online advertising. *Journal of Cybersecurity and Privacy*, 1(4), 804–832.
- [17] Haider, C. M. R., Iqbal, A., Rahman, A. H., & Rahman, M. S. (2018). An ensemble learning based approach for impression fraud detection in mobile advertising. *Journal of Network and Computer Applications*, 112, 126–141.
- [18] Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133, 113303.
- [19] He, L., Wang, X., Chen, H., & Xu, G. (2022). Online spam review detection: A survey of literature. *Human-Centric Intelligent Systems*, 2, 14–30.
- [20] Hussain, N., Mirza, H. T., Rasool, G., Hussain, I., & Kaleem, M. (2019). Spam review detection techniques: A systematic literature review. *Applied Sciences*, 9(5), 987.
- [21] Duma, R. A., et al. (2024). Fake review detection techniques, issues, and future directions. *Knowledge and Information Systems*.
- [22] Wu, Y., et al. (2021). Fraudulent traffic detection in online advertising with graph models. *Expert Systems with Applications*.
- [23] Kim, H., Lee, B. S., Shin, W.-Y., & Lim, S. (2022). Graph anomaly detection with graph neural networks: Current status and challenges. *IEEE Access*, 10, 111820–111829.
- [24] Pazho, A. D., et al. (2024). A survey of graph-based deep learning for anomaly detection in distributed systems. *IEEE Transactions on Knowledge and Data Engineering*.

- [25] Qiao, H., Tong, H., An, B., King, I., Aggarwal, C. C., & Pang, G. (2025). Deep graph anomaly detection: A survey and new perspectives. *IEEE Transactions on Knowledge and Data Engineering*.
- [26] Yuan, H., Yu, H., Gui, S., & Ji, S. (2023). Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5), 5782–5799.
- [27] Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*, 240, 122156.
- [28] Tardelli, S., Nizzoli, L., Tesconi, M., Cresci, S., et al. (2024). Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. *Proceedings of the National Academy of Sciences*, 121(20), e2307038121.
- [29] Graham, T., Hames, S., & Alpert, E. (2024). The coordination network toolkit: A framework for detecting and analysing coordinated behaviour on social media. *Journal of Computational Social Science*, 7(2), 1139–1160.
- [30] Milzner, M., Thiele, D., & Gong, B. (2025). Just the tip of the iceberg? State of the art of coordinated social media manipulation research. *Information, Communication & Society*.
- [31] Li, X., et al. (2024). A survey on LLM-based multi-agent systems. *Springer (journal)*.
- [32] Maldonado Andrade, D. J., Cruz, E., & Abad Torres, J. (2024). Multi-Agent Systems: A survey about its components, framework and workflow. *IEEE Access*.
- [33] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15.
- [34] Chatterjee, A., & Ahmed, B. S. (2022). IoT anomaly detection methods and applications: A survey. *Internet of Things*, 19, 100568.
- [35] Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10), e0241045.
- [36] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2018). Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4), 561–576.
- [37] Yang, K.-C., Ferrara, E., & Menczer, F. (2022). Botometer 101: Social bot practicum for computational social scientists. *Journal of Computational Social Science*, 5(2), 1511–1528.
- [38] Zouzou, Y., et al. (2024). Unsupervised detection of coordinated fake-follower behavior. *EPJ Data Science*.
- [39] Magelinski, T., Ng, L., & Carley, K. M. (2022). A synchronized action framework for detection of coordination on social media. *Journal of Online Trust and Safety*, 1(2).
- [40] Priyanka Verma, N. K. (2020). FEDUS: A comprehensive algorithm for web usage mining. *Journal of Information and Optimization Sciences*, 41(3), 835-854. doi:10.1080/02522667.2019.1616912
- [41] Verma, P., & Keswani, N. (2013). Web usage mining: identification of trends followed by the user through neural network. *Int. J. Inf. Comput. Technol*, 3, 617-624.
- [42] P. Verma and N. Keswani, "Comparative analysis of algorithms for identification of session on the basis of threshold value," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 3724-3730.
- [43] Sharma, M., Verma, P. (2022). Review: Benefits of Mining Fashion Data. In: Rathore, V.S., Sharma, S.C., Tavares, J.M.R., Moreira, C., Surendiran, B. (eds) *Rising Threats in Expert Applications and Solutions. Lecture Notes in Networks and Systems*, vol 434. Springer, Singapore. https://doi.org/10.1007/978-981-19-1122-4_48
- [44] Verma, P., & Keswani, N. (2017). User identification prerequisite for memorisation a type of personalisation. *International Journal of Services Technology and Management*, 23(4), 290-298.