Autonomous Mule Network Detection and Real-Time Asset Restitution in Instant Payment Ecosystems Via Behavioral Graph Analytics and Conditional Finality Protocols

Pratik Pol[1], Dr. Jyoti R. Jadhav[2]
[1,2]*School of Information Technology, Indira University*

***Abstract*—Instant payment rails settle transfers in seconds, which also lets organized fraud rings push stolen money through mule accounts before victims, banks, or investigators can react. Present anti-fraud tools mostly detect and alert; after a fraudulent transfer settles, recovery still depends on complaints, interbank correspondence, police involvement, and disputes lasting days. This paper moves the system boundary from detection to restitution. We present the Automated Real-Time Reversal and Restitution (ARRR) protocol, a policy-governed execution layer that fuses relational graph analytics, temporal velocity anomalies, device and session telemetry, and a conditional finality window of about 120 s. A dual-phase engine builds a dynamic financial-crime graph and produces one mule risk score; above a critical threshold, ARRR places a micro-lien on the receiving account, sweeps the funds back, and follows them downstream through partial liens and future-credit garnishment. In synthetic benchmarks against rule-based, supervised-ML, and graph-only baselines, full ARRR reaches 94.0% precision and 93.0% recall and shortens median recovery from over 24 hours to under 180 s. The paper also covers false-positive control, depleted-balance recovery, anti-reverse-scam checks, privacy-preserving data sharing, and sandbox governance.**

***Index Terms*—money mules; Mule-as-a-Service; instant payments; graph neural networks; payment finality; automated reversal; asset restitution; fraud detection; UPI; FedNow; SEPA Instant**

I. INTRODUCTION

Instant payment schemes such as India's Unified Payments Interface (UPI), the U.S. FedNow Service, and SEPA Instant Credit Transfer make account-to-account transfers available around the clock. FedNow, for example, settles individual payments within seconds and gives the recipient immediate access to the money [4]– [6]. That speed is commercially valuable, but it leaves very little time to step in when fraud occurs.

Conventional controls are mainly preventive: transaction limits, merchant rules, device fingerprinting, authentication challenges, and after-the-fact alerts. Once a payment settles it is generally treated as final. FedNow payments are described as final and irrevocable [4], and public UPI guidance notes that completed transfers cannot generally be recalled [9]. Recovery therefore turns into an administrative and legal process instead of an operation on the payment rail.

Mule-as-a-Service (MaaS) operations exploit this gap. A fraudster persuades a victim to authorize a payment through impersonation, remote-access software, fake investment sites, QR-code tricks, or coercive "digital arrest" calls. The money lands in one or more mule accounts, is split and forwarded, withdrawn as cash, or moved into stablecoins and other assets. Each minute the funds stay available to the mule lowers the chance of getting them back.

Our central premise is that recovery probability falls steeply with time. A process that waits for a complaint, an analyst, and interbank messages cannot keep pace with funds that move again within seconds. What is needed is not only faster detection but an autonomous, auditable, and legally bounded restitution pipeline.

The main contributions are:

- Detection-to-Restitution architecture: the system boundary shifts from "detect and alert" to "detect, contain, reverse, and restore" inside the clearing

layer.

- Dual-phase mule detection: graph topology of accounts, devices, identifiers, and money flows is combined with temporal velocity analysis.
- ARRR protocol: micro-liens, single-hop clawback, multi-hop propagation, and future-credit garnishment form one workflow.
- Conditional Finality Window (CFW): a risk-triggered delay of about 120 s, applied only to transactions with strong coercion or remote-access signals.
- Safeguards and evaluation: escrow, anti-reverse-scam checks, depleted-balance recovery, and governance, tested against rule-based, ML, and graph-only baselines.

II. RELATED WORK

Rule-based controls.

Early systems rely on fixed thresholds, velocity counters, geographic limits, and blacklists. They are cheap and transparent but easy to evade: transfers can be kept below limits, accounts and devices rotated, and activity spread across many small payments. A mule account also looks harmless in isolation; its role shows only through relationships such as shared devices, unrelated inbound senders, or fast outbound dispersion.

Supervised machine learning.

Gradient-boosted trees, random forests, logistic regression, and neural classifiers capture nonlinear links among amount, time, account age, merchant type, and device signals, and clearly beat fixed rules. However, tabular models treat each transaction or account independently, so they miss multi-hop propagation and shared infrastructure. They also depend on labels that are incomplete, because many victims never report and many mules are never confirmed.

Graph analytics.

Accounts, devices, phone numbers, e-mail addresses, IPs, merchants, and branches become nodes; transfers and shared logins or contacts become edges. Mule networks leave structural signatures: many unrelated senders paying a few receivers, minimal dwell time, downstream accounts sharing devices or recovery contacts, new accounts joining known fraud clusters, and a few controllers behind seemingly independent accounts. Graph neural networks learn node representations from both own attributes and neighborhood structure, so an account's risk reflects what it is connected to.

Recovery bottlenecks.

Victims notice fraud late; banks must triage complaints manually; funds cross institutions and need interbank or law-enforcement requests; finality rules protect recipients; and reversal usually needs evidence rather than suspicion.

In India the Indian Cyber Crime Coordination Centre (I4C) supports reporting and data sharing among agencies [7], [8]; in the United States FinCEN Section 314(b) gives participating institutions a safe harbor for sharing information on suspected illicit activity [1]–[3]. Both aid investigation, but neither is a sub-minute automated restitution rail.

III. PROPOSED METHODOLOGY

A. System Context and Threat Model

The system operates across payment operators, issuing and receiving banks, wallet providers, consent-bound data providers, financial intelligence and cybercrime bodies, and the people involved (victims, legitimate payees, merchants, mule operators).

The adversary is financially motivated and may use social engineering, remote-access tools, SIM-swap or account takeover, synthetic identities, rented mule accounts, rapid layering, cash-out and crypto off-ramps, and even false victim claims to reverse honest payments.

Because not every suspicious payment is fraud, four states are distinguished: low risk (proceed normally), escalation risk (brief hold and extra verification), critical mule risk (eligible for immediate lien and reversal), and ambiguous risk (funds go to verification escrow).

Fig. 1 shows the complete architecture, from data ingestion through detection and scoring to the ARRR layer and the governance layer.

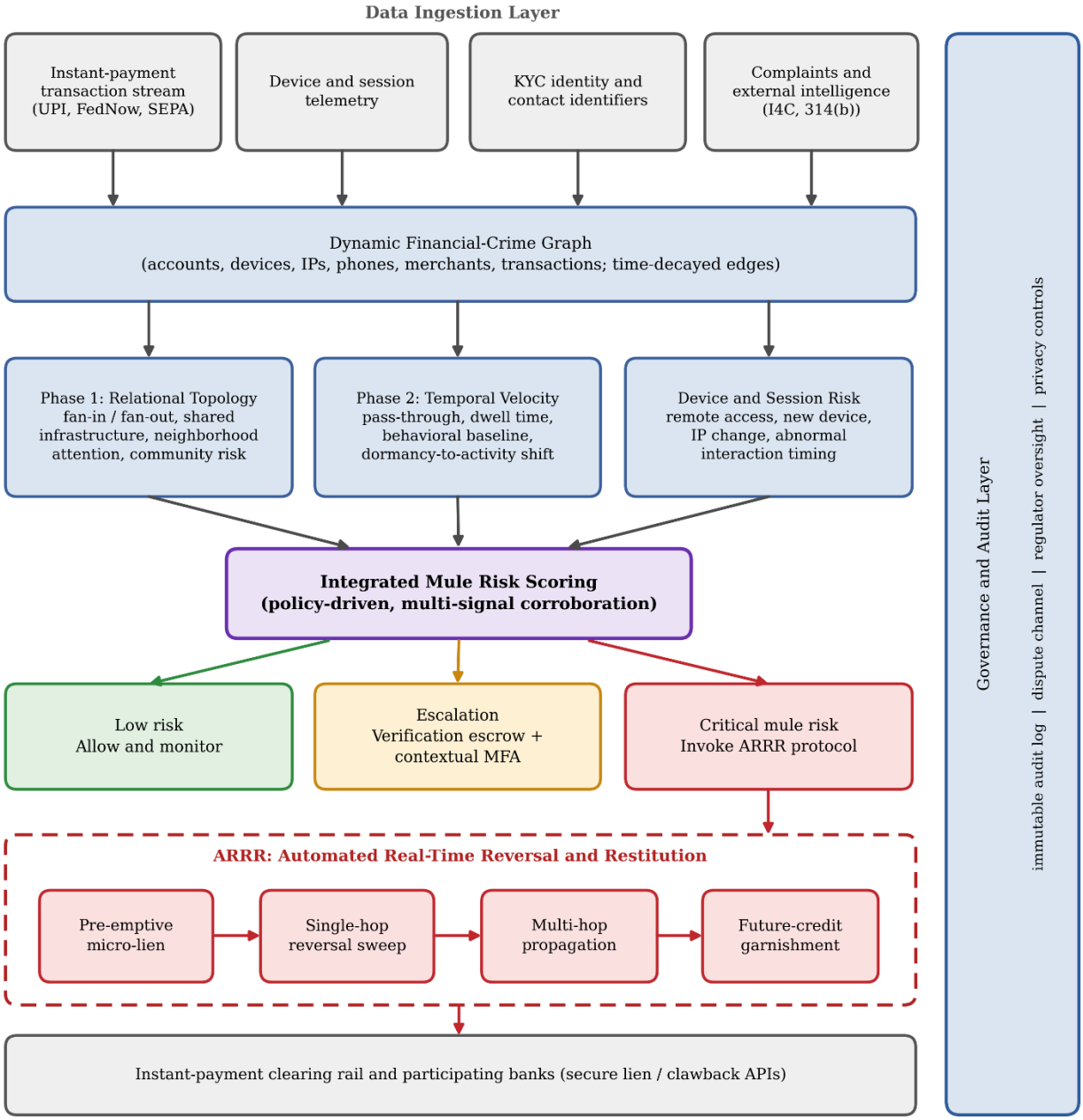

Fig. 1. Overall architecture of the proposed Detection-to-Restitution system.

B. Dynamic Financial-Crime Graph

The core data structure is a time-varying heterogeneous graph. Node types are accounts (bank, wallet, prepaid), persons or KYC identities, devices (handsets, browser fingerprints, emulators), network identifiers (IP, subnet, ASN), contact identifiers (phone, e-mail, UPI ID), merchants or payees (including QR codes), and transactions (payments, refunds, reversals, lien events). A payment is a directed edge carrying time, amount, channel, device context, and risk annotations; shared-device and shared-IP links are undirected; a phone number reused across accounts forms an identity-linkage edge. The graph updates continuously from new transactions, logins, complaints, and external intelligence, and old links lose influence over time so that stale associations do not permanently stigmatize an account.

Fig. 2 illustrates a typical mule network: fan-in from unrelated senders, fast fan-out to downstream mules and cash-out sinks, and shared device, phone, and IP identifiers.

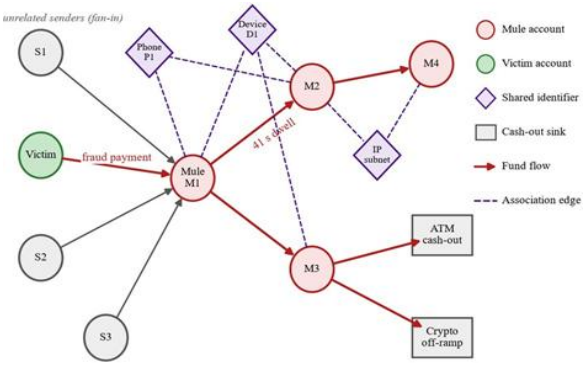

Fig. 2. Example mule network in the dynamic financial-crime graph.

C. Relational Topology Phase

This phase asks whether an account belongs to a mule-like community. It extracts eight structural indicators: fan-in concentration, fan-out dispersion, low dwell time, a dormant-to-active shift, shared infrastructure (device fingerprint, recovery phone, IP subnet, remote-access signature), circular or layered flows, new-account acceleration, and community-level risk from confirmed fraud reports or blocked neighbors.

Risk is not read from an account alone; neighborhood attention weighs the strength, recency, direction, and type of each link. Evidence is tiered. *Weak*: a shared public IP, a common merchant category, or a generic device model. *Moderate*: repeated contact identifiers, similar registration patterns, or overlapping beneficiaries. *Strong*: identical device fingerprints, remote-access tool signatures, confirmed fraud links, or repeated rapid pass-through. For instance, one shared café Wi-Fi address means little, whereas one unique device fingerprint behind several accounts that all receive money from strangers and forward it at once is strong proof of coordination.

D. Temporal and Velocity Anomaly Phase

This phase looks at how money moves. For pass-through detection it tracks the gap between credit and debit, the share of inbound money forwarded quickly, the number of beneficiaries per inbound payment, how closely outflows mirror inflows, the residual balance, and any abrupt change after inactivity. An account that receives money from many unrelated senders and forwards almost all of it within seconds or minutes is a strong pass-through candidate. Against per-account and peer-group baselines it also checks unusual hours, sudden value jumps, new beneficiaries, new geography or device, concurrent distant sessions, changed navigation behavior, screen-sharing during a payment, and abnormal copy-paste or form-filling speed. These session signals matter most for victim-side coercion, where a genuine customer authorizes a payment under remote control, urgency, or fear.

E. Integrated Mule Risk Scoring

Three dimensions are combined: graph structural risk, temporal velocity risk, and device and session risk. The result is not a plain sum. A policy-driven logic lets some high-confidence signals escalate a case directly while weaker ones must corroborate each other. Table I lists the action bands, and Fig. 3 shows how the dimensions map to them. The critical threshold is deliberately conservative: automatic reversal needs independent evidence streams to converge on suspicious structure, abnormal money movement, and victim-side coercion or takeover telemetry.

TABLE I Risk bands and system responses

Band	Typical evidence	System response
Low risk	Normal topology, stable behavior, trusted device	Allow and monitor
Escalation escrow	Mixed signals: new device, unusual velocity, partial graph link	Hold briefly; biometric or contextual MFA
Critical mule	Strong clustering, rapid pass-through, coercion or remote-access evidence	Micro-lien; start ARRR reversal

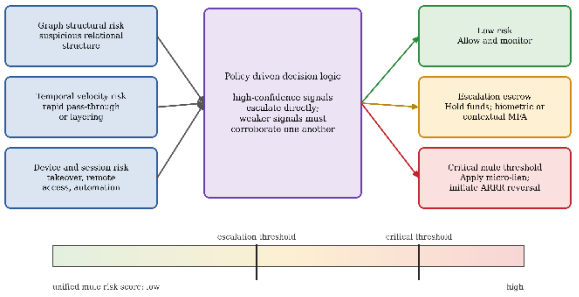

Fig. 3. Integrated mule risk scoring and action bands.

IV. THE ARRR PROTOCOL

A. Design Principles

ARRR rests on four principles: *speed with restraint* (act within seconds, but take irreversible steps only on multi-signal evidence); *containment before clawback* (first stop further dissipation); *proportional recovery*

(recover available funds without freezing unrelated assets); and *auditability and contestability* (every automated action leaves an immutable evidence trail and a dispute path). It is not a blanket power for banks to reverse any payment, but a narrow, risk-conditioned mechanism embedded in the rail and bound by scheme rules, consumer-protection law, and regulation. Fig. 4 gives the full decision and restitution workflow that the following subsections describe.

B. Pre-Emptive Micro-Lien

When a transaction crosses the critical threshold, the receiving institution gets a lien instruction through a secure rail API. The lien attaches only to the identified inbound amount, not the whole balance, so unrelated customer funds stay usable.

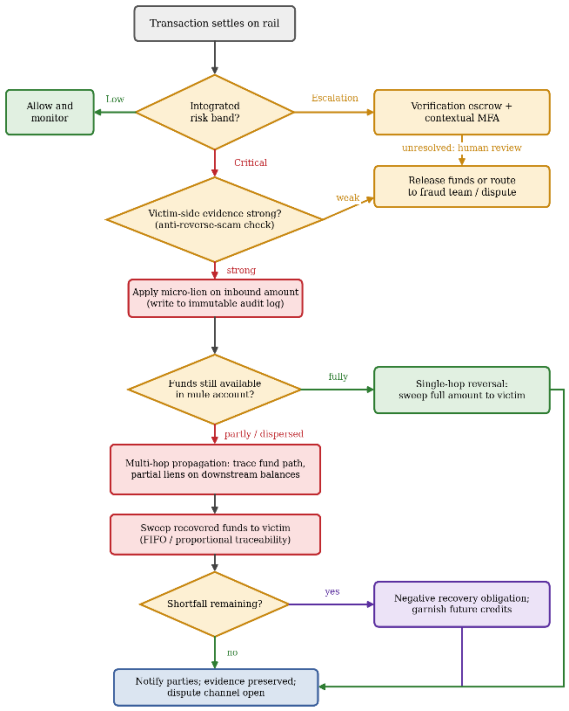

Fig. 4. ARRR decision and restitution workflow.

The steps are: (1) the engine emits a signed event with transaction identifiers and graph, temporal, and device-session evidence; (2) the rail validates it against scheme policy and the institution's authorization scope; (3) the receiving bank earmarks the amount, blocking withdrawal, onward transfer, or crypto purchase; (4) the bank returns an acknowledgment with balance, lien status, and timestamp; and (5) the event and rationale go to the immutable audit log. The micro-lien is the containment primitive that buys time for reversal without an immediate full-account freeze.

C. Single-Hop Autonomous Reversal

In the simplest case the victim pays a mule directly and the funds are still there. ARRR confirms the payment is within the CFW or otherwise eligible, verifies victim-side coercion or takeover evidence, confirms the receiver matches the mule profile, applies the micro-lien, issues an interbank clawback for the earmarked amount, sweeps the available amount to the victim's original account, notifies both banks, the operator, and (where required) the victim, and preserves evidence for disputes and reporting. Fig. 5 shows the message sequence. If only part of the amount remains, the available portion is recovered and depleted-balance handling begins.

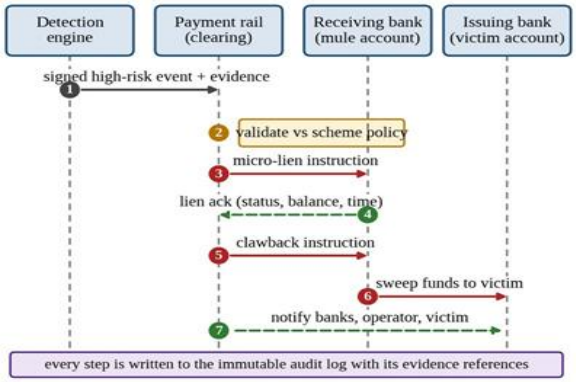

Fig. 5. Message sequence for micro-lien execution and single-hop autonomous reversal.

D. Multi-Hop Dynamic Propagation

If the mule forwards funds before the lien lands, ARRR reconstructs the directed fund-flow path from the victim's original payment, identifies downstream accounts holding derived funds, estimates the remaining traceable value at each hop after fees, partial withdrawals, and commingling, and places partial micro-liens on available balances. Accounts with high recoverable balances, short time since receipt, and strong linkage to the fraud path come first. Recovered value is returned through a controlled restitution settlement, and downstream accounts stay monitored for later credits. Since stolen and legitimate money may be mixed, a traceability policy (first-in-first-out, proportional allocation, or local legal rules) applies; the default is conservative partial recovery, with unresolved amounts sent to dispute and law-enforcement channels.

E. Conditional Finality Window

Finality is essential to confidence in instant payments, so a universal delay would defeat their purpose. ARRR instead defines a Conditional Finality Window: a brief, risk-triggered period, nominally 120 s, in which a settled transaction remains open to a narrowly defined high-risk review. It is activated only by strong signals such as active remote-access or screen-sharing software, abrupt IP or device changes mid-session, extreme urgency or abnormal navigation, a new-device login just before a large transfer, known scam infrastructure in the beneficiary graph, or victim-side signs of impersonation or coercion. During the window the recipient may see the credit, ordinary low-risk payments stay instantly usable, and the flagged payment is provisionally available but subject to a micro-lien if the critical threshold is met. If no critical evidence appears, the lien is released and finality becomes unconditional. Fig. 6 shows the timeline using the decision times measured in Section VI-E.

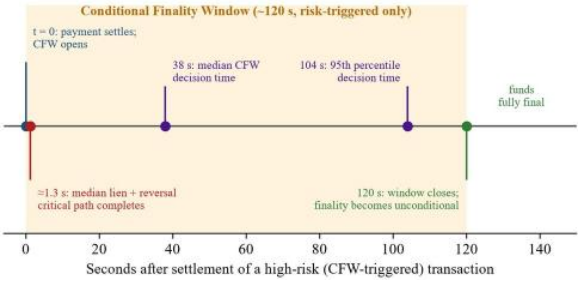

Fig. 6. Conditional Finality Window timeline.

F. Depleted-Balance Handling

If the money has already left, the case is not written off. ARRR computes the unrecovered amount, places a negative-balance lock or restitution obligation on the mule account, blocks closure, nominal ownership transfer, or withdrawal of linked funds where the law allows, garnishes future inbound credits up to the outstanding amount, extends the same logic to linked accounts only when a documented control relationship and legal authority exist, and notifies the holder with a dispute channel. A one-off recovery attempt thus becomes a persistent process, and a flagged account can no longer simply be abandoned after cash-out.

G. Anti-Reverse-Scam Verification

Autonomous reversal creates a new risk: a fraudster may pose as a victim to undo a legitimate payment. ARRR therefore checks the payer side first: remote-access tools in the session, abnormal navigation, abrupt IP, device, or location change, whether the payee matches a known scam or mule cluster, the claimant's history of disputed reversals, consistency with a genuine commercial relationship, and whether the payer can pass contextual multi-factor verification. If victim-side evidence is weak, the system does not reverse automatically; it uses escrow, asks for verification, or routes the case to a human fraud team, which blocks friendly fraud, buyer's remorse, and coordinated reverse-scam attacks.

H. Regulatory and Interbank Integration

A governance layer ties automated execution to legal authority. Section 314(b) can support privacy-preserving risk intelligence in the United States, while reversal itself stays under scheme rules and consumer-protection law [1]–[3]. In India a UPI deployment would need to interoperate with cybercrime complaint channels, bank fraud desks, NPCI-level scheme governance, and law-enforcement escalation [7], [8]. The model has five parts: a scheme-level ARRR rulebook (triggers, evidence standards, action limits, appeal rights); bank-level execution policy; regulatory oversight of reversal and false-positive rates, complaints, and recovery; a law-enforcement interface for evidence preservation and asset tracing; and independent audit of algorithmic decisions, drift, and disparate impact.

V. WORKED EXAMPLE

The following scenario is illustrative and is not drawn from the experiments. *Case A (single hop, CFW).* Asha is told by a caller posing as her bank to install a screen-sharing app. From a new device she sends INR 150,000 over UPI to account M1, opened nine days earlier, which has received funds from 11 unrelated senders in ten minutes and shares a device fingerprint with accounts M2 and M3. Screen-sharing plus a new-device login opens the CFW. The stage times in Table II are the median values of Table VI.

TABLE II Case A timeline (median stage latencies)

Time	Event or action
t = 0	Payment settles; CFW opens (remote access + new device)
+0.06 s	Graph update and graph, temporal, and session scoring finish; all three dimensions are high

+0.24 s	Critical threshold met; micro-lien on INR 150,000 at M1 acknowledged
+0.86 s	Interbank clawback instruction completes
+1.27 s	INR 150,000 restored to Asha; banks and operator notified; audit entry written

Case B (multi-hop). A second victim loses INR 80,000 with no visible coercion signal, so no CFW opens. Six minutes later M1 is flagged critical, but INR 50,000 has already moved to M2 and INR 30,000 remains in M1. ARRR liens INR 30,000 at M1 and INR 40,000 at M2 (INR 10,000 of the forwarded amount was withdrawn as cash), recovering INR 70,000; the missing INR 10,000 becomes a restitution obligation on M2, garnished from its future credits.

Case C (false-positive control). A freelancer receives many small payments from different clients. Fan-in is high, but the device is long-standing and trusted, no link to a fraud cluster exists, and outflows are not immediate. The score lands in the escalation band: a brief hold and contextual MFA, then release, with no lien.

VI. EXPERIMENTAL EVALUATION

A. Dataset and Protocol

Evaluation uses a synthetic but structurally realistic payment corpus that reproduces retail networks, organized fraud campaigns, and MaaS operations without exposing real customer data (Table III). It deliberately contains hard positives (low-value mule transfers, accounts without fraud history, newly onboarded mules, fraud in business hours), hard negatives (freelancers with many small payments, marketplace sellers with high fan-in, diaspora remittance receivers, customers traveling across jurisdictions), and adversarial cases (rotating devices, transfers fragmented below limits, false-victim claims).

TABLE III SYNTHETIC DATASET COMPOSITION

Dimension	Value
Simulated accounts	2,000,000
– legitimate / merchant / mule / confirmed fraud-linked	1,918,000 / 60,000 /18,000 / 4,000
Simulated transactions	48,000,000
Fraudulent transactions (prevalence)	240,000 (0.50%)
Single-hop / multi-hop / reverse-scam cases	132,000 / 84,000 / 24,000
Average / median fraud amount	12,500 currency units
Average mule hops; network size	2.7; 9.4 accounts
Mule dwell time vs. legitimate	41 s vs. more than 24 h

A time-based split is used: training on the first 70% of the timeline, validation on the next 10%, and testing on the final 20%, which prevents leakage of future information. The unit of evaluation is the fraud-to-mule path: a case counts as detected only if the originating flow and the associated mule account or propagation path are both identified. Metrics are precision, recall, F1, false-positive rate, asset-recovery rate, mean and median recovery latency, customer-friction rate, and wrongful-reversal rate.

B. Detection Performance

TABLE IV DETECTION PERFORMANCE COMPARISON

Framework	Full	Part.	None	Mean rec.	Median latency
Rule-based	11.0%	19.0%	70.0%	23.4%	>24 h
XGBoost ML	27.0%	31.0%	42.0%	46.8%	≈4 h
Graph-only	58.0%	27.0%	15.0%	74.1%	≈12 min
Full ARRR	76.0%	17.0%	7.0%	88.6%	<180 s

Relative to the rule-based baseline, full ARRR gains 40 points of precision, 25 points of recall, and 0.333 of F1, and cuts the false-positive rate by 95.9% (22.0% to 0.9%).

Against the XGBoost baseline the gains are 13 and 17 points, +0.151 F1, and an 87.1% lower false-positive rate. Against graph-only detection the classification gain is small (+1 point each, +0.010 F1, false positives down 40%), but median recovery time falls from about 12 minutes to under 3 minutes. Graph analytics therefore improves identification, while ARRR's main advantage is converting identification into timely restitution. Fig. 7 compares the four frameworks.

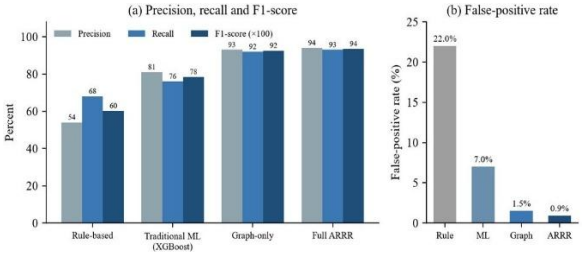

Fig. 7. Detection performance: (a) precision, recall, and F1; (b) false-positive rate.

C. Asset Recovery

TABLE V ASSET RECOVERY EFFICIENCY

Framework	Prec.	Rec.	F1	FPR	Recovery window
Rule-based	54.0%	68.0%	0.602	22.0%	>24 h, manual
XGBoost ML	81.0%	76.0%	0.784	7.0%	≈4 h, manual
Graph-only	93.0%	92.0%	0.925	1.5%	≈12 min, manual
Full ARRR	94.0%	93.0%	0.935	0.9%	<180 s, automated

Full ARRR returns 88.6% of stolen value on average and recovers everything in 76.0% of cases, leaving only 7.0% with no recovery. Three mechanisms drive this: micro-liens block immediate cash-out in 81.0% of critical-risk cases; multi-hop propagation recovers value from at least one downstream account in 64.0% of dispersed-fund cases; and future-credit garnishment adds a further 6.8% of value from depleted mule accounts over the 90-day observation window. Fig. 8 shows the outcome distribution and median latency.

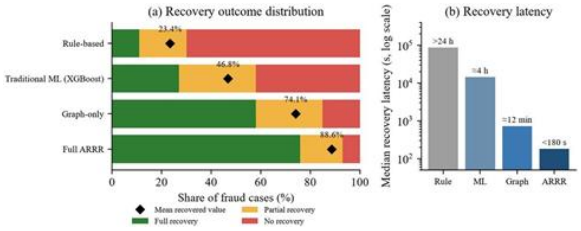

Fig. 8. Asset recovery: (a) recovery outcome distribution; (b) median recovery latency.

D. Latency

TABLE VI PIPELINE LATENCY AT 25,000 TPS (MS)

Stage	P50	P95	P99
Ingestion and graph update	8	19	31
Relational topology scoring	27	61	94
Temporal and velocity analysis	11	26	43
Device and session risk	9	22	38
Integrated risk decision	4	9	15
Micro-lien API execution	180	410	690
Interbank clawback instruction	620	1,240	1,870
Victim restitution settlement	410	930	1,410
End-to-end critical path	1,269	2,717	4,191

Under a simulated peak of 25,000 transactions per second, the median end-to-end recovery takes about 1.27 s, the 95th percentile 2.72 s, and the 99th percentile 4.19 s, all far below the 180 s target. The slowest stages are interbank clawback and restitution settlement, not model inference, which supports deploying ARRR as a native clearing-rail service rather than an external post-settlement analytics layer (Fig. 9).

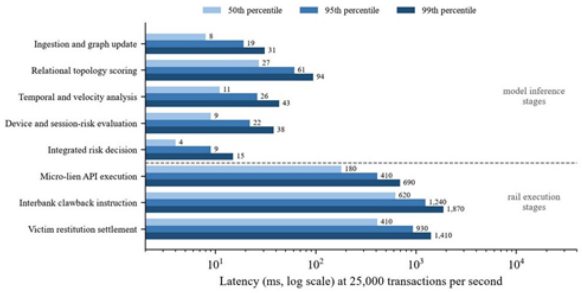

Fig. 9. Stage-wise latency distribution of the ARRR pipeline under peak load.

E. Conditional Finality Window Outcomes

The CFW is triggered for only 0.42% of all transactions. Of these, 58.0% are released without a lien, 24.0% go to escalation escrow, and 18.0% receive a critical micro-lien; 89.0% of lien cases end in successful reversal, and only 0.31% of reversals are later overturned. The median CFW decision takes 38 s and the 95th percentile 104 s, inside the 120 s window. The window thus works mainly as a short verification period rather than a broad settlement delay (Fig. 10).

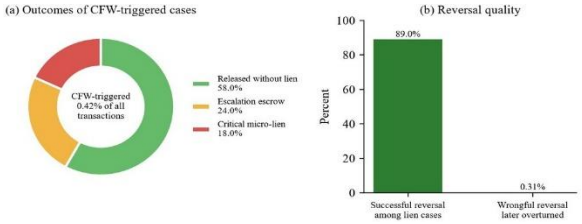

Fig. 10. CFW outcomes: (a) case outcomes; (b) reversal quality.

F. Ablation Study

TABLE VII ABLATION STUDY (ONE COMPONENT REMOVED)

Configuration	F1	FPR	Mean rec.
Full ARRR	0.935	0.9%	88.6%
– graph topology phase	0.814	4.8%	61.2%
– temporal velocity phase	0.865	2.6%	72.9%
– device/session telemetry	0.875	2.1%	76.4%
– multi-hop propagation	0.920	1.1%	69.5%
– micro-lien containment	0.925	1.2%	51.8%
-anti-reverse-scam checks	0.925	2.7%	90.2%

Graph topology is the largest detection contributor (F1 drops from 0.935 to 0.814), followed by temporal velocity (0.865). Micro-lien containment is the largest

restitution contributor: without it mean recovered value falls from 88.6% to 51.8% even though detection stays high. Removing multi-hop propagation lowers recovery to 69.5%. Without anti-reverse-scam checks recall rises slightly (94.0%) but the false-positive rate climbs to 2.7%, because ambiguous cases are no longer diverted to escrow (Fig. 11).

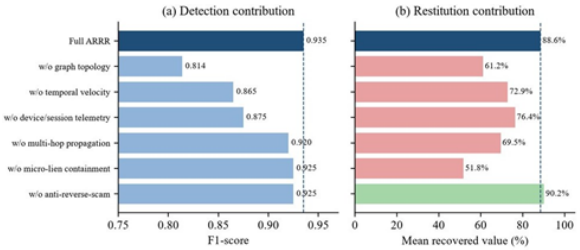

Fig. 11. Ablation study: (a) detection contribution; (b) restitution contribution.

G. Customer Friction

Full ARRR flags 0.9% of legitimate transactions, holds or challenges 0.6%, places a lien on 0.18% of legitimate accounts, and sees 0.31% of reversals overturned. The graph-only system scores 1.5%, 1.1%, 0.4%, and 1.8% on the same measures. False positives are released in under 5 minutes, versus about 35 minutes (graph-only), about 6 hours (ML), and more than 24 hours (rule-based), thanks to automated evidence review, contextual MFA, and scheme-level appeal workflows (Fig. 12).

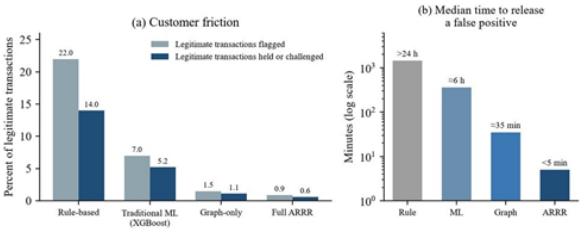

Fig. 12. Customer friction: (a) legitimate transactions flagged or held; (b) median time to release a false positive.

H. Robustness Under Adversarial Variation

Against adaptive mules, ARRR keeps higher recall than graph-only detection in every scenario: 91% vs. 88% for cash-out within 30 s, 90% vs. 86% for fragmentation, 89% vs. 84% for device and IP rotation, 87% vs. 79% for chains of five or more hops, and 90% vs. 81% for false-victim attempts. Its mean recovered value is 74.3%, 79.8%, 81.6%, 68.4%, and 87.1% respectively. Long layering chains hurt most, especially when funds become cash or digital assets before a lien applies; this gap marks the limit of automated restitution and motivates law-enforcement asset tracking and digital-asset intelligence (Fig. 13).

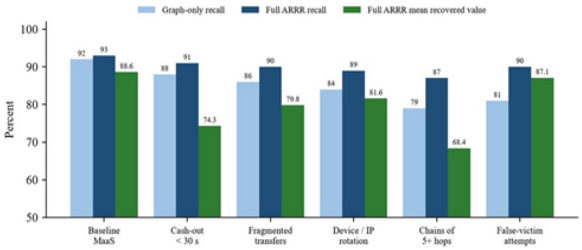

Fig. 13. Robustness of graph-only detection and full ARRR under adversarial variation.

I. Statistical Significance

Bootstrap resampling over 10,000 test iterations shows that all differences are significant at the 95% level. ARRR's F1 advantage is +0.333 (CI +0.318 to +0.347) over rule-based, +0.151 (+0.139 to +0.163) over ML, and +0.010 (+0.006 to +0.014) over graph-only. In mean recovered value the gains are +65.2 points (63.1 to 67.3), +41.8 points (39.7 to 43.9), and +14.5 points (13.2 to 15.8). The key result is that the marginal classification gain is small but the gain in recovered value is large and robust (Fig. 14).

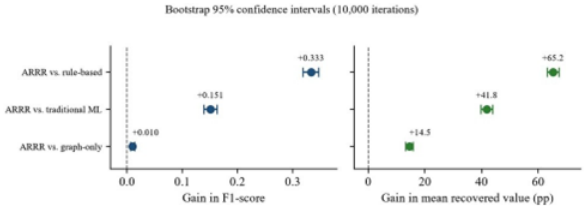

Fig. 14. Gains of ARRR over each baseline with bootstrap 95% confidence intervals.

J. Threat-Model Coverage

Detection recall and recovery are strongest where victim-side session evidence exists before dissipation: account takeover with remote access (95.0% recall, 91.0% recovery) and direct single-mule transfers (96.0%, 93.0%). Coercion-based authorized push payments reach 92.0% and 87.0%. Recovery weakens for three-to-five-hop chains (76.0%), immediate ATM cash-out (68.0%, since physical cash cannot be clawed back automatically), and crypto off-ramps (87.0% recall, 54.0% recovery, which needs exchange and virtual-asset-service-provider integration). False-victim reversal scams are detected at 90.0%, with anti-reverse-scam verification as the primary control (Fig. 15).

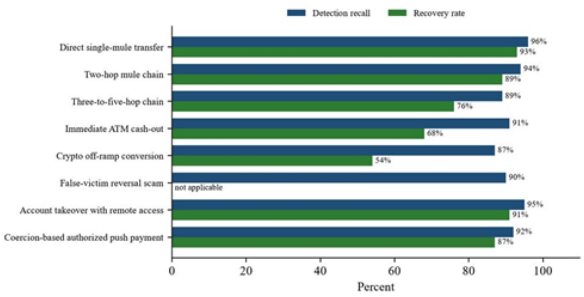

Fig. 15. Detection recall and recovery rate by fraud scenario.

VII. POLICY, GOVERNANCE, AND ETHICS

Payment finality.

ARRR does not abolish finality; it proposes a conditional exception that is time-bounded (about 120 s), risk-triggered, evidence-based, appealable, auditable, and proportionate. Consumer protection. Customers need prompt notice of any lien, hold, or reversal, a plain-language explanation, a human review channel with time-bound resolution, compensation for demonstrable error, protection against repeated wrongful flags, and accommodation for vulnerable users. A person may be a victim in one transaction and an unwitting mule in another, so one automated decision must not criminalize them permanently.

Privacy and fairness.

Cross-institution graphs should use risk tokens instead of raw identity documents, hashed identifiers where lawful, purpose-limited sharing, retention limits, encryption, access logging, differential privacy or aggregation where feasible, and independent privacy impact assessment; Section 314(b) likewise demands confidentiality and permitted-purpose use [1]–[3]. Fairness needs regular testing across demographic and geographic segments, monitoring of false-positive burden by cohort, human oversight of irreversible actions, model versioning and rollback, explainable evidence summaries, and independent red-teaming.

Sandbox adoption.

Before production, ARRR should be tested in a central-bank or scheme sandbox with a payment operator, several issuing and receiving banks, fraud-analytics providers, consumer-protection and data-protection representatives, and law-enforcement observers, measuring detection quality, false-positive burden, recovery, complaints, resilience, API reliability, audit readiness, and effect on payment speed. Rollout is phased (Fig. 16): escrow-only interventions, then micro-liens, then automated single-hop reversal, and finally multi-hop restitution after reliability is shown.

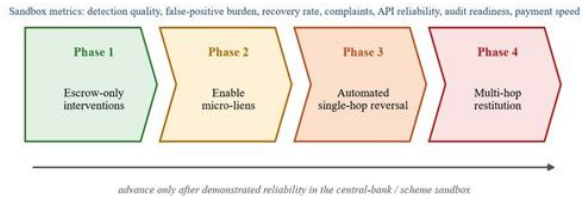

Fig. 16. Phased rollout of ARRR in a central-bank or scheme sandbox.

VIII. CONCLUSION AND FUTURE WORK

This paper recasts money-mule mitigation as autonomous, sub-minute restitution instead of delayed manual recovery. ARRR combines graph analytics, temporal velocity detection, device and session telemetry, and a conditional finality mechanism inside the payment infrastructure. In simulation, it delivers 94.0% precision and 93.0% recall while shrinking recovery time from more than 24 hours to under 180 s; the most important gain is not classification accuracy but the value preserved by immediate containment. By tying evidence-gated micro-liens and reversal sweeps to the first seconds after settlement, ARRR makes stolen funds traceable, recoverable, and attached to a lasting restitution obligation.

Future work includes cross-border interoperability (correspondent banking, remittance corridors, multi-currency settlement), digital-asset tracing across exchanges, stablecoin issuers, and DeFi off-ramps, federated graph learning so banks can improve detection without exposing raw data, robustness against mule networks built to evade graph detection, human-AI hybrid governance defining when automated restitution must defer to people, victim-centered design that protects coercion victims, elderly users, and digitally excluded groups, and formal evaluation of how conditional finality affects liquidity, merchant acceptance, trust, and systemic risk.

ACKNOWLEDGMENT

The author thanks her teacher guide for the guidance that made this research possible, and the Indira College

of Commerce and Science for allowing her to pursue this work.

REFERENCES

[1] Financial Crimes Enforcement Network, "Information Sharing Under Section 314(b)." U.S. Department of the Treasury. [Online]. Available: FinCEN Section 314(b) Information Sharing
[2] Financial Crimes Enforcement Network, "Section 314(b) Fact Sheet," Jun. 12, 2026. U.S. Department of the Treasury. [Online]. Available: FinCEN Section 314(b) Fact Sheet
[3] Financial Crimes Enforcement Network, "FinCEN Director Emphasizes Importance of Information Sharing Among Financial Institutions," Dec. 10, 2020. U.S. Department of the Treasury. [Online]. Available: FinCEN Information Sharing News Release
[4] Board of Governors of the Federal Reserve System, "FedNow® Service FAQs." [Online]. Available: Federal Reserve FedNow Additional Questions and Answers
[5] Board of Governors of the Federal Reserve System, "Additional Questions and Answers," *FedNow® Service*. [Online]. Available: Federal Reserve FedNow Additional Questions and Answers
[6] Federal Reserve Financial Services, "About the FedNow® Service." [Online]. Available: About the FedNow Service
[7] Ministry of Home Affairs, Government of India, "Indian Cyber Crime Coordination Centre (I4C)." [Online]. Available: Ministry of Home Affairs—Cyber Information Security Division
[8] Press Information Bureau, Government of India, "National Cybercrime Response Mechanism," Jul. 22, 2026. [Online]. Available: PIB—National Cybercrime Response Mechanism
[9] Reserve Bank of India, "Harmonisation of Turn Around Time (TAT) and Customer Compensation for Failed Transactions Using Authorised Payment Systems," Circular RBI/2019-20/67, Sep. 20, 2019. [Online]. Available: RBI—Failed Transaction Reversal and Compensation Framework