International Journal of Advanced Artificial Intelligence Research (IJAAIR)
eISSN: 3087-4270
Volume03 Issue11 | Page No: - 65-86

Received: 02 Oct 2026, Revised: 25 Oct 2026, Accepted: 09 Nov 2026, Published: 19 Nov 2026

DOI: https://doi.org/10.55640/ijaair-v03i11-05

Detecting Coordinated Inauthentic Behaviour Across Financial Technology and Media Platforms

Bharath Kandati

Independent Researcher,USA

ABSTRACT: Coordinated inauthentic behaviour is usually framed as a harder version of a familiar problem: given an account, a transaction, or a piece of content, decide whether it is fake. This paper argues that the framing is wrong, and that the error is expensive. Unusual is a property of an item. Coordinated is a property of a graph. A per-item review pipeline can be individually correct on every node in a campaign and still miss the campaign, because the evidence is in the edges, and the edges are the thing the adversary is actively hiding. The paper sets out three signatures that separate coordination from ordinary anomaly, describes at the level of signal families rather than features and weights an architecture for establishing that entities engineered to look independent are not, and sets out confirmation gates that let a system be fast to see networks and slow to act on them. Comparing a financial technology setting with a large-scale music streaming platform, it argues that the adversary's economics transfer almost perfectly while the defender's time budget does not: what differs is the reversibility window, meaning how long value remains recoverable after the act, and the due-process architecture available to a defender follows from it. The media claims are operator experience; the financial technology claims are informed observation from adjacent product work. The field keeps treating coordination as a bigger classification problem, when it is actually an economics problem, an accounting problem, and a governance problem wearing a classifier's clothes.

This is a cross-domain practitioner analysis: retrospective, written from a single operator vantage, with no controlled evaluation and no released operational metrics. Its claims are architectural rather than empirical, and Section 10 states what observation would overturn each of them.

KEYWORDS: coordinated inauthentic behaviour, trust and safety, graph-based fraud detection, adversarial adaptation, platform governance, streaming fraud, financial fraud detection, entity linkage, evidentiary standards

1. Introduction

Most published work on inauthentic activity on platforms asks a question of the form: is this thing fake? Is this account automated, is this transaction fraudulent, is this track machine generated. The question is tractable, it is easy to label, and it produces benchmarks. It is also, for the class of abuse that matters most commercially, the wrong unit of analysis.

Coordinated operations are not collections of individually suspicious items. They are collections of individually unremarkable items whose relationships are suspicious. A reviewer looking at one templated track sees a mediocre song and passes it; a reviewer looking at the campaign that produced it sees the operation instantly. The evidence never moved; the unit of inspection did. Any detection system whose unit of decision is smaller than the unit of the attack will be systematically, and quietly, wrong.

This paper is written from the operator's side of that problem. The primary vantage point is coordinated-fraud detection on a large-scale music streaming platform,

where the catalogue exceeds 250 million tracks and tens of thousands of new tracks are ingested daily, and where the abuse in question is the manufacture of catalogues and listening activity designed to extract royalty payments. A secondary vantage point, offered as informed observation rather than as operator experience, comes from several years of building small business banking products at a financial technology organisation, adjacent to the fraud and risk systems rather than inside them. The comparison in Section 5 is offered in that spirit, and the paper is explicit about which claims carry which weight.

Three commitments shape what follows.

The first is altitude discipline. Publishing a detection system's precise feature set is a courtesy to the people it detects. This paper therefore describes signal categories, connective methods, and design principles, and deliberately declines to specify features, thresholds, or weights. This is standard practice in published trust and safety work, and the categories plus the connective method are the transferable contribution in any case.

The second is that the employing platform is not named, and no product, vendor, tool, or colleague is identified. The paper refers to a large-scale music streaming platform and to a financial technology setting.

The third, and the one that most shapes the paper's character, is measurement honesty. A paper in this field that reports a precise harm-prevented figure is reporting a marketing estimate. The author holds no releasable operational numbers for this work: no networks-detected series, no account or content volumes beyond the catalogue and daily-ingestion scale stated above, no detection latency distribution, and no precision measurements. Rather than substitute estimates, illustrative figures, or a synthetic evaluation, the paper states its claims structurally and says plainly where a number would belong and why it is absent.

Those commitments define what kind of paper this is, which is worth naming rather than leaving to inference. This is a cross-domain practitioner analysis: a retrospective account of design decisions taken in production, written from a single operator vantage, with no controlled evaluation, no released metrics, and no benchmark. Its claims are architectural and are meant to be judged as architecture. The absence of an evaluation section is a property of that genre rather than a section that went missing, and Section 10 closes by stating, claim by claim, what observation would overturn the paper's central positions.

The contributions are:

- a practitioner taxonomy of what coordination looks like in platform data as distinct from ordinary anomaly (Section 3);

- an architecture-level account of signal construction and of how deliberately unrelated activity is linked (Section 4);

- a cross-domain comparison that isolates the defender's reversibility window, meaning how long value stays recoverable, as the axis on which financial and media platforms genuinely differ (Section 5);

- a treatment of adversarial adaptation as a rate rather than a state (Section 6);

- a five-gate confirmation model for acting against networks without acting against real users caught in them (Section 7);

- a composite case analysis of the campaign class that defeats short-window detection (Section 8);

- a field critique identifying five gaps, of which the last is that campaign-scale enforcement now operates something like a legal system at scale without having written down its rules of evidence (Section 9).

2. Related Work

Research on coordination on platforms has converged on a useful methodological insight: coordination is inferred from unexpected similarity of behaviour rather than from the content of any single item. Pacheco et al. formalise this as a general framework in which arbitrary behavioural traces are used to construct entity similarity networks whose dense regions are candidate coordinated groups [1]. Nizzoli et al. develop the complementary observation that coordination is a matter of degree rather than a binary label, and that thresholding a similarity network is itself a modelling decision with consequences for what gets surfaced [2]. Tardelli et al. extend the work to multiple coordination traces observed simultaneously,

arguing that operations are multifaceted and that single-trace detectors see only one facet [3].

This literature is the closest available analogue to the problem treated here, and its central methodological claim, that coordination is established by ruling out independence rather than by classifying items, is the claim this paper builds on. Two differences matter. First, most of this work targets information operations on social platforms, where the payoff is attention or narrative influence. The abuse examined here is directly monetary, and operations built to move money bend toward the payout function in ways that operations built to move opinion do not. Second, the academic work operates on public data with post hoc labels, while an operating platform must decide in production, under a precision constraint, against entities who will contest the decision.

The monetary case has a literature of its own, thin but directly on point. Drott's study of counterfeit attention in the streaming economy sets out the abuse economy this paper's primary vantage sits inside: spam uploads, click fraud, and manufactured or fake-artist catalogues assembled to divert payments from a royalty pool that is divided rather than grown [4]. It is an account of the incentive structure rather than of detection, and that division of labour is itself informative: the economics of streaming fraud have been described more carefully than the mechanics of detecting it, which is one reason the detection literature remains dominated by the attention-driven case.

The account-level literature on inauthenticity is mature. Cresci's retrospective on a decade of social bot detection documents the pattern that most concerns an operator: detectors that perform well on the generation of accounts they were trained against degrade against the next generation, and the field's progress is better described as a sequence of arms-race rounds than as cumulative accuracy gain [5]. That observation is the empirical basis for treating signal depreciation as a design parameter, which Section 6 does.

On the graph side, the industrial lineage is directly relevant. Beutel et al. detect lockstep behaviour as near-bipartite cores in a user-to-object graph, on the reasoning that legitimate populations do not act in unison on the same objects within the same windows [6]. Cao et al. generalise synchrony detection by clustering accounts on the similarity of their action timelines, explicitly to catch groups whose individual members are unremarkable [7]. Hooi et al. address the adversary's natural counter, camouflage, by giving dense-subgraph search a suspiciousness metric with provable bounds under the injection of legitimate-looking edges [8]. The Sybil-detection lineage supplies the complementary move on the identity side: Xiao et al. classify clusters of registered accounts rather than individual accounts, on the reasoning that fake accounts are created in batches and that the batch, not the account, is the object about which there is enough evidence to decide [9]. That is the same relocation of the unit of decision this paper argues for, arrived at from the registration side rather than the activity side. Each of these works answers the same structural question the present paper is organised around: how do you establish that a set of entities is not independent when each entity, examined alone, is fine.

On the financial side, graph learning over transaction and relationship networks has been applied to collusion and fraud-ring detection, using heterogeneous graph neural architectures to score groups rather than individual accounts [10]. Work on anti-money-laundering detection illustrates a different constraint: extreme class imbalance, a regulatory consumer of the output whose requirements set the precision bar, and an investigative posture that is deliberately post hoc rather than inline [11]. Both are useful points of comparison for Section 5, because the geometry of the problem is shared while the time available to act on it is not.

Two strands from outside the abuse literature carry more weight in practice than their citation counts in it would suggest. Dalvi et al. framed classification against an adaptive opponent as a game rather than as a stationary learning problem, which is the correct formal frame for a detection system whose signals are disclosed by its own enforcement actions [12]; the adversarial machine learning work that followed established how cheaply a deployed model can be evaded once its behaviour is observable, and how reliably defences built against a fixed attacker are re-broken by an attacker who adapts to them [13]. Gama et al.'s treatment of concept drift supplies the vocabulary for what happens next, although the abuse case is harder than the canonical drift setting because the drift is adversarial and is triggered by the detector's own decisions rather than by exogenous change [14]. Separately, the probabilistic record-linkage

tradition initiated by Fellegi and Sunter provides the inferential move that Section 4 relies on most heavily: deciding that two records refer to the same underlying entity by comparing the observed agreement pattern against what agreement would look like under independence [15]. Coordination detection is that inference applied to behaviour rather than to identifiers.

Finally, three literatures bracket the parts of the problem the field handles least well. Campos et al.'s study of unsupervised outlier-detection evaluation documents how badly measurement behaves when the ground-truth denominator is unknown, which is exactly the situation of a trust and safety team asked for a detection rate [16]. That situation has a formal name: it is a positive-unlabelled problem, in which only a biased subset of the positives is ever labelled and the size of the unlabelled positive mass is itself a quantity to be estimated rather than assumed away, and the survey literature on positive-unlabelled learning treats prevalence estimation as a first-class question with stated assumptions [17]. Section 9.2 argues that trust and safety reporting has that problem and proceeds as though it has neither. And Gorwa et al.'s analysis of algorithmic content moderation identifies the governance dimension that Section 9 argues is now the binding constraint: automated enforcement at platform scale raises questions of contestability, proportionality, and legitimacy that the technical literature does not address and that platforms are answering ad hoc [18].

3. What Coordination Looks Like in Practice

The practical starting point is that ordinary unusual behaviour and coordinated inauthentic behaviour do not look alike in the data, and that the difference is legible if you look at the right object.

Ordinary unusual behaviour is unusual in one dimension and it is independent. A real artist uploads forty tracks in one day because the back catalogue finally got digitised. A track spikes because it was used in a video that travelled. A new account behaves strangely because its owner does not yet know the conventions. Each of these is anomalous in isolation, and each is explained by a story about one actor doing one thing. Anomaly detection built for this case works well, because the case really is a matter of an item deviating from a distribution.

Coordination produces a different fingerprint, along three axes.

3.1 Templated variation

Legitimate content is diverse in ways that are hard to fake cheaply. Fraud operations produce variation from a template. Metadata differs by permutation rather than by authorship. Artist names are drawn from the same generative distribution. Artwork shares structure. Track lengths cluster where a generation tool's defaults sit. Individually, every item passes inspection. In aggregate, the catalogue has the fingerprint of a process rather than of a population.

The important refinement is that the tell is not similarity as such. Genuine populations are full of similarity: genres have conventions, scenes have aesthetics, producers have signatures. The tell is similarity in dimensions where genuine populations are messy. Real catalogues are noisy in the places where a template is tidy, and a detector calibrated on the wrong dimensions will flag a coherent scene and miss a coherent factory.

3.2 Synchronisation without a cause

Real-world events synchronise legitimate behaviour all the time. A release date, a playlist placement, a cultural moment: each produces correlated activity across accounts, and each has an observable trigger. Coordinated campaigns synchronise because they share an operator. Upload bursts appear across nominally unrelated accounts inside the same window. Streaming activity rises and falls together across catalogues that have no shared audience. Accounts created in batches activate in batches.

The operational rule is therefore not "correlation is suspicious" but "correlation without a locatable cause is suspicious." When behaviour correlates and the event that would explain it cannot be found, the missing cause is the finding. This is also the discipline that keeps the signal usable, because searching for the innocent trigger is cheap relative to the cost of acting on a legitimate community that happened to move together.

3.3 Economic directionality

Legitimate unusual behaviour is economically random. It does not consistently optimise anything.

Coordination exists to move value, so its patterns bend toward the payout function. Activity concentrates at monetisation thresholds. Engagement is distributed to maximise royalty extraction rather than to follow any plausible listening pattern. Catalogues are structured around what pays rather than around what an audience would want. If an incentive gradient can be read off the behaviour, someone is climbing it.

Of the three signatures, this one is the most durable, and Section 6 explains why: it is the signature an adversary cannot remove without giving up the profit that motivates the operation.

The distinction that governs system design is this: unusual is a property of an item, and coordinated is a property of a graph. Per-item review systematically misses coordination, because every node survives inspection. The evidence is not in any node. It is in the edges.

This is why the review model on the media platform moved from items to clusters. It is not a presentational change. It changes what a human is asked to adjudicate, it changes what evidence has to be assembled before the human sees it, and, as Section 7 discusses, it changes the stakes of being wrong.

3.4 The operational distinction

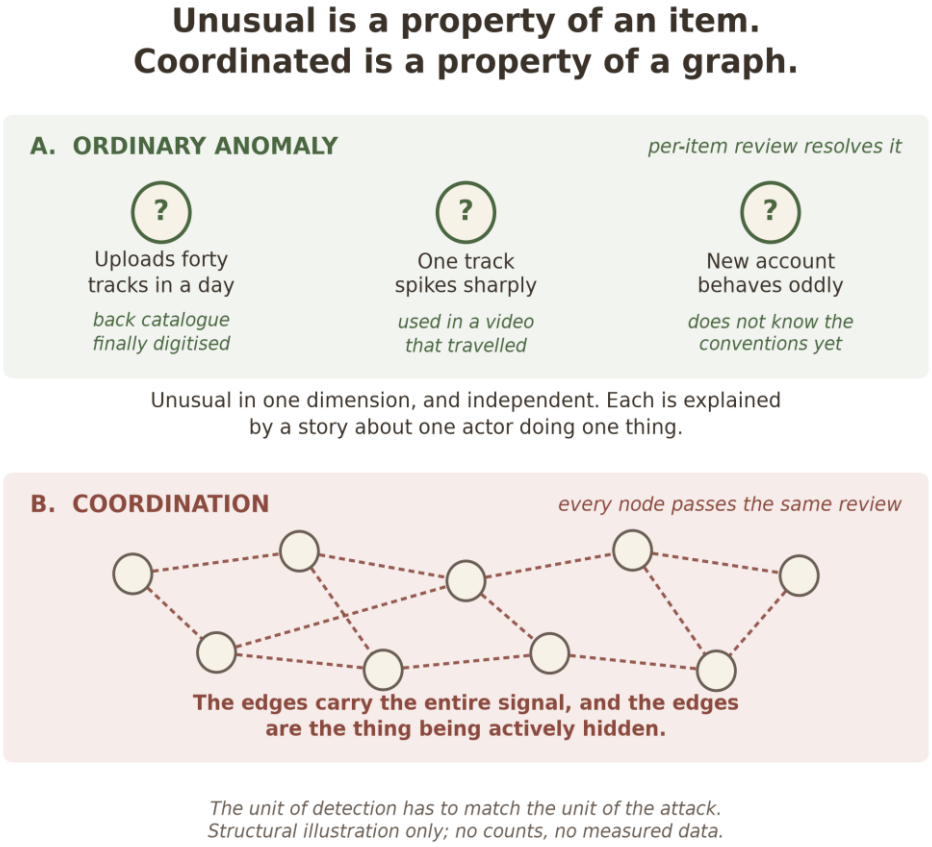


Figure 1. Ordinary anomaly versus coordination. In panel A, three accounts are each unusual in one dimension and each has an independent explanation; per-item review resolves them correctly. In panel B, every account is individually unremarkable and would pass the same review, while the relationships between them carry the entire signal. The figure renders the structural distinction described in Section 3.4. It is not derived from measured data and reports no counts.

4. Detection Approach and Signal Construction

This section is written at the architectural level, for the reason given in Section 1: features, thresholds, and weights are the parts an adversary most wants published, and publishing them converts a defensive asset into a specification for evasion. What remains is the level at which the work is transferable in any case, namely which families of signal are worth building, why the families are chosen for independence rather than for individual strength, and how activity engineered to look unrelated is linked.

4.1 Four signal families, chosen because they fail independently

The portfolio is organised into four families. The selection criterion is not predictive power in isolation. It is that the families have different failure modes, so that an adversary's defeat of one does not travel to the others.

Content signals are properties of the audio itself, scored by multiple detection approaches with deliberately different failure modes. Disagreement between those approaches is itself information, and is treated as such rather than being averaged away.

Behavioural signals describe how accounts act over time: upload cadence, catalogue growth patterns, activity rhythms. Behaviour is expensive to fake at scale, and the reason is economic rather than technical. Faking behaviour well means behaving like a real artist, and behaving like a real artist destroys the economics of the fraud.

Metadata and structural signals concern not what the metadata says but how it is shaped: its diversity, its consistency with the claimed identity, and whether variation across a catalogue looks authored or generated. This family is the cheapest for an adversary to rotate, which is precisely why it is never load bearing on its own.

Identity and provenance signals cover verification status, account history, distribution pathway, and accumulated track record. This is the anchor layer, because legitimate identity is slow to build and burns on first misuse. It is also the family with the longest memory, which matters for Section 4.2.

4.2 Connecting activity that is made to look unrelated

The adversary's core investment is making linked things look independent: separate accounts, separate distributors, staggered timing. The counter-investment is the recognition that independence is testable. Three methods carry most of the weight.

Shared-infrastructure convergence. Operations separate their identities but economise somewhere, because economising is the point of running an operation at scale. Wherever they economise, artifacts repeat: pathways, tooling, process. Any single shared artifact is coincidence. The same set of entities converging across multiple independent dimensions is a graph edge. The design rule that follows is that convergence is required across signal families rather than within one, because cross-family convergence is what is expensive for an adversary to prevent. Preventing it means economising nowhere.

Statistical impossibility of independence. It is not necessary to know how accounts are linked in order to establish that they are. Take the observed correlation, synchronised activity, templated variation, coupled growth, and ask how probable that pattern is under genuine independence. When the answer is vanishingly improbable, the burden flips. This is the workhorse of the architecture, and its value is that it detects the link without needing the mechanism, which matters because the mechanism is the part adversaries rotate. Formally this is the same inferential move as probabilistic record linkage, where an agreement pattern is scored against its likelihood under the hypothesis that two records describe different entities [15]; the difference is that the compared attributes are behavioural traces rather than identifiers.

Time as the unforgeable dimension. An operation can randomise almost everything going forward. It cannot retroactively reshape its history. Longitudinal structure, when accounts were created relative to one another, how catalogues grew, what happened after prior enforcement, is where campaigns that look independent today reveal the common origin they had yesterday. One pattern is worth stating because it is not obvious in advance: enforcement events are themselves revelatory. When action against one account produced correlated behaviour changes in nominally unrelated accounts, the reaction was the edge.

4.3 The binding rule

No single signal family, however alarmed, drives enforcement alone. Convergence earns automation; divergence earns a human; and the cluster, not the item, is the unit of decision.

The reason to state this as a rule rather than as a heuristic is that it defines the adversary's task. Defeating this architecture requires being independent in every dimension simultaneously, and continuing to be, which is another way of saying: it requires ceasing to be a

coordinated operation.

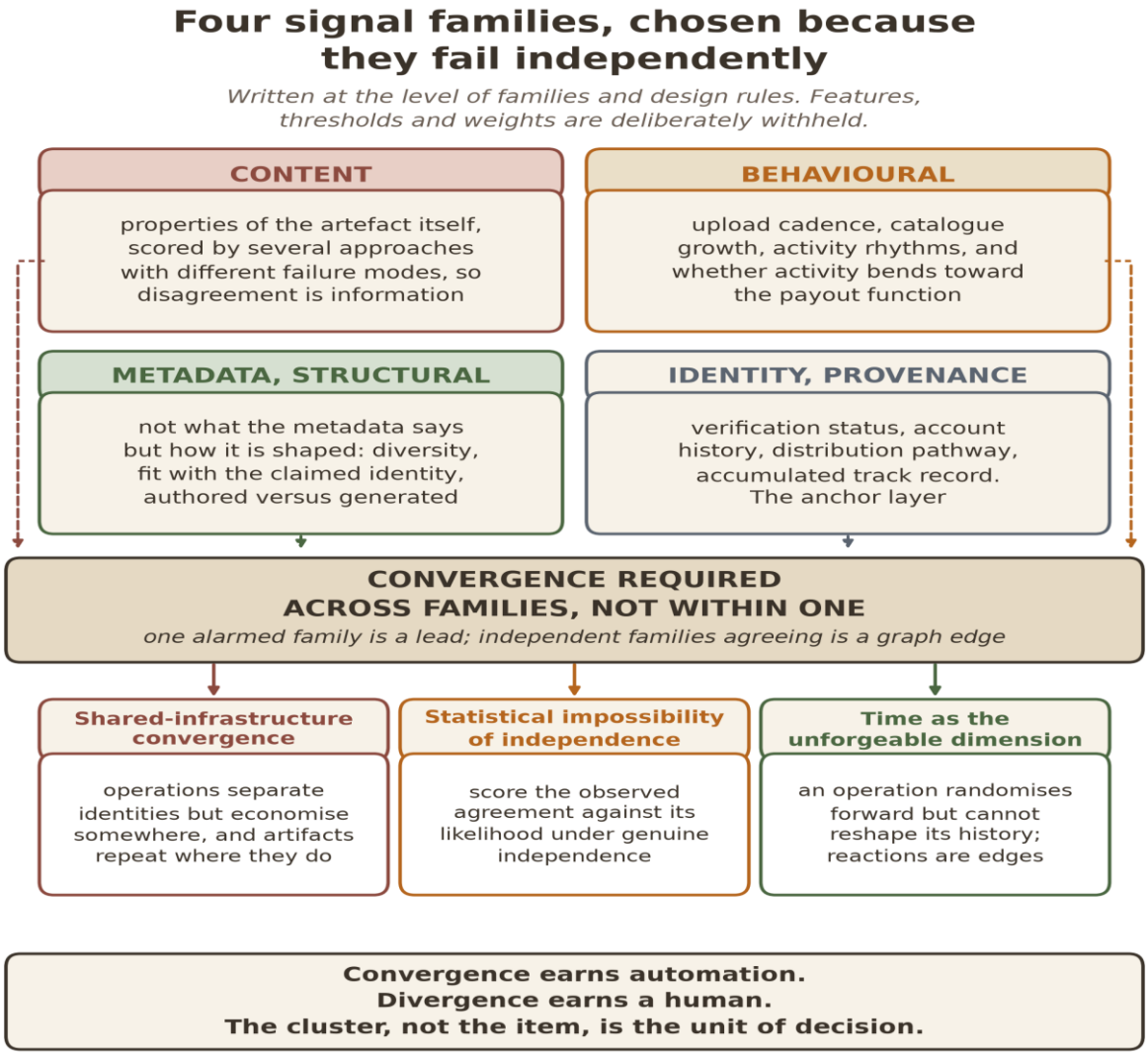


Figure 2. The four signal families and the convergence requirement. Each family observes a different aspect of the same entity and fails in a different way. A single alarmed family produces a lead, not a decision; an edge is asserted only where independent families converge on the same set of entities. The figure renders the architecture described in Sections 4.1 to 4.3. The families are drawn in the order they are introduced in Section 4.1; the figure asserts no ranking of any kind. The rotation-cost ordering discussed in Section 6 is a separate judgement, is not a measured ranking, and is given in Table 2 rather than here.

4.4 Coverage as an architectural requirement

Coverage is not a scale boast in this setting; it is a correctness condition. A coordination detector cannot sample, because the edges that carry the evidence may connect any two entities in the population. Sampling does not degrade a coordination detector gracefully. It tells campaigns where to distribute themselves.

Concretely, on the media platform, coverage was total by design: the full catalogue of more than 250 million tracks was scored, including a retroactive historical backfill of material that had never been scored, and the tens of thousands of new tracks arriving daily were decided at ingestion rather than after accrual. Detection operated over the full artist and uploader population associated with that catalogue, because cluster detection requires the whole graph rather than a sample of it. No account-population figure is published here; the author declines to estimate one.

The transactional analogue of a payment here is a

stream feeding royalty calculation. The scope of the work described is content and entities at catalogue scale, upstream of the royalty pipeline, delivering decisions before payout finality. The internals and volumes of that pipeline are outside the author's operating experience

and are not characterised.

4.5 What changed structurally when the unit of decision changed

Moving from item-level to cluster-level review changes the leverage of a human decision rather than its accuracy. One decision resolves an operation instead of a single track. The design consequence, stated as a structural property and not as a measured outcome, is that review demand is driven by the number of operations rather than by the volume of content: if the unit of decision matches the unit of the attack, decisions scale with operations rather than with content.

What that produced in practice is deliberately not reported. The author's operational account is that review demand grew more slowly than catalogue volume and that the band requiring human adjudication remained small, but both are comparative magnitudes, neither was retained as a figure, and neither is offered here as a result. The corresponding ratios, items resolved per reviewer decision before and after, or the share of traffic that ever touches a human, are not published because no releasable measurement of them exists.

5. Cross-Domain Comparison: Financial Technology and Media

A note on vantage point, because it determines how much weight each part of this section should carry. The author built and operated coordinated-fraud detection on the media side. The financial platform experience, several years building small business banking products, was adjacent to fraud and risk systems as a product engineering leader rather than as the person who built them. The media claims are therefore made with operator authority; the fintech claims are informed observation, and the comparison is offered in that spirit.

5.1 The deep difference is the reversibility window

The axis on which the two domains actually differ is the reversibility window: how long, after the act that moves value, that value remains recoverable by the defender. It is a spectrum rather than a switch, and where an instrument sits on it determines what kind of detection decision is available at all.

At the short end sit the instruments whose transfer is effectively final on completion. A wire, or a push payment on an instant rail, is gone once it lands, and recovery depends on the cooperation of a counterparty rather than on a rule the defender can invoke. For those instruments the attack and the value transfer are the same event, the decision has to fit inside the transaction path, the posture is interdiction, and the cost of friction is a declined legitimate payment that the customer notices immediately.

Further along sit instruments carrying a defined reversal mechanism. Card transactions can be disputed and reversed under network rules for a substantial window after authorisation, and clearing-house debits carry a return window, longer again for unauthorised consumer debits. Different in kind is the posture anti-money-laundering work occupies, which is deliberately post hoc: the obligation attaches after the transaction, and the product of the analysis is a case rather than a block [11]. A financial platform therefore does not run one posture. It runs interdiction on its final instruments, dispute and return on its reversible ones, and investigation over both.

A media platform sits at the far end of the same spectrum rather than in a different category. A fraudulent stream is worth a fraction of a cent, so the attack only works as a campaign accumulating over weeks, and payouts settle on cycles. Detection does not have to win in milliseconds; it has to win before the payout cycle closes, and the posture is attribution and clawback: connect the campaign, and un-move the money before or shortly after it moves.

Stated as a trade, the further along the window a defender sits, the more time is available for confirmation and the less atomic the decision becomes. What the financial platform has at the short end is an atomicity the media platform never gets. What the media platform has is a long window applying uniformly across the whole surface, and that uniformity is what makes a deliberate multi-gate confirmation process affordable rather than exceptional. The same detection idea therefore gets a different verdict in the two settings, because the binding constraint is not accuracy in either case; it is how long value stays recoverable.

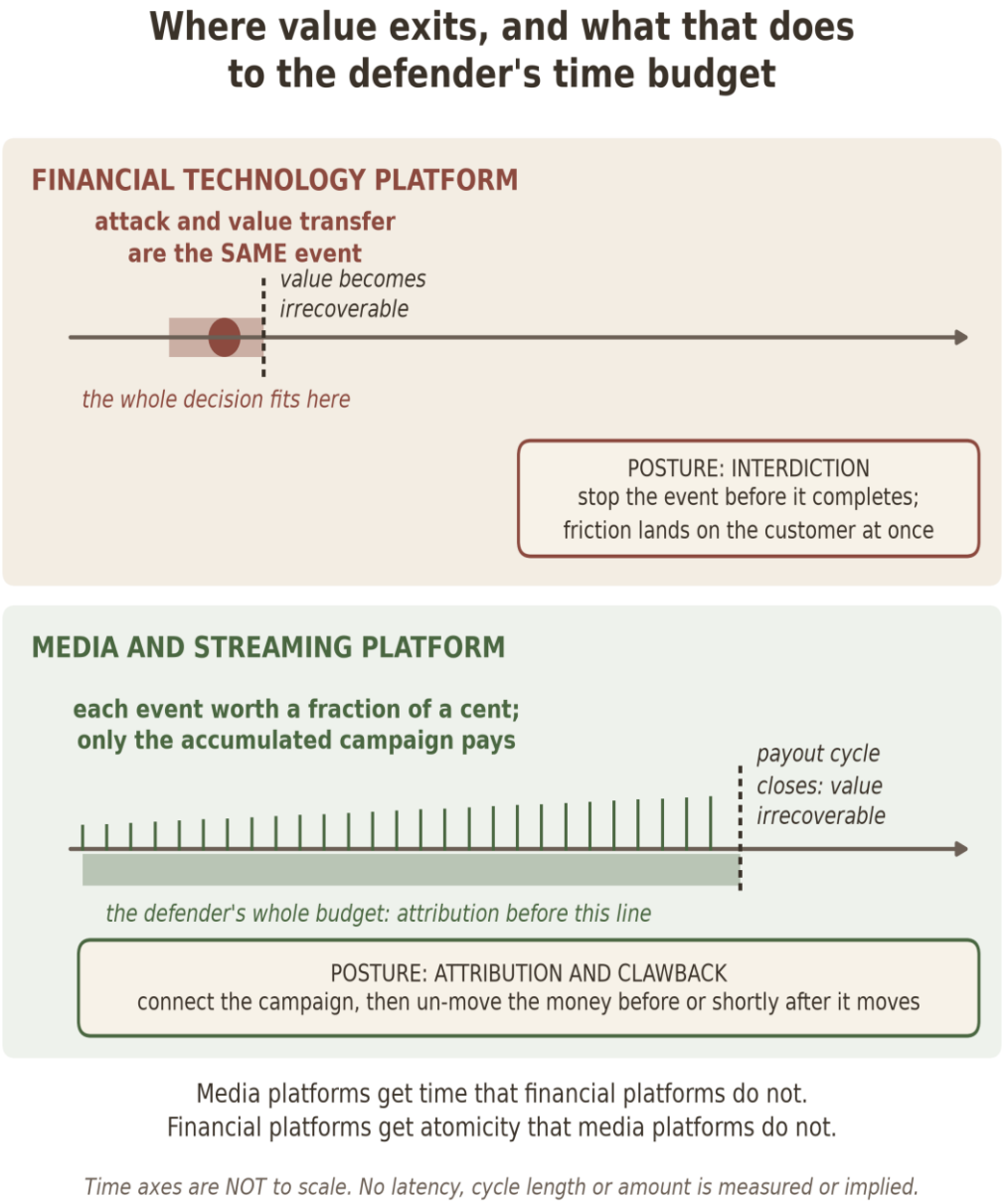


Figure 3. The reversibility window and the defender's resulting time budget. Above, the short end of the window: for instruments whose transfer is final on completion, the attack and the value transfer are the same event, so the entire decision must fit inside the transaction and the posture is interdiction. Below, the long end: value accrues in fractions of a cent across weeks and settles on a payout cycle, so the deadline is attribution before payout finality and the posture is attribution with clawback. The two paths are the ends of a spectrum rather than two categories; instruments carrying defined reversal mechanisms sit between them, as Section 5.1 sets out. The financial technology path is drawn from informed observation of adjacent product work rather than from operating fraud and risk systems. No latencies, cycle lengths, or amounts are measured or implied.

5.2 Two second-order differences

Identity anchoring differs at the entry point. Financial platforms have know-your-customer requirements, a regulatory identity floor that fraud has to defeat upfront. Media platforms have nearly frictionless account creation, so identity is earned post hoc rather than gated at entry. This is why the identity and provenance family of Section 4.1 behaves differently across the two domains: in fintech it is largely a gate, in media a slowly accumulating asset whose value to the defender comes precisely from the fact that it burns on first misuse.

The asymmetry of harms differs in texture rather than in magnitude. A false positive in fintech blocks someone's money now: acute, immediate, self-announcing, because the customer finds out at the point of sale and tells you. A false positive in media demonetises or flags an artist, which is slower to be noticed but reputationally deeper

and publicly combustible. Neither is the mild case, and a system tuned for one texture of harm will be badly calibrated for the other.

5.3 What transfers, which is most of it

Coordination is a graph property everywhere. Templated variation, synchronisation without cause, and economic directionality are properties of an operation rather than of a payout mechanism, and the graph-learning work on collusion and fraud rings in financial networks is organised around the same object: a group scored as a group, because the individual account carries too little evidence to decide on [10]. The payout function differs. The geometry does not.

The economic choke point. Both domains win not by detecting every item but by raising the cost of appearing independent past the margin the fraud earns. Fintech does this with velocity limits and step-up friction; the media system described here did it with cluster-level enforcement. Same lever, different handle.

Confidence-banded automation with humans on the uncertain slice. Its universality is not an accident. The cost of an error and the cost of a review are different currencies, so the only rational design is to spend review where the error cost is high and confidence is low.

The strictest consumer sets the bar. Engineering to the hardest requirement once is what lets an entire platform act on a signal, because a signal that only some consumers trust ends up re-derived, inconsistently, by everyone else.

5.4 The transferable thesis

The adversary's economics transfer almost perfectly between domains. What changes is the defender's reversibility window. Design the system around that window and the rest follows: interdiction where the transfer is final, dispute and return where a reversal mechanism exists, attribution and clawback where value accumulates, and a confirmation process whose depth is bounded by the window available rather than by the sophistication of the detector. Because the financial technology half of this comparison is informed observation rather than operator experience, the thesis is offered as a design principle to be tested and not as an established finding.

Dimension Financial technology platform Media and streaming platform Reversibility window Short and instrument-dependent: final on completion for wires and instant push rails, defined reversal windows for cards and clearing-house debits Long and uniform: value accrues across a campaign and settles on a payout cycle

Relationship of attack to value transfer At the short end, the attack and the transfer are the same event Slow and aggregated across a campaign

Value of a single fraudulent action Material on its own A fraction of a cent; only a campaign is worth running

Defender's deadline Inside the transaction path for final instruments, inside the reversal window for the rest Before the payout cycle closes

System posture Several at once: interdiction, dispute and return, and post hoc investigation over both Attribution and clawback

Identity anchoring Regulatory identity floor at entry, defeated upfront Nearly frictionless creation; identity earned post hoc through verification and track record

Texture of a false positive Blocks money now; acute, immediate, self-announcing Demonetises or flags an artist; slower to notice, reputationally deeper, publicly combustible

Economic choke-point instrument Velocity limits, step-up friction Cluster-level enforcement

Strictest consumer that sets the precision bar

The regulator The most precision-demanding downstream team

What is shared The three coordination signatures; group-level rather than item-level scoring; confidence-banded automation with humans on the uncertain slice The same three signatures and the same automation pattern, applied to catalogue and listening behaviour

Dimension	Financial technology platform	Media and streaming platform

Reversibility window	Short and instrument-dependent: final on completion for wires and instant push rails, defined reversal windows for cards and clearing-house debits	Long and uniform: value accrues across a campaign and settles on a payout cycle
Relationship of attack to value transfer	At the short end, the attack and the transfer are the same event	Slow and aggregated across a campaign
Value of a single fraudulent action	Material on its own	A fraction of a cent; only a campaign is worth running
Defender's deadline	Inside the transaction path for final instruments, inside the reversal window for the rest	Before the payout cycle closes
System posture	Several at once: interdiction, dispute and return, and post hoc investigation over both	Attribution and clawback
Identity anchoring	Regulatory identity floor at entry, defeated upfront	Nearly frictionless creation; identity earned post hoc through verification and track record
Texture of a false positive	Blocks money now; acute, immediate, self-announcing	Demonetises or flags an artist; slower to notice, reputationally deeper, publicly combustible
Economic choke-point instrument	Velocity limits, step-up friction	Cluster-level enforcement
Strictest consumer that sets the precision bar	The regulator	The most precision-demanding downstream team
What is shared	The three coordination signatures; group-level rather than item-level scoring; confidence-banded automation with humans on the uncertain slice	The same three signatures and the same automation pattern, applied to catalogue and listening behaviour

Table 1. Cross-domain comparison of coordinated abuse and the systems built against it. Media-side entries are drawn from operating experience; financial technology entries are informed observation from adjacent product work, as stated in Section 5. The table contains no measurements.

6. Adversarial Adaptation

Adaptation was the operating condition, not an edge case. The starting assumption of the design is uncomfortable and worth stating without softening: every enforcement action is also a disclosure. The adversary learns from what gets caught, rotates the cheap parts of the operation, and returns. A detection system that treats its current signals as durable assets is a depreciating one. This is the same dynamic the bot-detection literature has documented across successive generations of automated accounts [5], and the correct formal frame for it is adversarial rather than stationary learning [12], [13].

Building for that depreciation explicitly meant five things.

6.1 Depend on what is expensive to change, not on what is easy to observe

Every signal has a rotation cost for the adversary, and those costs differ by orders of magnitude. Metadata patterns rotate in an afternoon. Tooling and process rotate in weeks. But the economics of the operation, the need for scale, for templated production, for behaviour that bends toward the payout function, cannot rotate without abandoning the fraud's profitability.

The portfolio was therefore deliberately weighted

toward economically anchored signals, so that adaptation forced the adversary into the one move that actually defeats them: becoming too expensive to run. The sign that this is working is not a rising detection count. It is evasion attempts that make campaigns smaller. That is the criterion the portfolio was designed against, and it is stated here as a criterion: whether it was met is not reported, because no measurement of campaign size over time is available.

Signal family What it observes (Section 4.1) Adversary's cost to rotate it

Metadata and structural The shape of metadata across a catalogue Low. Patterns can be re-randomised in an afternoon, which is why this family is never load bearing alone Content Properties of the artefact itself Moderate. Requires changing generation tooling, and defeating several scorers with different failure modes at once

Identity and provenance Verification, account history, pathway, track record High. Legitimate identity is slow to build and burns on first misuse, so it cannot be rebuilt at the speed of enforcement

Behavioural and economic Cadence, catalogue growth, and the direction activity bends in Highest. Faking it well means behaving like a real artist and distributing activity as an audience would, which removes the profit that motivates the operation

Dimension	Financial technology platform	Media and streaming platform
Reversibility window	Short and instrument-dependent: final on completion for wires and instant push rails, defined reversal windows for cards and clearing-house debits	Long and uniform: value accrues across a campaign and settles on a payout cycle
Relationship of attack to value transfer	At the short end, the attack and the transfer are the same event	Slow and aggregated across a campaign
Value of a single fraudulent action	Material on its own	A fraction of a cent; only a campaign is worth running
Defender's deadline	Inside the transaction path for final instruments, inside the reversal window for the rest	Before the payout cycle closes
System posture	Several at once: interdiction, dispute and return, and post hoc investigation over both	Attribution and clawback
Identity anchoring	Regulatory identity floor at entry, defeated upfront	Nearly frictionless creation; identity earned post hoc through verification and track record
Texture of a false positive	Blocks money now; acute, immediate, self-announcing	Demonetises or flags an artist; slower to notice, reputationally deeper, publicly combustible
Economic choke-point instrument	Velocity limits, step-up friction	Cluster-level enforcement
Strictest consumer that sets the precision bar	The regulator	The most precision-demanding downstream team
What is shared	The three coordination signatures; group-level rather than item-level scoring; confidence-banded automation with humans on the uncertain slice	The same three signatures and the same automation pattern, applied to catalogue and listening behaviour

Table 2. Signal families ordered by the author's assessment of the adversary's rotation cost, per Section 6.1. Rotation cost is an analytical judgement grounded in operating experience, not a measured quantity. The extreme rungs, metadata at the bottom and behavioural economics at the top, are stated directly from that experience; the relative placement of the two middle rows is an inference rather than a judgement the author advances independently, and it is the part of this table most exposed to being wrong. The paper reports no measured cost, latency, or yield figures for any family.

6.2 Orthogonal families as rotation insurance

Because the four families fail independently, an adversary adapting to one typically becomes more visible

in another. Randomising metadata to defeat template detection produces unnatural diversity that behavioural signals catch. Staggering uploads to defeat synchronisation stretches the campaign across time, where longitudinal structure accumulates. The architecture converts the adversary's adaptation energy into new signal, which is a much better property than raw accuracy, because accuracy is a snapshot and this is a mechanism.

6.3 Enforcement reactions as detection

A signal appeared after every enforcement wave, and the team came to weight it heavily: the reaction itself. Nominally unrelated accounts going quiet, migrating pathways, or restructuring in correlated ways after one cluster was actioned. Adversaries reveal their graph precisely when they respond to being caught, because the response is coordinated by the same

operator that coordinated the campaign.

The operational corollary is a scheduling one: post-enforcement windows are collection opportunities, not

victory laps. A team that stands down after a successful action discards its highest-yield observation period.

6.4 Appeals and misses as the retraining loop

Overturned decisions indicated where confidence was miscalibrated. Fraud that surfaced late indicated which band it had slipped through. Both flowed back as labelled data on a cadence, because a model that ships once is a model an adversary gets to study at leisure. This is drift handling in the sense of Gama et al. [14], with the important difference that the drift is triggered by the detector's own actions rather than by exogenous change in the world, so the retraining cadence has to be coupled to the enforcement cadence rather than to the calendar.

The asset was never any individual model. It was the pipeline that re-derived models from fresh adversarial contact.

6.5 The review queue as early-warning radar

Before a new tactic shows up in enforcement outcomes, it shows up as drift in the uncertain band: a category flooding human review that used to resolve automatically. Queue composition was read as the leading indicator of adversary adaptation, and modelling investment was directed there. The reviewers were not only resolving cases. They were the sensor for where the automated system was going stale.

6.6 Effectiveness is a rate, not a state

The honest framing is that you do not stay ahead of adaptive adversaries by having better signals, because your signals decay the moment you use them. You stay ahead by having a faster loop, and by anchoring detection in costs the adversary cannot shed. One symmetry is worth conceding here rather than leaving to a reader: Section 9.2 faults the field for not instrumenting the adversary's cost curve, and this paper does not instrument it either. Naming the right quantity is not the same as having measured it.

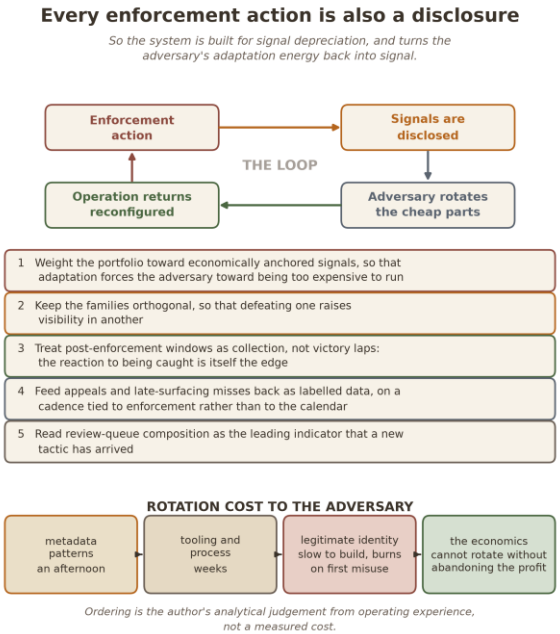


Figure 4. The adaptation loop and the five mechanisms that exploit it. Enforcement discloses; the adversary rotates the cheap parts and returns. The mechanisms of Sections 6.1 to 6.5 are placed at the point in the loop where each intervenes, so that the adversary's adaptation energy is converted into signal rather than only into evasion. The rotation-cost ladder on the right reproduces the ordering in Table 2, an analytical judgement rather than a measurement.

7. Confirming Coordination Before Action

This is the question the whole architecture exists to answer safely, and it is the question that cluster-level enforcement makes acute. A false positive at item level is one track. A false positive at cluster level is a real artist's entire catalogue swept into someone else's fraud case. Confirmation was therefore built as a series of gates, each designed to catch a different way of being wrong.

7.1 Gate one: convergence across independent signal families

No network was treated as coordinated on the strength of one signal family, however loud. The reason is not statistical conservatism. It is that legitimate music is full of coordination. Scenes, labels, collectives, and producers who work with many artists all correlate, and they correlate in ways a synchrony detector cannot distinguish from an operation.

The requirement was convergence: content signals, behavioural structure, and pathway evidence

independently agreeing. Innocent explanations tend to satisfy one family and fail the others. A record label looks coordinated behaviourally, but its content is diverse and its identity is anchored. A fraud operation is coordinated in every dimension at once. Convergence is therefore not merely stronger evidence; it is evidence with a different discriminative target.

7.2 Gate two: the innocent-explanation hunt

Before a cluster moved toward enforcement, an explicit analytical step ran adversarially against the case itself: what legitimate structure could produce this pattern? A distributor's batch migration produces synchronised uploads. A production house produces templated metadata. A viral moment produces correlated streaming.

The burden was on the case to rule out the legible innocent causes, not merely to accumulate suspicion. Statistical impossibility of independence is where an investigation starts, never where it ends. This gate is the one most often skipped in published methods, and it is the reason threshold selection in similarity networks is a substantive modelling decision rather than a tuning detail [2].

7.3 Gate three: human confirmation at the cluster level, with the case laid out

Confirmed-coordination decisions were human decisions. The reviewer received the assembled case: which families converged, at what confidence, and which innocent explanations were tested and failed. Critically, the reviewer examined the operation, where fraud is obvious and legitimacy is also obvious, rather than adjudicating individual items, where neither shows.

The attribution trail was not paperwork. It was the mechanism that made a human able to genuinely confirm rather than rubber-stamp a score. This distinction matters for any claim about human oversight of automated enforcement: a human who cannot see why the system decided is not oversight, and the literature on algorithmic moderation is right to treat that as a governance problem rather than an interface problem [18].

7.4 Gate four: blast-radius discipline inside confirmed networks

Confirmation of the network is not confirmation of every node. Real accounts get entangled in fraudulent graphs: a legitimate artist distributed by a dirty pipeline, a real producer who touched a bad catalogue.

Enforcement was therefore scoped to the evidence. Nodes carrying direct evidence were actioned. Peripheral nodes connected only by association were treated as leads rather than as targets. Guilt by graph adjacency was explicitly not a standard. This is the gate that most directly separates a detection result from an enforcement decision, and it is the one that dense-subgraph methods do not supply on their own, since a suspiciousness score for a subgraph says nothing about the standing of an individual node inside it [8].

7.5 Gate five: the appeal as the final instrument

Every actioned entity could contest, reaching a human with the full case visible. Overturns fed back as the highest-value calibration data the system produced. An appeal surviving adversarial review told us the gates worked. An overturn told us precisely which gate failed.

An appeals channel is therefore not only a fairness obligation; it is the only reliable source of labelled evidence about the system's own errors. Research on how appeals actually function on large platforms suggests the mechanism is often nominal rather than substantive [19], which, from a purely instrumental standpoint, means those platforms are also discarding their best calibration signal.

7.6 The design stance: asymmetric burden of proof

Underneath the five gates is a single stance. Suspicion was cheap to raise and expensive to act on. Automation could accumulate the case, but only convergence, plus a ruled-out innocent explanation, plus a human, could spend it.

The system was allowed to be slow to act on networks precisely because it was built to be fast to see them. Time spent confirming was affordable because detection ran ahead of the payout clock. This is the clearest instance of the thesis in Section 5.4: the reversibility window, not the defender's accuracy, is what determines which due-process architecture is even available. A system deciding inline on a final instrument cannot run a five-gate confirmation ladder before acting, and its protections have to be located elsewhere, in the reversal mechanism itself, in step-up friction, and in remediation after the fact.

A note on precision, and on what kind of claim it is. The wrongful-flag bound was set as a hard shipping constraint, and it remained the constraint the system was held to as detection expanded from item-level to cluster-level enforcement. That is a design commitment, and it is stated here as one. The paper does not report that the bound was observed to hold: establishing that would require the denominator Section 9.2 argues no trust and safety team has, and no such measurement is available here. The numeric bound is withheld under the confidentiality rule stated in Section 1, and the result the withheld figure would support is not claimed. What can be said is why the commitment mattered. Cluster enforcement raises the stakes of a false positive in the way this section opened by describing, so carrying the same artist-level harm bound into campaign-scale action is what made the expansion defensible at all. Whether it was met is exactly the claim this paper declines to make.

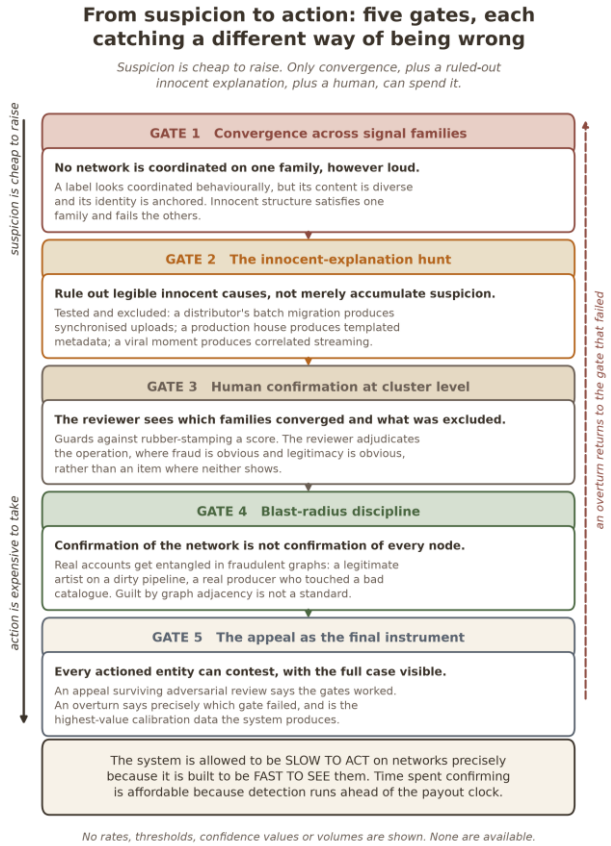


Figure 5. The confirmation ladder from suspicion to action. Each gate is designed to catch a different way of being wrong, and each names the innocent explanations that must be tested and excluded before the case advances. Suspicion is cheap to raise at the top of the ladder and expensive to spend at the bottom; the appeal at the end feeds calibration evidence back to the gate that failed. The figure renders the process described in Sections 7.1 to 7.6 and contains no rates, thresholds, or volumes.

8. Case Analysis

Disclosure. What follows is a composite of campaign patterns encountered in operation, with identifying details altered. The structure is faithful; the specifics are deliberately not. The magnitudes given are approximate and illustrative rather than counted, no dates are given, no volumes are attached, and the resolution narrated at the end of the section describes the shape in which a case of this class comes together rather than a measured outcome. It is presented as an illustration of a campaign class, not as a report on a single incident, and nothing quantitative should be inferred from it.

The campaigns that taught the most were not the loud ones. Burst operations, hundreds of tracks uploaded in a weekend with metadata stamped from an obvious template, are almost a gift. They are synchronised, they are templated, and cluster review resolves them in one sitting. The hard case is the opposite animal: the patient, distributed operation.

Picture a catalogue of AI-generated tracks spread across dozens of small artist accounts. Each created at a different time. Each with a plausible name, modest artwork, a handful of releases. Uploads staggered over

months, through more than one distribution pathway. Streaming activity kept deliberately unremarkable: no spikes, no threshold-hugging, volumes that look like the ordinary long tail. Every account, inspected alone, is indistinguishable from the vast population of real small artists the platform exists to protect.

Three properties made it hard to see, and they are worth separating because they fail different parts of a detection system.

It was engineered against the unit of inspection. Per-item and per-account review is exactly what the operation was shaped to survive. Every node was boring. The evidence did not live in any node; it lived in the edges, and the edges were the thing being actively hidden. This is the failure mode of Section 3.4 in its purest form.

It weaponised the defender's own false-positive discipline. Because the precision gate was strict by design, in order to protect real artists, accounts that looked like plausible small artists inherited the benefit of the doubt. The operation had, in effect, studied what legitimacy looks like at the long tail and manufactured it. The platform's most important protection was the operation's best camouflage. This is camouflage in the sense that graph-fraud work has formalised [8], but aimed at a policy rather than at a metric: the thing being exploited is not the scoring function, it is the deliberate conservatism sitting on top of it.

It starved every short-window signal. Synchronisation detection, burst detection, and velocity signals all depend on things happening close together in time. Patience defeated all of them simultaneously. In any given week, the campaign contributed nothing anomalous to the data. A detector with a short memory was not merely less accurate against this operation; it was structurally blind to it.

What eventually connected it was the one dimension patience cannot launder: history, read across signal families at once. Longitudinal structure showed catalogues growing in ways that were individually plausible but jointly improbable, with variation that was messy per account and suspiciously systematic across accounts, and consistency in exactly the places where genuine populations are noisy. Convergence across independent families, content signals agreeing with structural signals agreeing with distribution-pathway patterns, accumulated until the probability that these accounts were independent collapsed. And when the first subset was actioned, the reaction sealed it: behaviour changes rippled through accounts that had no visible relationship to the ones enforced. The response to enforcement drew the graph.

The lesson carried forward is a reframing. The burst campaign is a detection problem. The patient campaign is an accounting problem. It can hide in any window, but

it cannot hide in its own accumulated record. Which is why retention of behavioural history is not storage overhead in a trust system. It is the substrate the hardest detections are made from.

9. Discussion: What the Field Is Missing

Five gaps, roughly in order of how much they cost.

9.1 The field still ships item classifiers for a graph problem

Most detection investment, academic and industrial, goes into deciding whether this thing is fake: this track, this account, this transaction. But coordination is a property of edges, and the item framing systematically produces the failure mode described in Section 3.4: every node survives inspection.

The benchmarks make it worse. Datasets are labelled at the item level, so the literature optimises for the unit the adversary has already engineered around. Campaign-level benchmarks are what would move the field, and almost none exist. The reasons are structural rather than intellectual: platforms cannot share graph data, and nobody wants to publish their edges. The coordination-detection literature has made real progress precisely by moving to entity-similarity networks and away from item labels [1], [2], [3], and the absence of shared campaign-level evaluation data is the main thing preventing that progress from compounding.

9.2 Detection and economics are still separate conversations

The field measures success as detection rates. Operations measure success as margin. Those can both be winning simultaneously: a detector can resolve most of a campaign while the remainder still clears the operation's cost of production.

The correct objective function is making the operation uneconomic, which means the interesting metrics are the adversary's costs: what it takes to appear independent, to accumulate identity, to survive one enforcement cycle. Almost nobody instruments the adversary's cost curve. The evidence that this is the right frame is behavioural rather than statistical: the campaigns that stop are not the ones you detect best, they are the ones you make unprofitable.

There is a measurement corollary that deserves stating plainly, because it is where this field's reported numbers are least trustworthy. "Detection rate" implies a known denominator of fraud, and no honest trust and safety team has one. You cannot count the campaigns you did not see. The problem is the same one documented in unsupervised anomaly-detection evaluation, where the absence of a reliable ground-truth population makes conventional performance measures unstable and comparisons across settings close to meaningless [16]. It also has a formal name that trust and safety reporting almost never uses: this is a positive-unlabelled setting, in which only a biased subset of the positives is ever labelled and the size of the unlabelled positive mass is itself the quantity in question, and the prevalence-estimation literature treats that quantity as something to be estimated under stated assumptions rather than divided by [17]. A field that reported prevalence estimates with their assumptions attached would be making a smaller claim and a defensible one.

What can honestly be measured is coverage, the yield of cluster-level methods relative to item-level review on the same corpus, and late-surfacing fraud fed back as labelled data rather than counted as a rate against an unknowable base. Harm prevented is worse still: deterrence, operations that shrank, moved elsewhere, or never launched because the system existed, is likely the largest effect and cannot be honestly quantified. The proxies that would show the economic choke point working are that enforced operations do not reconstitute at prior scale and that evasion attempts trend toward smaller and costlier structures. Those are the right quantities to instrument, and this paper does not report them. By its own argument it could not report them credibly without the denominator it has just said nobody has.

9.3 Time horizons are mismatched to the patient adversary

Detection windows are short: sessions, days, velocity features. That is what is computationally comfortable and what streaming architectures encourage. But the hardest campaigns are built precisely to be innocent in every short window and guilty only in their accumulated history, as Section 8 describes.

Longitudinal behavioural retention is not storage overhead; it is where the hardest detections come from,

and the field mostly treats history as a cost centre to be aged out. Synchrony-based methods inherit this limitation structurally, since they compare activity inside a window and a sufficiently patient operation never puts two of its actions in the same one [6], [7].

9.4 Cross-platform coordination is detected by nobody because it is visible to nobody

Operations do not respect platform boundaries. The same actors monetise across streaming services, social platforms, and payment rails, and every platform sees only its slice of the graph. Recent work documenting coordinated operations that span platforms confirms the shape of the problem from the outside [20], but an external researcher assembling a cross-platform picture from public data is not a substitute for the platforms themselves being able to correlate evidence.

The claim to make here has to be stated narrowly, because the broad version of it is false. Cross-platform sharing infrastructure does exist for some harm classes: hash-sharing consortia operate for terrorist content, through the Global Internet Forum to Counter Terrorism, and for child sexual abuse material, which are working counterexamples to the idea that platforms can never stand up a trusted intermediary. Obligations are arriving from the regulatory side too, with the European Union's Digital Services Act creating a vetted-researcher data-access channel and a systemic-risk regime, the first lawful route by which anyone outside a platform can examine this class of problem at all.

What none of it reaches is coordination evidence, or the monetary-abuse case. A hash identifies a known artefact, and an artefact can be shared because it is the thing being suppressed. Coordination is a claim about relationships between accounts, and no platform can hand that over without exporting behavioural histories of real users, most of whom are not implicated. So the gap is specific rather than general: the consortium model works where the evidence is an object and fails where the evidence is a relationship, and there is nothing that plays for coordination the role suspicious-activity reporting and consortium data play in financial crime. The result is that every platform independently rediscovers the same operations, at full cost, with partial evidence, and the binding constraint is institutional rather than technical.

9.5 The field has no serious jurisprudence of coordination

Fraud concepts were inherited from a world of individual actors: this account, this transaction, guilty or not. Campaign-scale enforcement raises questions the field has not formalised. What evidentiary standard confirms a network? What does due process owe a node connected only by graph adjacency? When is correlation an indictment and when is it just a scene?

Practitioners answer these ad hoc, platform by platform, with gates each team improvises. Section 7 is one such improvisation, offered because the alternative is that nothing is written down at all. The governance literature has identified the general problem, that automated enforcement at scale creates obligations of contestability and proportionality that technical systems are not designed to discharge [18], and empirical work on appeals suggests the remedies platforms do offer often do not function as remedies [19]. But there is no equivalent, for coordination specifically, of an evidentiary standard.

The gap is not detection capability. It is that platforms are operating something like a legal system at scale without having written down its rules of evidence.

Gap	What is missing	What would close it
Item framing for a graph problem	Campaign-level labels and benchmarks	Shared campaign-level evaluation data, which means solving why it does not exist: platforms cannot share graph data and will not publish their edges
Detection versus economics	Instrumentation of the adversary's cost curve	Objective functions defined as making the operation uneconomic, and reporting that treats an unknown fraud denominator as unknown
Time horizon mismatch	Long-horizon behavioural retention treated as detection substrate rather than as storage cost	Longitudinal signal design, and retention policy set by evidentiary need rather than by storage economics
Cross-platform blindness	Lawful sharing of coordination evidence: a schema for relationships rather than for artefacts, and cover for the monetary-abuse case	Extension of the consortium and researcher-access

models that work for artefact-defined harms to evidence that is relational

No jurisprudence of coordination An evidentiary standard for confirming a network, and a doctrine for what is owed to a node connected only by adjacency Written, contestable rules of evidence for campaign-scale enforcement, developed jointly by practitioners, researchers, and regulators

Gap	What is missing	What would close it
Item framing for a graph problem	Campaign-level labels and benchmarks	Shared campaign-level evaluation data, which means solving why it does not exist: platforms cannot share graph data and will not publish their edges
Detection versus economics	Instrumentation of the adversary's cost curve	Objective functions defined as making the operation uneconomic, and reporting that treats an unknown fraud denominator as unknown
Time horizon mismatch	Long-horizon behavioural retention treated as detection substrate rather than as storage cost	Longitudinal signal design, and retention policy set by evidentiary need rather than by storage economics
Cross-platform blindness	Lawful sharing of coordination evidence: a schema for relationships rather than for artefacts, and cover for the monetary-abuse case	Extension of the consortium and researcher-access models that work for artefact-defined harms to evidence that is relational
No jurisprudence of coordination	An evidentiary standard for confirming a network, and a doctrine for what is owed to a node connected only by adjacency	Written, contestable rules of evidence for campaign-scale enforcement, developed jointly by practitioners, researchers, and regulators

Table 3. The five gaps of Section 9 and what closing each would require. The entries are the author's analytical assessment, not measured findings.

10. Limitations

This paper is a cross-domain practitioner analysis, and its limitations should be read as constraints on what kind of claim it can support rather than as caveats on results, since it reports no results. It is retrospective, it is written from a single operator vantage, no controlled evaluation was performed, and no operational metrics are released. Each of those is unpacked below, and the section closes with what evidence would overturn the paper's central claims.

It is practitioner experience across two domains, not a study. The media-side claims come from building and operating a coordinated-fraud detection system. The financial technology claims come from adjacent product engineering leadership rather than from operating fraud and risk systems, as stated in Section 5. The cross-domain comparison is therefore asymmetric in authority by construction, and readers should weight it accordingly.

No operational metrics are released, and none are estimated. The author holds no releasable underlying numbers for this work, as set out in Section 1, and nothing beyond the catalogue and daily-ingestion scale stated in Section 4.4 is published. Consequently this paper reports no detection rates, no false-positive rates, no latencies, no yields, and no harm-prevented figures. Where a number would ordinarily appear, the paper states the structural claim and says that the figure is absent. Three claims are affected directly: the wrongful-flag bound in Section 7.6 is stated as a design commitment and explicitly not as a measured outcome, the review-leverage claim in Section 4.5 is stated as a structural property without a before-and-after ratio, and the adaptation criterion in Section 6.1 is stated as a criterion the portfolio was designed against rather than as an effect that was observed. None has been substituted with an estimate. The reasoning is set out in Section 9.2.

No controlled evaluation was performed. There is no held-out benchmark, no ablation across signal families, and no comparison against a baseline detector under matched conditions. The claims about which signals are durable and which rotate cheaply, summarised in Table 2, are analytical judgements grounded in operating experience, not measured quantities. The absence of campaign-level benchmarks noted in Section 9.1 is exactly what prevents them from being tested here.

The case in Section 8 is a composite. It is assembled from patterns encountered in operation, with identifying details altered. Magnitudes in it are approximate and illustrative, and the resolution it narrates is the shape of how such a case comes together rather than a measured outcome. It should be read as a description of a campaign class, not as an incident report, and nothing quantitative should be inferred from it.

Altitude discipline limits reproducibility. The withholding of features, thresholds, and weights described in Section 1 is a real cost to the reader, knowingly accepted. It means the architecture is transferable as a design but is not reproducible as an implementation from this paper alone.

The vantage point is single-platform and single-domain-of-abuse. The media observations come from one platform and one abuse economy, royalty extraction. Coordination on platforms whose payout function is attention rather than money may behave differently in the economic-directionality signature specifically, which is the signature this paper leans on most heavily. The coordination literature cited in Section 2 works largely in that attention-driven setting, and the extent to which the two regimes share signal structure is an open question rather than something established here.

What would falsify the central claims. Since none is tested here, each is stated with its own defeater. The rotation-cost ordering in Table 2 falls to evidence that operations rotate economically anchored behaviour as readily as they rotate metadata, or that behavioural signals depreciate no more slowly than metadata signals against the same adversary. The orthogonality claim in Section 6.2 falls to evidence that one adaptation, a single change of generation tooling for instance, degrades several families at once rather than shifting visibility from one family to another; correlated failure would reduce the architecture to an ordinary ensemble and remove its central guarantee. The reversibility-window thesis in Section 5.4 falls to a defender at the short end of the window who nonetheless runs a deep multi-gate confirmation process before acting, or to a long-window defender whose confirmation depth turns out to be limited by something other than the window. And the cluster-as-unit claim in Section 3.4 falls to an item-level detector that resolves patient distributed campaigns at parity with a cluster-level one on the same corpus, which is precisely the comparison the missing campaign-level benchmarks prevent anyone from running.

11. Conclusion

Coordinated inauthentic behaviour is not a harder instance of the fake-item problem. It is a different problem with a different unit, and a system that decides at the item level will be individually correct on every node of a campaign and still miss the campaign, silently, because nothing in its own outputs records the miss. Everything else in this paper follows from that. Establishing coordination is an inferential problem about the improbability of observed agreement rather than a classification problem about content, which is why the signal portfolio should be selected for orthogonal failure modes rather than for individual strength, and why the durable signals are the ones anchored in costs the adversary cannot shed without abandoning the operation's profitability. Because campaign-scale action

can sweep up real users, confirmation has to be gated rather than scored. And because the two domains compared here differ in the reversibility window rather than in the geometry of the abuse, the depth of confirmation a defender can afford is set by that window; the financial technology side of the comparison is informed observation, and the thesis is offered to be tested rather than as settled.

The argument this paper most wants to leave behind is the one it opened with. The field keeps treating coordination as a bigger classification problem, when it is actually an economics problem, an accounting problem, and a governance problem wearing a classifier's clothes. Economics, because the objective is the adversary's margin rather than the detector's accuracy, and almost nobody instruments the adversary's cost curve. Accounting, because the patient campaign is innocent in every short window and guilty only in its accumulated record, which makes behavioural history a detection asset rather than a storage cost to be aged out. Governance, because platforms are running campaign-scale enforcement against real people without a written standard for what confirms a network or what is owed to a node connected only by association.

The platforms that get the third of those wrong will not fail on recall. They will fail on legitimacy. That is a harder failure to detect from inside a detection system, and it is the one this paper would most like the field to start measuring itself against.

References

1. D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 455-466, 2021. doi: 10.1609/icwsm.v15i1.18075

2. L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated Behavior on Social Media in 2019 UK General Election," Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 443-454, 2021. doi: 10.1609/icwsm.v15i1.18074

3. S. Tardelli, L. Nizzoli, M. Avvenuti, S. Cresci, and M. Tesconi, "Multifaceted online coordinated behavior in the 2020 US presidential election," EPJ Data Science, vol. 13, art. 27, 2024. doi: 10.1140/epjds/s13688-024-00467-0

4. E. Drott, "Counterfeiting Attention in the Streaming Economy: Spam, Click Fraud, and Fake Artists," in Streaming Music, Streaming Capital, Durham, NC: Duke University Press, 2023, pp. 144-192. doi: 10.1215/9781478027874-005

5. S. Cresci, "A decade of social bot detection," Communications of the ACM, vol. 63, no. 10, pp. 72-83, 2020. doi: 10.1145/3409116

6. A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "CopyCatch: stopping group attacks by spotting lockstep behavior in social networks," in Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 119-130. doi: 10.1145/2488388.2488400

7. Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering Large Groups of Active Malicious Accounts in Online Social Networks," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 477-488. doi: 10.1145/2660267.2660269

8. B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "FRAUDAR: Bounding Graph Fraud in the Face of Camouflage," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 895-904. doi: 10.1145/2939672.2939747

9. C. Xiao, D. M. Freeman, and T. Hwa, "Detecting Clusters of Fake Accounts in Online Social Networks," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, 2015, pp. 91-101. doi: 10.1145/2808769.2808779

10. S. Ghosh, R. Anand, T. Bhowmik, and S. Chandrashekhar, "GoSage: Heterogeneous Graph Neural Network Using Hierarchical Attention for Collusion Fraud Detection," in Proceedings of the 4th ACM International Conference on AI in Finance, 2023, pp. 185-192. doi: 10.1145/3604237.3626856

11. M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and

J. Lorentzen, "Detecting money laundering transactions with machine learning," Journal of Money Laundering Control, vol. 23, no. 1, pp. 173-186, 2020. doi: 10.1108/JMLC-07-2019-0055

12. N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 99-108. doi: 10.1145/1014052.1014066

13. B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317-331, 2018. doi: 10.1016/j.patcog.2018.07.023

14. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, art. 44, 2014. doi: 10.1145/2523813

15. I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210, 1969. doi: 10.1080/01621459.1969.10501049

16. G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," Data Mining and Knowledge Discovery, vol. 30, no. 4, pp. 891-927, 2016. doi: 10.1007/s10618-015-0444-8

17. J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," Machine Learning, vol. 109, no. 4, pp. 719-760, 2020. doi: 10.1007/s10994-020-05877-5

18. R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," Big Data & Society, vol. 7, no. 1, art. 2053951719897945, 2020. doi: 10.1177/2053951719897945

19. C. Are, "'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation," Information, Communication & Society, vol. 28, no. 11, pp. 1997-2014, 2025. doi: 10.1080/1369118X.2024.2396621

20. M. Minici, F. Cinus, L. Luceri, and E. Ferrara, "Uncovering coordinated cross-platform information operations: Threatening the integrity of the 2024 U.S. presidential election," First Monday, vol. 29, no. 11, 2024. doi: 10.5210/fm.v29i11.13831