

Plus Factors on the Ledger: Testing Evidence of Coordinated Demand in Live AI-Agent Payment Markets

Marley, ChainWard (chainward.ai)

Apart AI Collusion Research Sprint, October 23 to 25, 2026. Track 2: Detection and Audit of Collusion (research question 2.3)

Abstract

What should count as convincing evidence that AI agents collude? We test candidates in two live AI-agent markets with public payments: x402 on Base and TermiX's agent-hiring escrow on BNB Chain. Labels for 123 cases (a payee in a 7-day window; 20 families) come from ChainWard investigations published before the sprint. We chose a rule over eight pre-registered features on 10 development families, froze it in a commit and tested it on 10 held-out families. The rule flags a payee when at least 42% of its payment value comes from payers its own money reached within four transfers, exchanges excluded. On the test half it caught 38 of 42 coordinated cases and flagged 1 of 6 others (family-weighted balanced accuracy 0.82, 95% interval 0.56 to 1.00, permutation $p = 0.024$). Evidence antitrust would call parallel conduct separated poorly: AUC 0.63 for a shared funder, 0.47 for concentration, 0.70 for timing. The labels rest partly on the same funding evidence; a rule on features they did not use scored 0.67 ($p = 0.14$). After the freeze, the unchanged rule flagged 34% of 750 x402 payee-weeks above \$10 (30% of the volume) and 14% of 9,131 Set and Earn payees. Of 25 flagged payees no decode had examined, two blinded judges called 22 and 20 coordinated; of 15 unflagged, 4. End to end, synthetic rings in the raw data were recovered exactly. Here, money returning to the seller behaved as a plus factor; shared funders, concentration and timing did not.

1 Introduction

AI agents now pay each other. With x402, a server answers an HTTP request with a price and the client pays by signing a USDC transfer that a facilitator settles on-chain (Coinbase, n.d.; Kim et al., 2020). On TermiX, agents registered in the ERC-8004 identity registry hire other agents through an escrow that pays the seller minus a 2% fee (De Rossi et al., 2025; [D7]). These markets reward measured demand: leaderboards rank x402 sellers by the USDC they receive [D4], TermiX awards points for jobs [D7], and BNB Chain's Set and Earn campaign rewards participants whose agents other wallets hire [D5]. That invites a seller to fund its own buyers, or a group of wallets to hire each other. ChainWard's investigations ("decodes") published before the sprint found both: a seller whose buyers' first USDC came from the seller, six wallets that paid only each other, and 48 wallets that hired only within their own group [D1]–[D6]. Fabricated volume bought rankings on crypto exchanges in the same way (Cong et al., 2023).

Research question 2.3 of this sprint asks what should count as convincing evidence of collusion. Antitrust law met the problem long ago: similar behaviour by competitors (parallel conduct) is lawful on its own (Bell Atlantic v. Twombly, 2007), and inferring an agreement takes additional evidence, the plus factors (Kovacic et al., 2011; Berger et al., 2025). Most work on collusion between AI agents studies simulated games in which the experimenter knows who colludes (Calvano et al., 2020; Fish et al., 2024; Lin et al., 2024; Agrawal et al., 2025; Nakamura et al., 2026). We take the opposite trade: weaker ground truth (published investigations) for real markets with a complete public record of actions. A payment cannot be reworded after the fact.

We ask which ledger evidence separates coordinated from independent demand, and whether a rule chosen on half of the cases holds on the other half and on the rest of the market. Our contributions are **a labelled case set from live agent markets** (123 cases in 20 families, each tied to a published quote and to the evidence the investigation used, so that circularity can be measured); **a pre-registered test** (split and hypotheses committed before any feature was computed, the rule frozen in a commit before the held-out half was scored, with bootstrap intervals and an exact permutation test); **controls** (synthetic rings injected into real markets, including end to end through the raw transfers, shuffled payment times, sensitivity to the free settings); **an out-of-sample run, added after the freeze** (every x402 payee on Base above \$10 a week and every Set and Earn payee, and a blinded spot-check of flagged and unflagged payees no decode had examined); and **an answer to 2.3 in antitrust terms**:

money that returns to the seller and buyers without outside money act as plus factors, while a shared funder, concentrated demand and timing behave like parallel conduct.

Terms. A payee’s demand is *coordinated* when the money paying it comes from the payee itself or circulates in a closed group. A ledger cannot tell one operator with many wallets from colluding agents: both leave the same money paths, and a ranking or incentive program is misled the same way, so we do not separate them. The units are wallets, not people or models: both markets are built for AI agents, but the chain does not show which wallets an LLM agent runs.

Preexisting work and AI assistance. The case labels come from six ChainWard decodes published Sep 27 to Oct 13, 2026 [D1]–[D6]; a seventh [D7] is background. ChainWard’s hire check and seller-demand check (paid API endpoints) and its Base node also predate the sprint. The checks compute nothing here, the node is only a data source, and the decodes’ own scripts and data are not used (their published counts serve only as checks): the chain was pulled again. Everything in the repository and this report is new sprint work: the case study was built on Oct 23, 2026 (UTC), and the work marked as added after the freeze on Oct 23 to 24. The code, the analysis and this report were written with AI coding agents; the team is responsible for every claim.

2 Related work

Collusion without talk. Q-learning pricing algorithms learn supracompetitive prices without communicating (Calvano et al., 2020); LLM pricing agents reach them quickly, and prompt wording moves them (Fish et al., 2024); LLM agents divide multi-commodity markets (Lin et al., 2024); seller communication raises collusion in LLM double auctions (Agrawal et al., 2025). Independent bandit learners can produce collusive-looking outcomes with no coordination at all (Douglas et al., 2024), which is the evidentiary problem in miniature. Collusion is one of three failure modes in the taxonomy of multi-agent risks by Hammond et al. (2025).

Detection and audit. Bonjour et al. (2022) detect colluding players from the mutual information between their actions. Colosseum (Nakamura et al., 2026) measures collusion through actions, as regret against the cooperative optimum, and compares it with what agents say; agents that plan collusion in text often pick non-collusive actions (“collusion on paper”). Agents can also hide coordination inside innocuous text (Motwani et al., 2024). Both results favour evidence from actions, and a payment ledger is such a record: complete and public, whatever the agents said.

Evidentiary standards. Under US antitrust law, parallel conduct is as consistent with independent decisions prompted by common perceptions of the market as with conspiracy, so a claim needs facts suggesting a preceding agreement; the commentators the Supreme Court cites look for parallel behaviour that would probably not result from chance, coincidence or independent responses to common stimuli (Bell Atlantic v. Twombly, 2007). Kovacic et al. (2011) define plus factors as actions and outcomes, beyond parallel conduct, that are largely inconsistent with unilateral conduct but consistent with explicit coordination, and propose ranking them by probative strength. Berger et al. (2025) ask how this framework applies when competitors coordinate through shared pricing algorithms. The sprint notes that ML work on collusion detection largely re-derives this doctrine without naming it. Plus factors are evidence of an agreement among competitors; we borrow their logic, not their legal category: here coordination links a seller with its buyers, or one operator’s wallets.

On-chain forensics. Address clustering infers common control of Ethereum accounts from usage patterns such as deposit-address reuse (Victor, 2020). Wash-trading studies find closed cycles of trades among accounts whose positions do not change (Victor and Weintraud, 2021) or statistical irregularities in reported volume (Cong et al., 2023). Our loop share applies the cycle idea to payments for services. What is new here is the evaluation: a frozen rule scored on held-out labelled cases from live agent markets and then on the whole market, positive and negative controls, and a mapping of each kind of evidence onto plus factors and parallel conduct.

3 Methodology

3.1 Markets and data

Base (x402). An x402 payment on Base is a USDC transfer that the payer authorizes with an EIP-3009 signature and a facilitator submits. We pulled every USDC Transfer and AuthorizationUsed log on Base from 2026-09-01

00:00 to 2026-10-12 23:59:59 UTC (blocks 50,715,727 to 52,530,126): 157,753,459 transfers, with no gaps and no thin segments. Of the 38 segments, 24 came from two public endpoints (at most 2 requests a second each) and the rest from ChainWard's node; 72 random 40-block ranges from the public segments were re-read from the node and matched row for row.

BNB Chain (TermiX). We pulled every log of TermiX's two escrow contracts from Sep 17 to Oct 8, 2026 (blocks 122,313,771 to 126,344,213), querying each chunk twice (0 disagreements), and the ERC-8004 owner of every agent hired. Our count of hires in the Set and Earn window, 10,325, equals the published count, as do all 7 daily counts; on 30 sampled settlements the escrow paid the agent's owner every time. Plain USDC and USDT transfers on BNB Chain could not be pulled at the allowed pace from keyless endpoints, so the BNB Chain evidence is the market's own hire graph.

3.2 Cases and labels

A case is one payee (an x402 seller's payTo address, or the owner of a hired agent) in one 7-day window: four Base windows from Sep 15 to Oct 13 and one BNB Chain window, Oct 1 to 8. A case enters only if a published decode makes a statement about that payee covering that window, and cases with no payments are dropped (3 were). Labels: **C** coordinated (the decode reports a loop, a closed ring or buyers funded by the seller); **I** independent (positive evidence such as many distinct funders or buyers who also pay other sellers); **N** no loop found (a weak negative that the decodes say is not a clearance); **L** linked, no loop (a strong common link, such as one wallet funding many buyers or an on-ramp, that the decode explicitly did not count as a loop). The detection target is C against the rest. The set has 98 C, 12 L, 8 N and 5 I cases (59 on Base, 64 on BNB Chain); Appendix C lists the families. Each family records the decode's quote and the evidence it used, so we can say which features re-derive the label. The two BNB Chain rings are rebuilt from the decodes' published descriptions and reproduce the published counts (48 wallets, 218 orders, 216 settled; 14 wallets, 38 hires inside, none in, one out). Defined by the closure of their hiring, they test the pipeline more than the evidence.

3.3 Definitions

All definitions were fixed before any case was scored. The *lookback* is the 14 days before the window. A *payment* on Base is a USDC transfer into the payee inside the window that the payer authorized with EIP-3009, directly or through a facilitator proxy that forwards it in the same transaction; on BNB Chain it is a TermiX order created for the payee's agent (payer = hirer, value = budget). *Forwarding*: a transfer X to Z that follows Y to X in the same transaction and carries 50% to 100% of it counts as Y to Z, so proxies and routers are passed through. A *hub* is an address with at least 1,000 distinct counterparties in lookback plus window (on BNB Chain, in the hire graph); hubs (exchanges, routers, bridges, custodial accounts) are never walked through. The *reach set* is every wallet reachable from the payee through non-hub wallets in at most $K = 4$ transfers in lookback plus window (on BNB Chain, through settled hires); Figure 1 shows it on three cases. The walk stops after 200,000 visited wallets; the 10 cases that reached this cap did so at hop 5 or 6, so the $K = 4$ features are complete. In the hypotheses, committed before any feature was computed, a plus factor is behaviour that makes sense only under coordination, above all paying fees to move one's own money in a circle; parallel conduct is similarity that shared infrastructure also produces (the same funder, a few payers, the same timing). *outside_funding_share* asks where the payers' money came from, so it sees loops one transfer longer than K .

Table 1: The evidence features (src/features.py) and the role the hypotheses gave each before any value was computed. Timing features subtract a null in which each payer’s payment times are shifted circularly inside the window (mean of 200 shifts).

Feature (role)	Definition
loop_share (plus factor, H1)	Share of payment value from payers in the payee’s reach set
loop_share_1hop	Same with $K = 1$: payers the payee paid directly
outside_funding_share	Share of the value flowing into payers (lookback plus window) from wallets outside the payee, the payers and the reach set; 1 if payers received nothing
shared_funder_share (parallel conduct, H2)	Value share of the payers behind the most common first funder, counting only funders of 2 or more payers; _nonhub: a hub never counts
payer_hhi (parallel conduct, H3)	Herfindahl index of payment value across payers
timing_coupling (timing, H4)	Share of payments within 1 hour after any transfer out of the payee
relay_share (timing, H4)	Share of payments within 10 minutes after the payer received money
Exploratory, not in the rule search	pair_reciprocity_timing, payer_synchrony (added after a debug run; src/features.py)

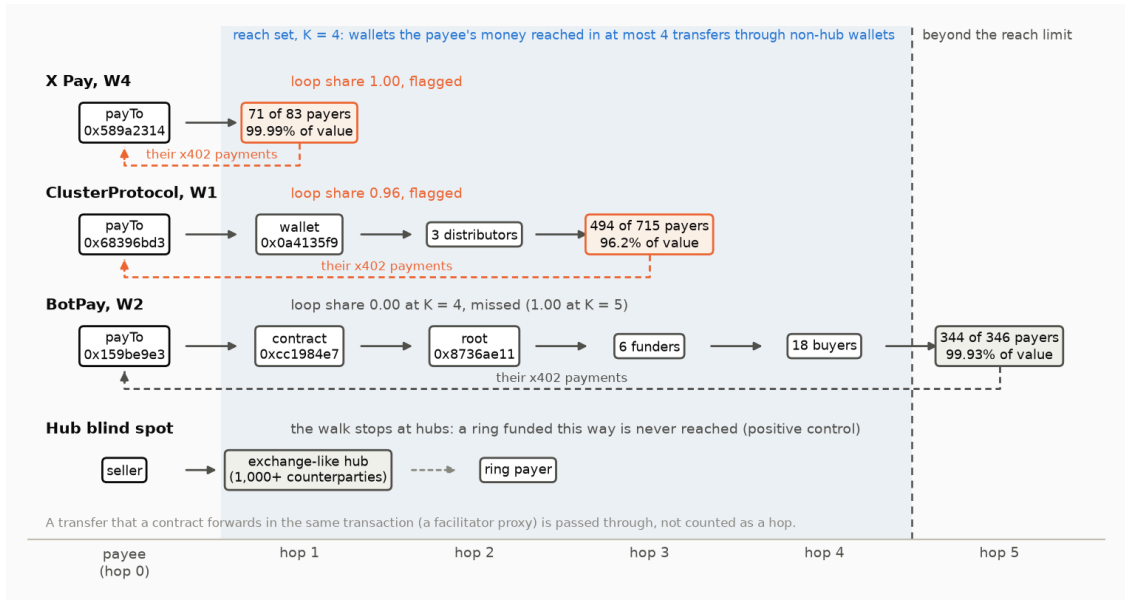


Figure 1. The reach walk, with worked examples from the committed case tables (added after the freeze; src/fig_reach_walk.py). X Pay’s payers are wallets its payTo funded directly (hop 1); ClusterProtocol’s sit three transfers out, behind one wallet and three distributors; BotPay’s sit five transfers out, one beyond the reach limit, so the frozen rule misses it. The walk stops at hubs, and a proxy that forwards a transfer in the same transaction is passed through. A ledger cannot tell one operator with many wallets from colluding agents: both leave these paths.

3.4 Protocol

1. The case registry and the split were committed first (dd05819, Oct 23 06:01 UTC): whole families, stratified by coordinated $\times$ chain with a seeded shuffle (seed 20261023), 10 development families (4 coordinated, 75 cases) and 10 test families (5 coordinated, 48 cases).
2. Hypotheses H1 to H5 and the positive-control prediction were committed at 06:50 UTC (d765207), before any feature was computed; features were then computed without printing test-half values.
3. The rule search read the development half only: every single-feature threshold and every two-feature AND or OR rule over the eight pre-registered features, scored by family-weighted balanced accuracy (each family weighs 1). Ties go to fewer features, then the wider margin (distance to the nearest development value), then the name. A second rule was chosen from the features the decodes did not use for labelling (payer_hhi, timing_coupling, relay_share), and the best threshold for each single feature was recorded.
4. The rules were frozen in commit f99f40eb42b69b7d90dcf76305db08a6b42f92d2 (11:10 UTC) and scored on the test half (563d219, 11:43 UTC). Commit times are local to the repository and not externally timestamped.

5. After the test-half results, a plan for the whole-market run, the spot-check and the end-to-end control was committed before any of their values was computed (eae4f78, 15:52 UTC; one amendment, ee4f211, 16:00 UTC). The frozen rule and its code were not changed.

Uncertainty comes from a bootstrap over families within each class ($B = 5,000$) and an exact permutation test over the 252 ways of choosing 5 coordinated test families. Deviations, all in the repository: a pre-freeze smoke test on a one-day slice printed feature values for three development cases and payment counts for test cases, and showed BotPay’s payers (a test family) five transfers from its payTo (no test-half feature value was seen; $K = 4$ was kept); the exploratory features came after it; the 48-wallet ring was first rebuilt by id range alone (rejected before any feature); a fine positive-control grid was added after the pre-set grid.

3.5 Controls

Positive control. Into the real data of each market the decodes did not label coordinated or linked (labels I or N, at least 5 payers; 12 markets once Losbeto, which the rule already flags, is excluded at reporting time; with it, every route shows 1 of 13 flagged) we inject a ring of 5 synthetic payers carrying a share ϕ of the seller’s payment volume (0.2% to 50%, 10 seeds per setting). The seller pays them directly, through 2 or 4 hops of relays, through 5 hops (one beyond K), through an exchange-like hub, or not at all, with one outside wallet funding all 5 (parallel conduct). This injection edits the per-case tables and sets each payer’s distance by hand, so it does not test the reach walk. **End to end** (added after the freeze), the ring is written as USDC transfers and EIP-3009 authorizations into the data the extraction reads, for three markets the rule does not flag, six routes (direct; two, three or four relays; via a real hub in two transactions; forwarded by it in one) and five shares, two just below and above the predicted cut-off; the unchanged extraction and features then run on a scratch database that reads the base data read-only. **Negative control.** Each payer’s payment times are shifted circularly inside the window (3 seeds), destroying timing dependence but keeping payers, amounts and the graph. **Sensitivity.** The frozen thresholds are applied to features recomputed with $K = 1$ to 6 and with hub thresholds of 300 and 3,000.

3.6 After the freeze: the whole market and a blinded spot-check

Market. On Base, every payee with at least \$10 of x402 payments in a week, an x402 payment being a payment as defined above whose transaction an address on x402scan’s public facilitator list sent (128 addresses at commit f205f8e, the list the decodes used; senders read from ChainWard’s node, exactly for payees with up to 1,000 payment transactions in the week and from 1,000 value-weighted draws for larger ones). The floor was set before any count: below it, payees are mostly developer tests and dust. Facilitators are excluded, and a proxy that forwards a payment in the same transaction credits its final recipient. On BNB Chain, every owner of an agent hired on TermiX in the Set and Earn window, no floor. The frozen extraction and features run unchanged (timing nulls 20 draws; the rule uses no timing).

Spot-check. Eligible are Base payees the decodes did not examine: not printed in any decode, not a case, never in a week’s top 30 by volume (the decodes traced the top sellers). We drew 25 payees the rule flags and 15 it does not (seed 20261024), one week each, shuffled into one list. Two separate AI agent sessions judged every payee from on-chain evidence alone, with read-only tools over the same transfers; neither saw a feature value, a flag, the rule or the stratum. Each gave a verdict (coordinated, independent or unclear), an estimated coordinated share and its reasoning, and both verdict files were committed before the key was joined. We also compare the rule with the one public label source ChainWard did not write that we found: Agent402’s seller index, whose Base router marks a seller `settlement_self_funded` when payments made with USDC the seller had sent the payer leave too little history (read Oct 23).

4 Results

4.1 The frozen rule on the test half

All 40 rules kept on the development leaderboard scored balanced accuracy 1.00, so the tie-break decided: `loop_share >= 0.4235`. In words, flag a payee when at least 42% of its payment value comes from payers that its own money reached within four non-hub transfers. The threshold sits midway between the development cases, which all lay near 0 or near 1 (Appendix D, Figure D1). On the test half (Table 2) the primary rule’s balanced accuracy of 0.82 is unlikely under chance: of the 252 ways to choose 5 coordinated families among the 10, 2.4%

score at least as well ($p = 0.024$). Unweighted, the test result is TPR 0.90 and FPR 0.17; over both halves the rule flagged 94 of 98 coordinated cases and 1 of 25 others. Outside funding share alone did better on the test half (0.90), but it tied with loop share on the development half and lost the margin tie-break; we report it as found and do not switch rules after seeing the test half.

Table 2: Frozen rules on the test half (10 families, 5 coordinated; 42 coordinated and 6 other cases). Rates are family-weighted, intervals are 95% bootstrap intervals over families, and p is the exact permutation p -value. The last two rows are best single-feature thresholds from the development half, frozen with the primary rule; Appendix D, Table D1 has the other five.

Rule (frozen on the development half)	Dev BA	Test TPR	Test FPR	Test BA [95% CI]	TP/FN/FP/TN	p
Primary: loop_share ≥ 0.4235	1.00	0.84	0.20	0.82 [0.56, 1.00]	38/4/1/5	0.024
Non-circular: timing_coupling ≥ -0.001939 AND relay_share ≥ 0.09579	0.93	0.54	0.20	0.67 [0.40, 0.91]	17/25/1/5	0.14
outside_funding_share ≤ 0.7537	1.00	1.00	0.20	0.90 [0.70, 1.00]	42/0/1/5	
payer_hhi ≥ 0.01951	0.75	0.58	1.00	0.29 [0.13, 0.45]	35/7/6/0	

4.2 Which evidence separates

The two plus-factor features separated the labelled cases on both halves: family-weighted AUC 1.00 and 0.91 for loop share on the development and test halves, 1.00 and 1.00 for outside funding share (Appendix D, Table D2 and Figure D2). The parallel-conduct features did not: a shared first funder scored 0.58 and 0.74, and payer concentration 0.71 and then 0.21, pointing the wrong way on the test half, where its rule flagged all six non-coordinated cases. Timing coupling (0.83, then 0.62) carried signal on the development half that mostly did not carry over.

4.3 Errors

Misses. All four false negatives are BotPay’s main payTo (0x159be9e3...) in its four weeks, with 298 to 346 payers each. Its payers sit five transfers from the payTo, one beyond the reach limit (Figure 1), as hypothesis H5 stated before scoring (written after a debug slice showed it); at $K = 5$ their loop share is at least 0.999 in every week. Outside funding share, which looks one transfer further, caught all four (at most 0.054).

False positive. Losbeto in W4 (0xd5ba9711..., label N): one payer carried 97% of its payment value (\$158.79), and in lookback plus window the seller sent that same wallet \$406.17. The decode read the seller’s outflows as payouts to people. By the ledger, the payee and its main payer pay each other; what the payments were for (refunds, winnings, wages or self-funding) is not on-chain. The development half holds the mirror case: F07, whose seller sent 91% of its takings back to its buyers [D1], is labelled C but contested, since it may be a game working as designed.

4.4 Positive controls

When the seller’s money reached the injected ring within four transfers, the ring was flagged in 100% of runs at 50% of the seller’s volume and in 0% at 30% or less (Appendix D, Figure D3). A ring with share ϕ in a market with existing loop share b has loop share $\phi + (1 - \phi) \cdot b$, so it is flagged once $\phi \geq (0.4235 - b)/(1 - b)$: 31.5% to 42.4% across the 12 markets. Over five hops or behind an exchange-like hub the ring was never flagged at any size, nor were five payers funded by one outside wallet with no loop. The rule misses partial rings below about a third of a seller’s volume: the development set held none, so the margin rule put the threshold midway between 0.05 and 0.80.

End to end, through the raw transfers, the extraction found every synthetic payment and put every ring payer at its designed distance in all 90 runs (hop 1, 3, 4 or 5; not reached behind the hub; hop 1 when the hub forwarded in the same transaction). Measured loop share equalled $\phi + (1 - \phi) \cdot b$ on routes within four transfers and $(1 - \phi) \cdot b$ on the others, to within 1.1×10^{-16} , and the flag flipped between $\phi^* - 0.02$ and $\phi^* + 0.02$ (ϕ^* the cut-off above) exactly as predicted, with $b = 0, 0$ and 0.158 as committed for the three markets. The base data was read-only throughout and its row counts did not change (Appendix E, Table E2).

4.5 Timing, sensitivity and the hypotheses

Timing is the evidence the decodes did not use. Circularly shifting each payer’s payment times took the coordinated cases’ mean timing coupling from 0.36 to -0.02 and the non-circular rule’s flag rate on them from 0.50 to 0.00, while the primary rule flagged the same cases (Appendix D, Figure D4): the signal is real but weak (test 0.67, $p = 0.14$), with thresholds inside the null’s sampling noise. Relay share is high in non-coordinated markets too (0.32, against 0.41): checkout flows fund a wallet and pay at once (Korus, 0.998). At $K = 1$ to 6 the frozen threshold gave test TPR 0.59, 0.59, 0.80, 0.84, 1.00 and 1.00 and FPR 0.20, except 0.40 at $K = 6$ (CheapTokens’ loop share rises from 0.16 to 0.75); hub thresholds of 300 and 3,000 changed nothing. $K = 5$ would have caught BotPay, but K was fixed before scoring. H1 held. H2 (a shared-funder rule flags the L cases) failed as stated: the frozen threshold (0.857) sat above every development non-coordinated case, so the rule flagged no L case, and AX1’s seeding wallet, expected not to be a hub, is one (hub-excluded share 0.03, not 0.47); only its premise held (AUC 0.63). H3, H4, H5 and the positive-control prediction held.

5 After the freeze: the whole market

Table 4: The frozen rule on every x402 payee on Base with at least \$10 of x402 payments in a week, and on every Set and Earn payee on BNB Chain (added after the freeze; results/market_summary.json). The last column is ChainWard’s own published share of weekly x402 volume in verified loops (a floor, same method [D2]): the labeller’s number, given as context only.

Market, week	Payee-weeks	Volume	Flagged	Volume flagged	Decodes’ loops
Base W1 (Sep 15 to 22)	216	\$126,527	89 (41%)	49.4%	$\geq 46.8\%$
Base W2 (Sep 22 to 29)	190	\$230,801	56 (29%)	26.8%	$\geq 25.7\%$
Base W3 (Sep 29 to Oct 6)	179	\$110,254	71 (40%)	44.2%	$\geq 40.6\%$
Base W4 (Oct 6 to 13)	165	\$207,363	42 (25%)	12.7%	
Base, all weeks	750	\$674,944	258 (34%)	29.5%	
BNB Chain (Oct 1 to 8)	9,131	\$495,424	1,276 (14%)	14.6%	

Base. 750 payee-weeks (431 payees) passed the \$10 floor; their x402 volume in W1 to W3 is 98.7%, 99.1% and 101.0% of the decodes’ published weekly totals [D2]. The rule flags 258 payee-weeks (34%), 135 of the 431 payees in some week, and 29.5% of the volume; no reach walk was capped before hop 4. One payTo is 67.5% of W4: \$140,000 in two payments, the round trip through a high-throughput account that the decodes count apart (family F19, label L). The rule does not flag it; without it, 39.1% of W4 is flagged. Ranked by weekly volume, as x402scan and Agent402 rank sellers, 25.3% of the top ten’s volume and 29.4% of the top fifty’s is flagged. On Agent402’s published top ten of Oct 13 [D4] the rule flags ranks 1, 5, 8 and 10 in W4, 39.7% of the board’s USDC: X Pay and keyring-agent (self-funded per the decode), Losbeto (the test-half false positive) and Storm Station 247 (not traced); it misses BotPay (rank 7, five hops). **BNB Chain.** All 62 ring wallets are flagged, and 1,214 of the other 9,069 payees: 14.6% of order budgets, against the 90.1% the decode classes as cycling and linked, mostly by timing [D6]. With only the hire graph visible, money that leaves through exchanges or plain transfers is never seen coming back, so there the rule’s share is a floor.

Spot-check. Of the 25 flagged payees no decode had examined, judge A called 22 coordinated (88%, 95% interval 69% to 97%) and judge B 20 (80%, 59% to 93%; two unclear); of the 15 unflagged, each called 4 coordinated (27%, 8% to 55%). The judges agreed on 38 of 40. Both called three flagged payees independent, all the same error: the payee had sent its payers small refunds or change (once hours after the payment), and the reach walk counts a payer the payee paid whatever the amount or the order. The four misses are the blind spots: two payers whose only USDC came from the payee in May, before the data starts, and two wallets of BotPay’s ring at five hops. Post hoc, 48 of the 135 flagged payees are paid by that one ring, flagged in all 86 of their weeks in W1 to W3 and in none of their 21 in W4, when every one’s main payer sat at hop 5 (Appendix E): counts of flagged payees overstate the number of operators.

A public label source. Agent402’s index gives a settlement verdict to 58 payees of the Base market. Of the 9 it marks self-funded, the rule flags 7 in some week; of the 49 others, 19 (37 of 58 agree). Agent402 nets one hop over its own history to Oct 23 and marks a seller only when what remains falls below its floor, so a seller with

enough outside demand is never marked: this is agreement, not accuracy. We know of no public label source for Set and Earn.

6 Discussion

What should count as convincing evidence of collusion? On a payment ledger, our answer has three parts.

1. *Count money that comes back, and buyers without money of their own.* Loop share and outside funding share separated the labelled cases on both halves (test balanced accuracy 0.82 and 0.90), loop share behaved as designed on synthetic rings written into the raw data, and on the whole market it flagged about the share of weekly volume the decodes had counted by hand. In these data they behave like plus factors (but see Circularity below). An independent buyer spends its own money; a seller that funds its buyers pays fees to move its own money in a circle, which only pays if something rewards measured demand (a leaderboard rank, campaign rewards, points). That conduct is largely inconsistent with unilateral conduct and largely consistent with explicitly coordinated action, which is how Kovacic et al. (2011) define a plus factor.
2. *Treat a shared funder, a few large payers and synchronized timing as grounds to look, not to conclude.* Shared infrastructure can produce all three without any agreement, as the decodes read these cases: an app seeding embedded wallets (AX1), checkout wallets funded through an on-ramp (Korus), an account both sides may use (the round-trip family), a few large customers (SurplusIntelligence). If so, these are independent responses to common stimuli, in the phrase Twombly quotes; they separated poorly (AUC 0.47 to 0.71), partly because the decodes counted only loops as coordination. Set and Earn’s terms exclude both wallets with a common funding source and wallets that transact circularly [D5]; our results support the second test much more than the first.
3. *Even a plus factor is not proof of intent.* Payouts, refunds or wages make loops (Losbeto and the spot-check’s three false positives; F07 may be a game), and the ledger shows where money went, not who controls a wallet or why. A market can treat a loop as shifting the burden: the seller explains it, or its demand is discounted.

Circularity. The labels were mostly reached by tracing where buyers’ money came from, close to what loop share and outside funding share measure (Appendix C). The test shows that a frozen, mechanical version of that reasoning reproduces the decodes’ judgments on unseen families, over every payment rather than the samples the decodes traced; it does not validate the evidence independently. The rule on features the decodes did not use (0.67, $p = 0.14$) does not lean on the labels’ evidence, and it is weak. The spot-check judges were blinded and never saw the rule, but they are AI agents reading the same money paths, so they check the rule’s reading of a ledger, not its truth; the positive controls are label-free but synthetic.

For AI collusion research. Ledgers give action evidence that agents cannot edit and that needs no access to their messages, a complement to monitors that read talk, which can overstate collusion (“collusion on paper”, Nakamura et al., 2026) or miss it, as with Colosseum’s hidden collusion or steganographic talk (Motwani et al., 2024). A published threshold invites evasion; a sturdier standard asks for independence a ring cannot cheaply fake (Appendix A).

7 Conclusion

On two live markets built for AI agents, a rule frozen before it saw the test half reproduced the decodes’ split of coordinated from independent demand by asking one question: did the payee’s own money come back as its payments? Whether the buyers had outside money did as well; a shared funder, a few large payers and synchronized timing, the ledger’s parallel conduct, did not. Run unchanged on the whole market, it flagged a third of Base x402 payee-weeks and 30% of their volume, and blinded judges called 80% to 88% of the flags they checked coordinated; the three both rejected were refunds, which weighting a loop by the money that comes back, in time order, would clear. This suggests a screening standard for markets that rank or reward agents by demand, with known limits: one team’s labels that partly share the rule’s evidence, ten test families, and blind spots at exchanges, beyond four transfers and outside the data. The next step is labels the rule did not inform, from another team or a market’s own enforcement.

Code and Data: <https://github.com/saltxd/plus-factors-on-the-ledger>

References

- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025). Evaluating LLM agent collusion in double auctions. [arXiv:2507.01413](https://arxiv.org/abs/2507.01413). <https://arxiv.org/abs/2507.01413>
- Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007). <https://www.courtlistener.com/opinion/145730/bell-atlantic-corp-v-twombly/>
- Berger, L. F., Weiner, M. L., and Jiang, W. (2025). Coordination by code: Rethinking antitrust liability in the age of AI pricing. *Antitrust* (ABA Section of Antitrust Law), 40(1), Fall 2025. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/
- Bonjour, T., Aggarwal, V., and Bhargava, B. (2022). Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022)*, PMLR 180, 223–232. <https://proceedings.mlr.press/v180/bonjour22a.html>
- Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
- Coinbase (n.d.). x402: A payments protocol for the internet, built on HTTP (specification and reference implementation). <https://github.com/x402-foundation/x402> (moved from github.com/coinbase/x402, now a fork), accessed Oct 23, 2026.
- Cong, L. W., Li, X., Tang, K., and Yang, Y. (2023). Crypto wash trading. *Management Science*, 69(11), 6427–6454. <https://doi.org/10.1287/mnsc.2021.02709>
- De Rossi, M., Crapis, D., Ellis, J., and Reppel, E. (2025). ERC-8004: Trustless Agents. Ethereum Improvement Proposals. <https://eips.ethereum.org/EIPS/eip-8004>
- Douglas, C., Provost, F., and Sundararajan, A. (2024). The illusion of collusion. [arXiv:2411.16574](https://arxiv.org/abs/2411.16574). <https://arxiv.org/abs/2411.16574>
- Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2024). Algorithmic collusion by large language models. [arXiv:2404.00806](https://arxiv.org/abs/2404.00806). One-page abstract in *Proceedings of the 27th ACM Conference on Economics and Computation (EC '26)*. <https://arxiv.org/abs/2404.00806>
- Hammond, L., Chan, A., Clifton, J., et al. (2025). Multi-agent risks from advanced AI. Cooperative AI Foundation, Technical Report #1. [arXiv:2502.14143](https://arxiv.org/abs/2502.14143). <https://arxiv.org/abs/2502.14143>
- Kim, P. J., Britz, K., and Knott, D. (2020). ERC-3009: Transfer With Authorization. Ethereum Improvement Proposals. <https://eips.ethereum.org/EIPS/eip-3009>
- Kovacic, W. E., Marshall, R. C., Marx, L. M., and White, H. L. (2011). Plus factors and agreement in antitrust law. *Michigan Law Review*, 110(3), 393–436. <https://doi.org/10.36644/mlr.110.3.plus>
- Lin, R. Y., Ojha, S., Cai, K., and Chen, M. F. (2024). Strategic collusion of LLM agents: Market division in multi-commodity competitions. [arXiv:2410.00031](https://arxiv.org/abs/2410.00031). <https://arxiv.org/abs/2410.00031>
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., and Schroeder de Witt, C. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. https://papers.nips.cc/paper_files/paper/2024/hash/861f7dad098aec1c3560fb7add468d41-Abstract-Conference.html
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. In *NeurIPS 2026 Evaluations and Datasets Track* (to appear). [arXiv:2602.15198](https://arxiv.org/abs/2602.15198). <https://arxiv.org/abs/2602.15198>
- Victor, F. (2020). Address clustering heuristics for Ethereum. In *Financial Cryptography and Data Security (FC 2020)*, Lecture Notes in Computer Science, vol. 12059, Springer, 617–633. https://doi.org/10.1007/978-3-030-51280-4_33
- Victor, F., and Weintraud, A. M. (2021). Detecting and quantifying wash trading on decentralized cryptocurrency exchanges. In *Proceedings of the Web Conference 2021 (WWW '21)*, 23–32. <https://doi.org/10.1145/3442381.3449824>
- Preexisting ChainWard decodes (case labels and background, written before the sprint)**
- [D1] ChainWard (2026, Sep 27). The #2 x402 seller on Base funded its own buyers. <https://chainward.ai/decodes/x402-on-base>
- [D2] ChainWard (2026, Oct 6). x402 on Base, two weeks later: The loops kept running. <https://chainward.ai/decodes/x402-on-base-two-weeks-later>
- [D3] ChainWard (2026, Oct 9). Six x402 wallets paid only by each other now rank 3 to 8 on Base. <https://chainward.ai/decodes/x402-closed-six>

- [D4] ChainWard (2026, Oct 13). Three of the top ten x402 sellers on Agent402’s board pay their own buyers.
<https://chainward.ai/decodes/x402-top-ten-self-funded>
- [D5] ChainWard (2026, Oct 5). Set and Earn, week one: Who is hiring whom on BNB Chain.
<https://chainward.ai/decodes/set-and-earn-week-one>
- [D6] ChainWard (2026, Oct 8). Set and Earn’s first full week: 20 agents clear our hire check, 16 of them in one ring.
<https://chainward.ai/decodes/set-and-earn-week-one-closed>
- [D7] ChainWard (2026, Oct 2). TermiX collected \$507K in job fees. 99% came from wallets hiring each other.
<https://chainward.ai/decodes/termix-on-chain>

Data sources used after the freeze (not literature)

x402scan’s facilitator list: github.com/Merit-Systems/x402scan, `packages/external/facilitators`, commit `f205f6e6fa96924adecb3619e5bf6edbf1e1e697` (Aug 17, 2026). Agent402’s seller index: <https://agent402.tools/api/index>, read Oct 23, 2026, 16:04 UTC; its Oct 13 top ten as printed in [D4].

Appendix A: Limitations and Dual-Use / Ethical Considerations

What this establishes, and what it does not. On 10 held-out families from two live markets, a rule frozen in advance reproduced published judgments of coordinated demand with family-weighted balanced accuracy 0.82 (95% interval 0.56 to 1.00), and the evidence it uses behaved as designed on synthetic rings, end to end. Run on the whole market after the freeze, it flagged about the share of volume the decodes had counted by hand. It does not establish that its flags are right where no one has looked closely, that it carries over to other markets, tokens or chains, or that the operator of any flagged wallet meant to deceive anyone.

Limitations.

- *Circularity of labels.* Most labels were reached by tracing where buyers’ money came from, close to what `loop_share` and `outside_funding_share` measure (Appendix C). Without that evidence the frozen rule scored 0.67 ($p = 0.14$). The two BNB Chain rings are defined by the closure of their hiring, so their loop share is high almost by construction.
- *Small test set.* 10 test families, 5 of them coordinated; the interval for balanced accuracy runs from 0.56 to 1.00. The test negatives are 6 cases in 5 Base families. All L families and the only BNB Chain negative family fell in the development half, so the result on linked-without-loop cases (0 of 12 flagged) is in-sample, and no BNB Chain negative was tested.
- *The threshold misses partial rings.* A ring below 31.5% to 42.4% of a seller’s volume (depending on the market) is never flagged. The development cases sat near 0 or near 1, so the margin rule put the threshold far from where small rings sit.
- *Blind spots, by construction.* Loops through exchanges and other hubs, and loops longer than four transfers (BotPay), are invisible to the reach walk (positive control: 0% at every size, end to end as well). Only USDC on Base is followed: other tokens, ETH and other chains are not. BNB Chain evidence is TermiX’s hire records only, so funding through exchanges or plain transfers is unseen there, and the whole-market share on BNB Chain is a floor. A payment is an EIP-3009-authorized transfer, the default way x402 settles USDC, so other EIP-3009 transfers to a payee also count, and x402 payments settled another way are missed.
- *The market run’s population.* x402 is defined by x402scan’s facilitator list at one commit, so facilitators not on it are missed. For payees with more than 1,000 payment transactions in a week, x402 volume is estimated from 1,000 value-weighted draws of transaction senders.
- *Label quality.* One team wrote every case label. Decodes publish striking cases, so partial rings and borderline sellers are under-represented; N is a weak negative; F07 is contested (it may be a game); the rings were rebuilt from published descriptions, though they reproduce the published counts. The spot-check judges are AI agent sessions, not human investigators, and Agent402’s verdict is a router gate over a different period.
- *Money, not control.* The ledger shows where money went, not who controls a wallet or what a payment was for. A seller paying a customer back for a game, a refund or a contractor’s work makes the same loop as self-funding (Losbeto against F07).
- *No talk.* Payment ledgers carry no messages, so case type (c) of question 2.3, collusive talk that never changes behaviour, cannot be tested here. The nearest analogue is a visible link without a loop (the 12 L cases, 0

flagged), which is not talk and is in-sample.

- *Timing noise and timestamps.* The timing rule’s thresholds sit inside the null’s sampling noise; its outputs change with the number of null draws. The pre-registration rests on local commit times, which are not externally timestamped.

Dual use. Published thresholds can teach rings to stay under them: keep a ring below about a third of a seller’s volume, route the seller’s money through an exchange or another hub, add a fifth hop, or fund the buyers from wallets the seller never paid. The positive controls measure the first three directly (0% detection). We publish anyway, for four reasons. Each evasion has a cost: exchange round trips cost fees and time and, on regulated exchanges, tie wallets to identities; longer chains cost fees; outside funding means bringing capital that cannot simply be recycled; a smaller ring makes less fake demand. A market can require positive evidence that a ring cannot cheaply fake, such as outside funders and buyers who also pay other sellers, instead of the mere absence of a detected loop. Defenders can vary K , the hub threshold and the cut-off privately per market; the protocol, not the number, is the contribution. And cycle detection in transfer graphs is already public in the wash-trading literature. Every Base case payee was named, in full or by prefix, in a decode before the sprint; the per-case tables add only the public payment records of those payees. The 62 BNB Chain ring wallets appear only under pseudonyms, with none of their agent ids, orders or transactions: the decodes describe these rings without listing them (one names four members as hiring examples), and the code can rebuild them from public escrow logs, which the repository does not include. The two other BNB Chain case payees, owners of two agents the decode cleared, appear by address. The whole-market run adds per-payee outputs, addresses and numbers only, for 431 Base payees and 9,131 BNB Chain payees (ring wallets again under pseudonyms), most of which no decode examined, and the spot-check adds two AI judges’ readings of 40 payees’ public money paths; a flag or a verdict there is a measurement of where money went, not a finding about any seller.

Privacy and ethics. Only public on-chain data, published decodes and public third-party listings are used. Cases are wallets and contracts, not people, and we make no claim about who controls them. No private label sets, tip clusters or paid wallet lists are used. A label of “coordinated” can hurt a seller’s reputation if it is wrong; we report every error, and the one false positive (Losbeto) is a case where the ledger cannot say why money moved, so we do not claim it is self-funded.

LLM Usage Statement. AI coding agents wrote the code and this report under the author’s direction; the author chose to enter, approved the approach and decided what to publish. Two further AI agent sessions served as the blinded spot-check judges (Section 3.6). Every number was checked against the result files and every reference against its source in separate review passes before submission. The author is responsible for every claim, number and reference.

Competing interests. ChainWard sells on-chain wallet checks built on similar evidence, and its published investigations supplied the case labels.

Appendix B: Reproduction

Code, case registry, per-case evidence tables and results: <https://github.com/saltxd/plus-factors-on-the-ledger>. The raw pulls are not included: Base’s is about 4.5 GB, and BNB Chain’s contains wallet lists our decodes chose not to publish; the scripts in `pull/` re-pull both (below). Nor are the per-transaction tables of the 63 cases that contain a ring wallet (the 62 ring cases and one F20 case with a ring wallet among its payers), whose agent ids, order ids, blocks and times would identify the ring wallets; their case-level features and control rows are (data/withheld/), so the commands below run from the committed data and give the committed results. Rebuilding those tables needs the raw BNB Chain pull and `data/raw/bsc/ring_spec.json`, which holds the agent-id range and the example hire that the decodes print. Python 3.12; `pip install -r requirements.txt`. Every number in this report is read from `results/` (generated by `src/report.py`) or from `data/`.

```

# From the committed per-case tables (data/cases*) and data/withheld, no chain access needed
python src/run_features.py --sensitivity          # all features; test-half values not printed
python src/select_rule.py                      # development half only -> results/frozen_rule.json
git show f99f40e:results/frozen_rule.json      # the rules as frozen before test scoring
python src/evaluate.py                        # test half, bootstrap, AUCs, errors
python src/controls.py                       # positive control, shuffle, permutation test
python src/controls.py --which boundary        # fine grid near the threshold
python src/sensitivity.py                     # K = 1..6, hub >= 300 and >= 3000
python src/report.py                          # case table, figures, results/results.md
pytest tests/

# From the chain: Base (any archive endpoint with logs for Sep 1 to Oct 12, 2026)
python pull/base_usdc_pull.py --rpc <URL> --start 50715727 --end 52530126 --out data/raw/base
python scripts/check_coverage.py --raw data/raw/base --start 50715727 --end 52530126
python src/extract_base.py --db data/base.duckdb --raw data/raw/base --load
python src/extract_base.py --db data/base.duckdb --out data/cases --hub 1000 --k 6
python src/extract_base.py --db data/base.duckdb --out data/cases_h300 --hub 300 --k 6
python src/extract_base.py --db data/base.duckdb --out data/cases_h3000 --hub 3000 --k 6

# BNB Chain (the raw pull is not included; these lines re-pull it and rebuild the tables;
# first write data/raw/bsc/ring_spec.json from the decodes, see the README)
python pull/bsc_termix_pull.py anchors --start 122313771 --end 126344213 --seg 50000
python pull/bsc_termix_pull.py escrow --start 122313771 --end 126344213
python src/extract_bsc.py --stage orders
python pull/bsc_termix_pull.py owners --ids data/raw/bsc/agents.txt --block 126344213
tail -n +2 data/bsc_settle_sample.csv | cut -d, -f1 > data/raw/bsc/settle_sample.txt
python pull/bsc_termix_pull.py receipts --txs data/raw/bsc/settle_sample.txt
python src/extract_bsc.py --stage cases

# Added after the freeze: whole market, spot-check, end-to-end control, reach-walk figure
python pull/x402scan_facilitators.py          # facilitator list at commit f205fbc
python src/market_base.py candidates          # payees with >= $10 of payments; senders to read
python pull/base_tx_senders.py --pairs data/raw/base_senders/pairs_candidates.txt.gz \
--rpc <URL> --start 51320527 --end 52530126 # on ChainWard's node, about 15 minutes
python src/market_base.py population          # -> results/market_base_population.csv
python src/market_base.py extract && python src/market_base.py features --csv
results/market_base_features.csv
python src/market_bsc.py                     # -> results/market_bsc_features.csv
python -I pull/agent402_index.py --raw data/raw/agent402 --out data/raw/agent402/base_verdicts.csv
python src/market_summary.py                 # -> results/market_summary.json
python src/spotcheck.py unblind --features results/market_base_features.csv --key
results/spotcheck_key.csv
python scripts/diag_botpay_ring.py           # post hoc ring count, after unblinding
python src/positive_control_e2e.py           # -> results/positive_control_e2e*
python src/fig_reach_walk.py                 # -> results/fig_reach_walk.png

# This report (pandoc and Google Chrome)
bash report/build_pdf.sh

```

Appendix C: Case families, windows and the evidence behind each label

Table C1: Windows (UTC, end exclusive).

Window	Chain	Dates	Blocks
W1	Base	Sep 15 to Sep 22, 2026	51,320,527 to 51,622,926
W2	Base	Sep 22 to Sep 29	51,622,927 to 51,925,326
W3	Base	Sep 29 to Oct 6	51,925,327 to 52,227,726
W4	Base	Oct 6 to Oct 13	52,227,727 to 52,530,126
SE	BNB Chain	Oct 1 to Oct 8	125,000,755 to 126,344,213

Table C2: Case families. “Decode evidence” is the evidence the decode used to reach its label (label_basis in cases/cases.yaml); “closest feature” is the feature that re-derives it (results/circularity.csv). Flagged = cases the frozen primary rule flags.

Family	Chain	Half	Labels (cases)	Flagged	Decode evidence	Closest feature	Source
F01-clusterprotocol	Base	test	C 3	3 of 3	seller to buyer funding path, seller outflows to one hub	loop share, no matching feature	D1, D2
F02-meridian-six	Base	test	C 18	18 of 18	closure of payment set, pair matrix, no outside payments	no matching feature, outside funding share	D1, D2, D3
F03-meridian-seventh	Base	dev	C 3	3 of 3	seller to buyer funding path	loop share	D2, D3
F04-botpay	Base	test	C 5	1 of 5	seller to buyer funding path, payment tree	loop share	D1, D2, D4
F05-keyring-pingpong	Base	dev	C 4	4 of 4	direct reciprocity, seller funds buyers	loop share, loop share 1hop	D1, D2, D4
F06-xpay	Base	test	C 2	2 of 2	seller funds buyers, first funder is seller	loop share, shared funder share	D2, D4
F07-payback-ed9fcd0d (contested)	Base	dev	C 1	1 of 1	direct reciprocity	loop share 1hop	D1, D2
F08-se-ring48	BNB	dev	C 48	48 of 48	closure of hiring set, shared funder	loop share, shared funder share	D5, D6
F09-se-group14	BNB	test	C 14	14 of 14	closure of hiring set, shared funder	loop share, shared funder share	D6
F10-clashofcoins	Base	dev	I 1	0 of 1	many distinct funders	shared funder share	D1
F11-blockrun	Base	dev	I 2, N 1	0 of 3	buyers pay other sellers, amount diversity, no seller funding	loop share, no matching feature	D1, D2, D4
F12-stableenrich	Base	test	I 1, N 1	0 of 2	buyers pay other sellers	no matching feature	D1, D4
F13-claw402	Base	test	I 1	0 of 1	many distinct funders	shared funder share	D1
F14-surplusintelligence	Base	test	N 1	0 of 1	no loop in seller outflow trace	loop share	D4
F15-losbeto	Base	test	N 1	1 of 1	no loop in seller outflow trace	loop share	D4
F16-cheaptokens	Base	test	N 1	0 of 1	no loop in seller outflow trace	loop share	D4
F17-ax1	Base	dev	L 1, N 1	0 of 2	shared funder without seller path	shared funder share	D1, D4
F18-korus	Base	dev	L 1	0 of 1	shared high-throughput funder	shared funder share	D2
F19-roundtrip-ee7ae85f	Base	dev	L 10	0 of 10	money returns through high-throughput account	loop share	D1, D2
F20-se-clear-outside	BNB	dev	N 2	0 of 2	no funding link pairwise	loop share	D5, D6

Appendix D: Tables and figures from the case study

Table D1: Frozen rules on the test half (10 families, 5 coordinated; 42 coordinated and 6 other cases). Rates are family-weighted, intervals are 95% bootstrap intervals over families, and p is the exact permutation p-value. The last seven rows are the best single-feature thresholds from the development half, frozen with the primary rule and shown for comparison (the main text’s Table 2 shows four of these rows).

Rule (frozen on the development half)	Dev BA	Test TPR	Test FPR	Test BA [95% CI]	TP/FN/FP/TN	p
Primary: loop_share >= 0.4235	1.00	0.84	0.20	0.82 [0.56, 1.00]	38/4/1/5	0.024
Non-circular: timing_coupling >= -0.001939 AND relay_share >= 0.09579	0.93	0.54	0.20	0.67 [0.40, 0.91]	17/25/1/5	0.14
outside_funding_share <= 0.7537	1.00	1.00	0.20	0.90 [0.70, 1.00]	42/0/1/5	
loop_share_1hop >= 0.125	1.00	0.59	0.20	0.69 [0.40, 0.98]	33/9/1/5	
timing_coupling >= 0.04143	0.84	0.59	0.20	0.70 [0.44, 0.91]	17/25/1/5	
relay_share >= 0.09579	0.72	0.64	0.40	0.62 [0.33, 0.88]	26/16/2/4	
shared_funder_share >= 0.8571	0.75	0.20	0.00	0.60 [0.50, 0.80]	2/40/0/6	
shared_funder_share_nonhub >= 0.6354	0.75	0.20	0.00	0.60 [0.50, 0.80]	2/40/0/6	
payer_hhi >= 0.01951	0.75	0.58	1.00	0.29 [0.13, 0.45]	35/7/6/0	

Table D2: Single-feature AUC, family-weighted, oriented so that higher means more coordinated (outside funding share reversed); 0.5 is chance. Dev = development half, all = all cases. *Exploratory.

Feature	Dev	Test	All	Feature	Dev	Test	All
Loop share (K = 4)	1.00	0.91	0.97	Payer concentration (HHI)	0.71	0.21	0.47
Outside funding share	1.00	1.00	1.00	Timing coupling	0.83	0.62	0.70
Loop share (1 hop)	1.00	0.74	0.87	Relay share	0.61	0.57	0.61
Shared first funder	0.58	0.74	0.63	Pair reciprocity timing*	0.84	0.51	0.66
Shared funder, hubs excluded	0.58	0.82	0.71	Payer synchrony*	0.67	0.40	0.49

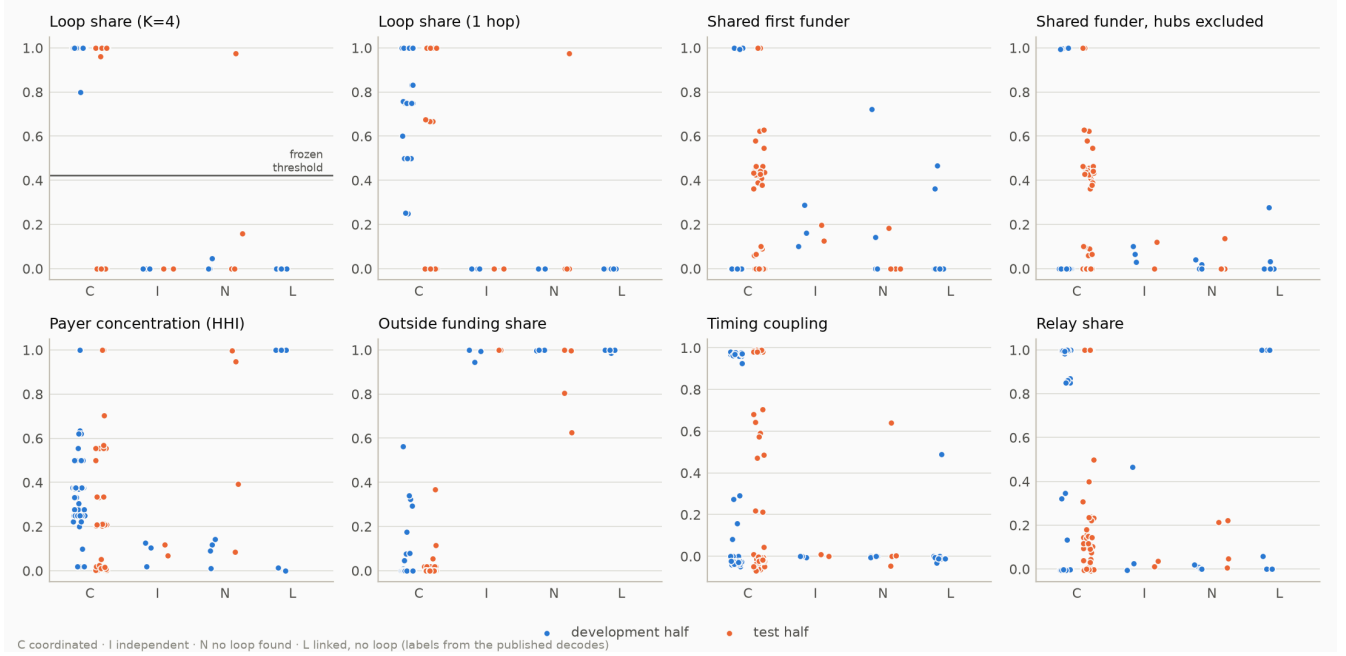


Figure D1. Every pre-registered feature by label, development half (blue) and test half (orange); the line in the first panel is the frozen threshold. C coordinated, I independent, N no loop found, L linked without a loop.

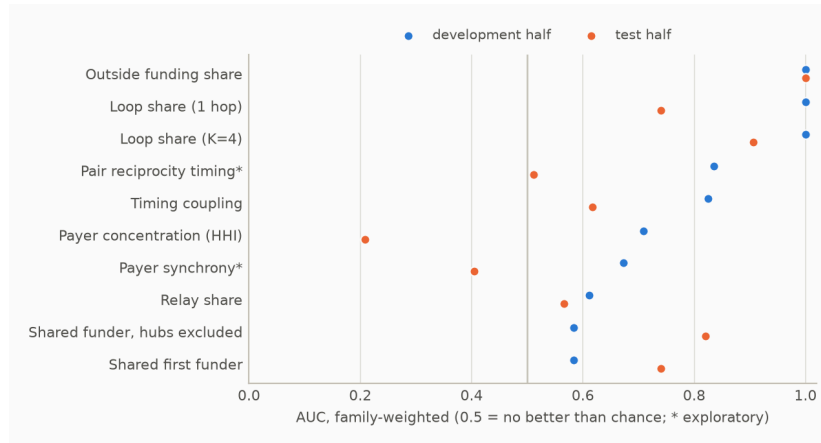


Figure D2. Single-feature AUC (family-weighted) on the development and test halves; 0.5 is chance.

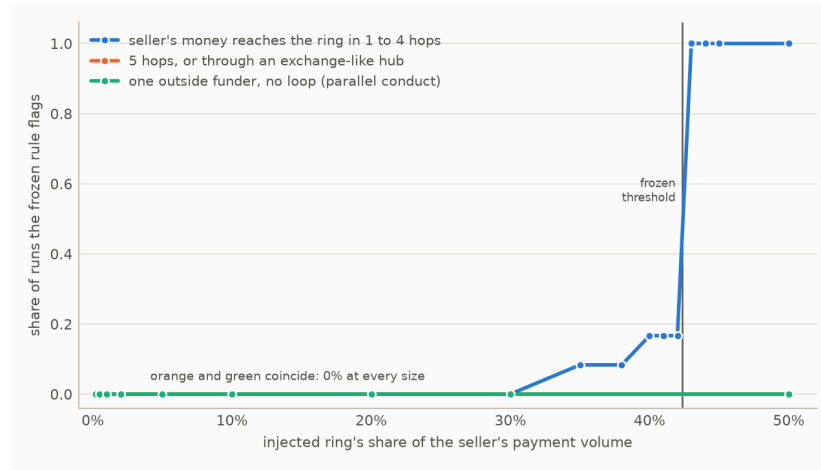


Figure D3. Positive control: share of injection runs the frozen rule flags, by the injected ring's share of the seller's payment volume (12 real markets, 10 seeds per setting; points between 30% and 50% come from the fine grid). The orange and green lines coincide at 0%.

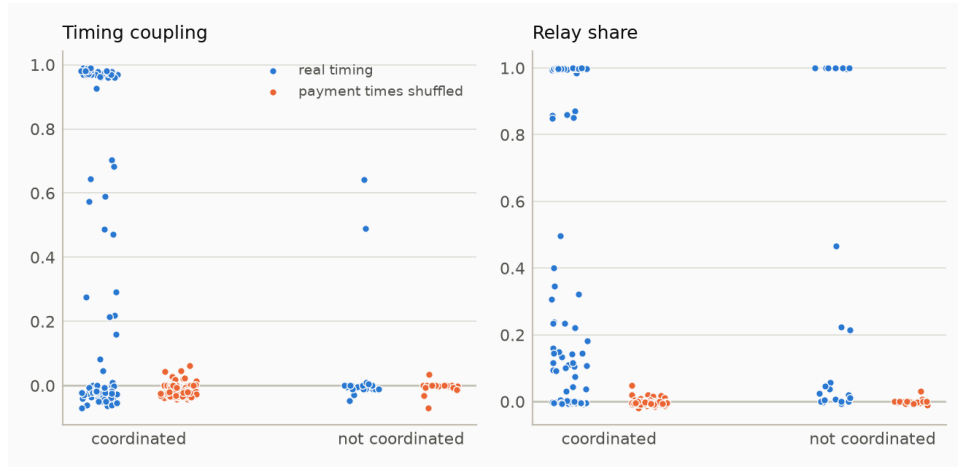


Figure D4. Negative control: timing features with real (blue) and circularly shifted (orange) payment times, all cases, coordinated and not. Shifting each payer's payment times inside the window removes the coordinated cases' timing signal (mean timing coupling 0.36 to -0.02) and leaves the primary rule's output unchanged.

Appendix E: After the freeze, details

Table E1: The frozen rule on the whole market, by weekly x402 volume, by weekly rank and by kind of BNB Chain payee (results/market_summary.json); Base bands are by x402 volume in the week. The last column is the share of volume the frozen outside_funding_share ≤ 0.7537 flags.

Payees	Payee-weeks	Volume	Flagged	Volume flagged	Outside-funding rule
Base, \$10 to \$100	548	\$16,147	179 (33%)	35.9%	32.3%
Base, \$100 to \$1,000	135	\$40,006	46 (34%)	32.1%	40.6%
Base, \$1,000 or more	67	\$618,791	33 (49%)	29.2%	29.2%
Base, weekly top 10	40	\$545,058	21 (52%)	25.3%	25.3%
Base, weekly top 50	200	\$658,402	80 (40%)	29.4%	29.9%
Base, all	750	\$674,944	258 (34%)	29.5%	29.9%
BNB Chain, ring wallets	62	\$5,534	62 (100%)	100.0%	100.0%
BNB Chain, others	9,069	\$489,890	1,214 (13%)	13.6%	19.5%
BNB Chain, all	9,131	\$495,424	1,276 (14%)	14.6%	20.4%

Table E2: End-to-end positive control (results/positive_control_e2e.csv): three markets, five ring shares per route; b is each market's loop share measured by the same pipeline with no ring.

Route	Designed distance	Measured (all ring payers)	Expected loop share	Runs	Largest error	Flag as predicted
seller to payer	hop 1	hop 1	$\varphi + (1 - \varphi) \cdot b$	15	1.1e-16	15 of 15
two relays	hop 3	hop 3	$\varphi + (1 - \varphi) \cdot b$	15	1.1e-16	15 of 15
three relays	hop 4	hop 4	$\varphi + (1 - \varphi) \cdot b$	15	1.1e-16	15 of 15
four relays	hop 5	hop 5	$(1 - \varphi) \cdot b$	15	5.6e-17	15 of 15
via a hub, two transactions	not reached	not reached	$(1 - \varphi) \cdot b$	15	2.8e-17	15 of 15
forwarded by the hub, one transaction	hop 1	hop 1	$\varphi + (1 - \varphi) \cdot b$	15	5.6e-17	15 of 15

Table E3: The 40 spot-check payees after unblinding (results/spotcheck_unblinded.csv, which also holds every reasoning). Judge columns: C coordinated, I independent, U unclear, then the judge's estimated coordinated share.

Sample	Payee	Week	Rule	Loop share	Outside funding	Judge A	Judge B
S01	0x4ad9fcc1	W2	no	0.00	1.00	I 0.00	I 0.00
S02	0x6bab1d63	W3	flag	1.00	0.95	I 0.00	I 0.00
S03	0x0bd02144	W2	flag	1.00	0.55	I 0.00	I 0.00
S04	0x7f629403	W2	no	0.00	1.00	I 0.00	I 0.00
S05	0x1c656644	W2	flag	1.00	0.76	C 1.00	C 1.00
S06	0x836be278	W4	no	0.00	1.00	U	U
S07	0xbc5bafda	W3	no	0.00	1.00	C 1.00	C 1.00
S08	0x837c40e2	W3	flag	0.90	0.54	C 0.70	C 0.70
S09	0xd593832c	W3	flag	0.96	0.91	C 0.95	C 0.95
S10	0x85af15f3	W1	flag	1.00	0.00	C 1.00	C 1.00
S11	0x768d3897	W1	flag	1.00	0.00	C 1.00	C 1.00
S12	0xb7c76e6a	W1	no	0.00	1.00	I 0.00	I 0.00
S13	0xe6dab093	W1	flag	1.00	0.00	C 1.00	C 1.00
S14	0xd8ba61a0	W4	flag	1.00	0.27	C 1.00	C 1.00
S15	0xce668a6e	W1	flag	1.00	0.00	C 1.00	C 1.00
S16	0x17d38086	W1	flag	1.00	0.00	C 1.00	C 1.00
S17	0xaf6ee95c	W4	no	0.00	0.00	C 1.00	C 1.00
S18	0x0a84297d	W3	flag	1.00	0.00	C 1.00	C 1.00
S19	0x33e59490	W1	no	0.00	1.00	I 0.00	I 0.00
S20	0x788b4ca1	W1	flag	1.00	0.90	C 0.97	C 0.98
S21	0x4df66b6c	W2	no	0.00	1.00	I 0.00	I 0.00
S22	0x4cd50150	W1	no	0.00	1.00	I 0.00	I 0.00
S23	0xc34f3d05	W1	no	0.00	1.00	I 0.00	I 0.00
S24	0x07e079b1	W1	no	0.00	1.00	I 0.00	I 0.00
S25	0x430a232c	W3	flag	1.00	0.00	C 1.00	C 1.00
S26	0xf1cbc619	W1	no	0.00	1.00	I 0.00	I 0.00
S27	0x8060f6e3	W3	flag	1.00	0.00	C 1.00	C 1.00
S28	0x434ca49f	W1	flag	1.00	0.00	C 1.00	C 1.00
S29	0x3a876a77	W3	flag	1.00	0.00	C 1.00	C 1.00
S30	0x91410228	W1	flag	0.69	0.48	C 0.69	C 0.69
S31	0x96368628	W4	flag	0.92	0.23	C 0.90	C 0.92
S32	0x8aab9a6d	W3	flag	0.94	0.84	C 0.94	C 0.94
S33	0x6454a645	W4	flag	1.00	0.12	C 0.95	C 0.94
S34	0xf3ecfe1d	W3	flag	1.00	0.99	C 1.00	U
S35	0x0748cc42	W3	no	0.00	1.00	C 1.00	C 1.00
S36	0xeb931657	W2	no	0.00	1.00	I 0.00	I 0.00
S37	0x4f180b77	W1	flag	1.00	0.86	I 0.00	I 0.00
S38	0x5976376b	W3	flag	0.88	0.99	C 0.88	U
S39	0xb13b783a	W4	no	0.00	0.00	C 1.00	C 1.00
S40	0x0440383b	W3	flag	0.91	0.86	C 0.50	C 0.50