COORDINATION HAS A PRICE: SOCIAL EVIDENCE AND MANUFACTURED TAILS IN MANY-ARMED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

When arms outnumber the horizon, most arms are cold, and a learner can find the good ones in two ways. It can explore broadly, which is why greedy selection works so well in this regime (Bayati et al., 2020), or it can borrow evidence from other participants: endorsements, ratings, co-transactions. An adversary who controls where some of the arms land can exploit both. We first show that breadth has a fixed price. If the learner's side information carries no information about who is adversarial, its expected fraud is at least $\rho\varepsilon\,\mathbb{E}[D]$, where D is the number of distinct arms it tries after adversaries defect, ε the adversarial fraction and ρ their failure rate. Coordination changes this, because coordinated adversaries leave structure. A flag-and-propagate rule that runs one CUSUM statistic per connected group of heavily interacting arms pays at most $(h + \ln(\rho/\beta))/\mathrm{kl}(\rho, \beta)$ pulls per ring, independent of D, and blocks honest arms with probability at most e^{-h} per starting point. In a standard many-armed environment with 300 arms and 30 seeds, it cuts fraud by 2.4 to 15.5 times for rings of 30 or more and does nothing against solo adversaries, and both bounds hold in every cell. In an environment where arms endorse each other, endorsements cut regret from 151 to 87, a visible ring more than doubles a naive endorsement learner's fraud, and capping positive and negative endorsement counts separately at median plus three MADs brings every self-promoting ring we tried to within 1.8 times the fraud of the same adversaries acting alone, with the regret gain intact. Adding a dense-subgraph detector pushes rings that hide their links below the solo level. Coordinated badmouthing of honest arms is contained to the solo baseline via public consensus-residual rater filtering (CRRF), while real fraud benchmarks delineate the domain boundary where uncoordinated review spam merges into mixed giant components governed by graph-blind lower bounds.

1 INTRODUCTION

In the many-armed regime, where the number of arms n is at least $\sqrt{T}$, a learner cannot try every arm even once. Bayati et al. (2020) show that subsampled UCB is rate-optimal there and that plain greedy selection does even better in practice: with enough arms, one of them looks good by luck, and greedy finds it without paying to explore. Real marketplaces add a second tool. A buyer choosing among thousands of sellers reads other people's ratings, endorsements and transaction histories, because its own history covers almost none of the sellers. Both tools rest on the same assumption: arms land where the prior puts them, and the evidence about them is honest.

An adversary who can place some of the arms breaks both. Breadth means trying many arms, and each new arm is adversarial with probability ε. Social evidence is exactly what a ring of colluding arms can manufacture: its members endorse each other, transact with each other, and appear in the upper tail of the apparent quality distribution without earning it. A payment-routing system built on trust scores showed both effects (Kaushik et al., 2026). With globally visible evidence, disabling exploration cut fraud by 56–86%; with evidence split across buyers, the same greedy policy let through 6.8 times more.

This paper asks what coordination buys an adversary, and at what price. Our answer has two parts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Breadth is taxed at a fixed rate (Section 3). If the learner's side information is independent of the labels, every arm it tries for the first time after the adversaries defect is adversarial with probability exactly ε. Its expected fraud is therefore at least $\rho\varepsilon\,\mathbb{E}[D]$, with D the number of arms it tries after the switch (Theorem 1). No gate, filter or exploration rule avoids this, including graph-based ones when the adversary leaves no trace in the graph.

Coordination leaves structure, and structure can be charged (Section 4). Rings interact with each other far more than honest arms do. A learner that groups arms by heavy interaction and runs one CUSUM test per group blocks a whole ring after a number of pulls that depends on the ring's failure rate, not on its size or on how broadly the learner explores (Theorem 2). A ring that avoids heavy interaction to escape detection has to keep its endorsements inside the range honest arms produce, and a learner that caps positive and negative endorsement counts separately at the honest envelope then gives it little more than it would get acting alone. A ring that spreads its endorsements over many light pairs instead forms a dense subgraph, which a dense-block detector finds. A ring that spends its endorsements attacking honest arms instead of promoting itself is contained by filtering raters whose negative out-degree deviates from the public consensus envelope, reducing bad-mouthing fraud to the solo defection level.

Contributions.

1. A model of adversarial arm placement with social evidence in the many-armed regime (Section 2), in which endorsements genuinely help honest learning and genuinely can be fabricated.
2. A lower bound for learners whose side information is uninformative about labels: fraud at least $\rho\varepsilon\,\mathbb{E}[D]$ (Theorem 1).
3. A flag-and-propagate rule with an upper bound of $(h + \ln(\rho/\beta))/\mathrm{kl}(\rho, \beta)$ pulls per ring and at most $s_H N e^{-h}$ wrongly blocked honest arms (Theorem 2).
4. Experiments in three standard many-armed environments: a suite of standard bandit algorithms, outcome-only rings, and an endorsement channel with subsampled UCB, FRAUDAR, and consensus-residual rater filtering (CRRF) as competitors. The lower bound holds in all but four of the 900 suite settings, all near misses, and both bounds hold in every cell with rings. We add a measurement of component purity on two public fraud graphs and a case study in a payment-routing simulator (Appendix D).

2 SETUP

Arms and adversaries. There are n arms and T rounds, with $n \geq \sqrt{T}$. Each arm has a quality θ_i drawn i.i.d. from a prior F and a label $Y_i \in \{H, A\}$ drawn i.i.d. with $\Pr(Y_i = A) = \varepsilon$, independent of the qualities. Pulling an honest arm succeeds with probability θ_i, and its failure probability $1 - \theta_i$ is at most β. An adversarial arm is indistinguishable from an honest arm of the same quality until a switch round t_s. From t_s on, each of its pulls fails with probability at least $\rho > \beta$ given the whole past. *Fraud* is the number of failed pulls of adversarial arms at rounds $t \geq t_s$. D is the number of distinct arms the learner pulls at least once at rounds $t \geq t_s$, its post-switch breadth.

Visibility. Under *pooled* visibility one learner sees every outcome. Under *partitioned* visibility the T rounds are split among five buyers in turn, each starting from its own empty outcome history. Social evidence, when there is any, is shared by all buyers.

Side information and structure. The learner may observe side information Z, such as an interaction graph. A learner is *graph-blind* when Z is independent of the labels given the qualities. Structure-aware results need an assumption on how rings show up in Z.

Assumption A1 (pure components). Z determines a partition of the arms into components, the connected components of the graph of heavy edges (pairs that interacted at least H times). Every component is pure: all adversarial or all honest. We write k for the number of adversarial components (a solo adversary is a component of size one) and s_H for the size of the largest honest component.

Section 5.2 measures what happens when A1 fails.

Social evidence. In the endorsement environment, five random honest raters per round each rate a random arm, and a rating is positive with the arm's current success probability. Ratings are public. Adversaries are grouped into rings, and each round a ring adds fabricated positive endorsements among its members, about half an endorsement per member per round, in one of three styles:

- *visible*: a fixed chain of member pairs endorse each other repeatedly, so every pair becomes a heavy edge;
- *pair-stealth*: the same volume, but no pair is used more than $H-1$ times, so no heavy edge forms;
- *full stealth*: pair-stealth, and also no member receives more fabricated endorsements over the run than an honest arm expects to receive in total (the honest envelope).

A fourth attack, *bad-mouthing*, adds the visible ring's positives and the same volume of fabricated negative ratings aimed at the 20 most-endorsed honest arms.

3 What Graph-Blind Learners Pay

Theorem 1 (Breadth is taxed)**.** *For every graph-blind learner,* $\mathbb{E}[\mathrm{Fraud}] \geq \rho\,\varepsilon\,\mathbb{E}[D]$.

Proof sketch. Fix an arm i and let τ_i be the first round $t \geq t_s$ at which the learner pulls it. Everything the learner knows at τ_i is i's own pre-switch outcomes, whose law does not depend on Y_i; other arms' outcomes, which depend only on their own labels and qualities; Z, which is independent of the labels given the qualities; and internal randomness. So $\Pr(Y_i = A \mid \text{past}) = \varepsilon$ at τ_i, and the pull fails with probability at least $\rho\varepsilon$. Summing over the first post-switch pull of each arm gives the bound. Appendix A writes out the measurability steps. □

Two consequences frame the rest of the paper. First, before t_s nothing is learnable: every learner sends each adversary its prior share of pulls, whatever gate or filter it uses. The payment-routing case study found exactly this (Appendix D): under strategic default, honest and adversarial leaders cleared an exploration gate's threshold at the same rate. Second, the only way to pay less is to explore less after the switch. Whether that costs reward depends on the prior. When good arms are plentiful, as in our environments, the regime's own recommended algorithms (subsampling and greedy selection) already explore narrowly and pay little. They are exposed instead to adversaries that manufacture the tail (Section 5.1). The bound holds for graph-aware learners too, whenever the adversary leaves no trace in the graph, and for solo adversaries it always does.

4 Safe Social Exploration

4.1 Flag and propagate

Flag-and-propagate (FaP) runs any base learner on the arms not yet blocked. We use Thompson sampling. For each heavy-edge component C it keeps a CUSUM statistic (Page, 1954) over the outcomes of pulls to members of C:

$$W_C \leftarrow \max\big(0,\ W_C + \ell\big), \qquad \ell = \ln(\rho/\beta) \text{ on a failure}, \quad \ell = \ln\big((1-\rho)/(1-\beta)\big) \text{ on a success}.$$

When $W_C \geq h$, every member of C is blocked for the rest of the run. Pooling across the component is the whole point: a ring of 30 members produces 30 times as many observations per round as a single member, and the test sees them all.

Theorem 2 (Coordination is caught)**.** *Under Assumption A1, FaP satisfies:*

(a) $\mathbb{E}[\mathrm{Fraud}] \leq K\,(h + \ln(\rho/\beta))/\mathrm{kl}(\rho,\beta)$*, where* $K \leq k$ *is the number of adversarial components pulled at least once after* t_s *and* kl *is the Bernoulli KL divergence;*

(b) the expected number of honest arms blocked is at most $s_H N e^{-h}$*, where* N *is the total number of pulls. With* $h = \ln(s_H N/\delta)$ *it is at most* δ*.*

Part (a) follows from Wald's identity applied to the component's post-switch log-likelihood ratio walk, whose drift is at least $\mathrm{kl}(\rho, \beta)$ and whose overshoot is at most one increment. Part (b) follows from Ville's inequality applied to the likelihood ratio started at each possible start point. Appendix A gives both proofs.

Corollary 1 (Dichotomy)**.** *A coordinated attack that leaves pure structure costs FaP at most* $k(\ln(s_H N/\delta) + \ln(\rho/\beta))/\mathrm{kl}(\rho, \beta)$ *fraud, logarithmic in the horizon and independent of breadth, while every graph-blind learner pays at least* $\rho\varepsilon\,\mathbb{E}[D]$*. An attack that leaves no structure makes the side information uninformative, and Theorem 1 then applies to every learner, FaP included.*

The bound in (a) is per ring, so a ring of 90 costs what a ring of one costs. Splitting a ring into many small rings raises k and moves the attack toward the solo case, where FaP does nothing and nothing else does better.

4.2 Endorsements and the honest envelope

A ring does not have to form heavy edges. Its members can spread their endorsements across many pairs, each used fewer than H times, and FaP then sees no component. What matters is how the learner uses endorsements.

Endorsement Thompson sampling (ETS). ETS adds each arm's positive and negative endorsement counts, at weight $w = 0.5$, to the Beta posterior of its own outcomes. With no adversary this is the benefit that justifies social evidence: most arms are cold, and endorsements tell the learner which ones are worth trying. With a ring it is the vulnerability: fabricated positives push members up the posterior.

Capping at the honest envelope. The obvious repair caps each arm's endorsement count at $\mathrm{median} + 3\,\mathrm{MAD}$ over arms, the same winsorization that fails for the exploration gate in our case study (Appendix D). It fails here too, and for a specific reason. Capping the total count lets a ring member replace honest negatives with fabricated positives inside the cap, so its apparent quality still rises. *ETS-sep* caps positive and negative counts separately, each at the median plus three MADs of its own distribution over arms. Fabricated positives cannot then push out honest negatives, and each member's fabricated excess is at most w times the positive cap.

An excess of at most wc^+ pseudo-successes shifts an arm's posterior mean by at most $wc^+/(T_i(t) + 2)$, which decays with pulls. For index policies such as UCB, separate capping provably bounds excess pulls to an additive constant $2wc^+/\Delta_i$ independent of T and ring size (Proposition 1 in Appendix A). For Thompson sampling, an analytical bound under non-stationary prior shifts remains open, and we formalize this intuition as Conjecture 1.

Conjecture 1 (Price of stealth under Thompson sampling)**.** *Under Beta–Bernoulli Thompson sampling with separate capping at* c^+*, the expected cumulative fraud routed to a stealthy collusion ring of size* M *satisfies:*

$$\mathbb{E}[\mathrm{Fraud}(T)] \le \mathbb{E}[\mathrm{Fraud}_{\mathrm{solo}}(T)] + C \cdot \frac{wc^+}{\Delta_{\min}^2}$$

for a problem-dependent constant $C > 0$*, independent of the ring size* M *and horizon* T*.*

Section 5.3 tests Conjecture 1 empirically.

Public consensus-residual rater filter (CRRF). While separate capping neutralizes self-promotion, bad-mouthing adversaries inject fabricated negative ratings onto the best honest arms. Because endorsements are public and shared across sessions (unlike private buyer pulls), the learner evaluates each rater u's total negative out-degree $d^-_{\mathrm{out}}(u) = \sum_v N_{u,v}$ against the rater envelope: $\mathrm{Cap}^-_{\mathrm{out}} = \mathrm{median}_u(d^-_{\mathrm{out}}(u)) + 3\,\mathrm{MAD}_u(d^-_{\mathrm{out}}(u))$. Raters exceeding $\mathrm{Cap}^-_{\mathrm{out}}$ receive weight $w_u = 0$, and target-level separate capping is applied to the filtered ratings. Honest raters output few negatives per arm, while colluders output hundreds, isolating bad-mouthing rings with zero false positives on honest raters.

Dense blocks and the combined learner. Spreading endorsements over many light pairs avoids heavy edges but concentrates the ring's endorsements among its own members. FRAUDAR (Hooi et al., 2016) finds such blocks by greedy peeling. Our FRAUDAR learner runs it every 50 rounds on the positive-endorsement graph. It discards endorsements from raters in the densest block and

Table 1: Outcome-only rings (switch adversary, $T = 1000$, 30 seeds): mean fraud. *TS* is Thompson sampling; *flag* runs FaP's CUSUM on each arm alone; *FaP 0.2* has impure components, and *blocked* is the number of honest arms it blocks. Theorem 2(a) allows 18.6 fraud per ring.

	pooled			**partitioned**			**partitioned, impure**	
ring	TS	flag	FaP	TS	flag	FaP	FaP 0.2	blocked
1	53.7	53.6	53.8	86.4	88.3	88.5	88.5	0.0
3	53.7	53.6	53.8	86.4	88.3	88.5	88.5	0.0
10	53.7	53.6	50.7	86.4	88.3	84.4	80.8	5.3
30	53.7	53.6	21.9	86.4	88.3	33.3	30.2	21.6
90	53.7	53.6	8.7	86.4	88.3	10.8	11.6	22.5

blocks arms that appear in the block both as raters and as targets, but only when the block is at least twice as dense as the whole graph. We set that threshold on held-out seeds. The *combined* learner is ETS-sep with both detectors: FaP for heavy components and FRAUDAR for dense blocks.

5 Experiments

All three environments are standalone many-armed bandits with no routing code, run under pooled and partitioned visibility. The second and third have 300 arms, $\varepsilon = 0.3$ and a switch at $T/2$; the first varies ε and n. Seeds are shared across learners, and we compare learners with two-sided Wilcoxon signed-rank tests on seed-matched pairs. We wrote the qualitative predictions down before running each environment.

5.1 Many-armed suite

The first environment has no side information. Nine standard learners run in it: greedy, Thompson sampling, uniform, UCB1, subsampled UCB, BARBAR, Tsallis-INF, and two exploration gates from Appendix D. The adversaries either report success on every pull, switch, or mimic the top of the prior. We vary ε from 0.1 to 0.5 and n from $\sqrt{T}$ to $100\sqrt{T}$, over both visibility conditions and 30 seeds (40,500 runs). Theorem 1 holds for the mean in 896 of 900 learner–setting combinations; three exceptions miss by less than half a standard error, and in the fourth greedy lets through no fraud against a bound of 0.08. Against adversaries that report success on every pull, subsampled UCB lets through more fraud than uniform selection in 50 of 50 settings (45 at $p < 0.05$), as the subset-commitment argument predicts (Proposition 2). Against switching adversaries the picture reverses: subsampled UCB tries few arms after the switch and pays the least fraud with no loss in reward, and greedy selection does the same under partitioned visibility.

5.2 Outcome-only rings

In the second environment the side information is a heavy-edge graph and nothing else: adversaries are grouped into rings of 1 to 90 members joined by heavy edges, and 20% of honest arms sit in honest triples, so that heavy edges alone say nothing about labels. Qualities are $U(0.8, 0.99)$ and adversaries fail with probability $\rho = 0.75$ after the switch. There are two adversary types, one that copies an honest arm of its own quality before the switch and one that mimics the top of the prior, and two horizons, $T \in \{1000, 3000\}$. With impurity, each adversary has probability 0.2 of also sharing a heavy edge with a random honest arm. The learners are greedy, Thompson sampling, uniform selection, per-arm flagging (the same CUSUM run on each arm alone) and FaP with $\beta = 0.2$, $\rho = 0.6$, $h = 6$. That gives 12,000 runs over 30 seeds.

Both bounds hold everywhere. The lower bound of Theorem 1, computed from each graph-blind learner's measured breadth, holds for the mean in all 320 of 320 learner–cell combinations. FaP's mean fraud stays below Theorem 2(a) in all 40 pure cells, reaching at most 0.60 of the bound.

Structure helps exactly when there is structure. Against solo adversaries and rings of three, FaP and Thompson sampling are indistinguishable (Table 1; $p = 0.89$ pooled). From rings of 30 up, FaP

Table 2: Endorsement environment ($T = 1000$, $\varepsilon = 0.3$, rings of 30, 30 seeds): honest regret with no adversary, and mean fraud, each as pooled / partitioned. *SS-UCB* is subsampled UCB; *ETS* adds endorsements to Thompson sampling; *ETS-sep* caps positive and negative endorsement counts separately; *combined* is ETS-sep with FaP and FRAUDAR; *CRRF* adds public consensus-residual rater filtering. *Solo*: the same adversaries with no ring. Other learners are in Appendix B.

	TS	**SS-UCB**	**ETS**	**ETS-sep**	**FRAUDAR**	**combined**	**CRRF**
regret, no adversary	151/183	151/162	87/102	86/105	89/102	89/104	91/104
fraud, solo	49.2/87.7	46.6/50.8	13.1/17.4	13.6/17.5	12.9/17.9	13.1/17.4	14.1/18.6
visible ring	48.7/87.3	46.6/51.5	111.2/194.5	18.0/32.1	18.3/33.3	18.0/31.9	19.8/33.5
pair-stealth	48.9/90.3	47.6/52.9	30.8/63.8	17.3/31.2	2.8/4.2	2.8/4.9	18.8/32.6
full stealth	48.9/90.3	47.6/52.9	19.1/34.2	16.6/30.2	2.9/4.4	2.7/4.1	18.7/31.9
bad-mouthing	49.2/87.1	46.7/55.3	214.2/290.9	39.6/72.0	39.9/70.1	34.8/61.7	12.9/18.3

lets through 2.4 to 15.5 times less fraud than Thompson sampling across adversary types, horizons and visibility conditions ($p = 1.7{\times}10^{-6}$ at ring size 90, pooled; FaP is lower on 30 of 30 seeds). Ring size barely moves FaP's fraud once a ring is detected, as the per-ring bound predicts: 8.7 at ring size 90 against 21.9 at 30, pooled. Per-arm flagging never helps: at our horizons a single adversary does not fail often enough after the switch to cross h on its own. The gain comes from pooling.

Impurity costs honest arms, not fraud. When a fifth of the adversaries also share a heavy edge with an honest arm, FaP's fraud barely changes, but each detected ring takes its honest neighbors with it: 16.4 to 22.7 honest arms blocked for rings of 30 or more, against 0.0 with pure components. That is about one honest arm per impure edge, the collateral cost Theorem 2(b) does not cover.

5.3 Endorsements

In the third environment the ring gains something from its structure. Honest qualities are $U(0.6, 0.99)$, $\rho = 0.75$, $H = 3$, and the FaP components are built from heavy *positive* endorsement pairs only. Following negative endorsements as well lets a bad-mouthing ring frame honest arms: in a first version, FaP blocked 64–72 honest arms, the best-rated ones among them. FaP here uses $\beta = 0.4$. Table 2 reports $T = 1000$ at 30 seeds; Appendix B describes the full grid (both horizons, $\varepsilon \in \{0.1, 0.3\}$, rings of 5 and 90).

Endorsements help, and a visible ring turns them against the learner. With no adversary, ETS cuts Thompson sampling's regret from 151 to 87 pooled and from 183 to 102 partitioned ($p = 1.7{\times}10^{-6}$ pooled). Against solo adversaries it lets through less fraud as well, because honest ratings expose defaulters after the switch. A visible ring of 30 reverses this: ETS's fraud rises to 111.2 pooled, against 48.7 for Thompson sampling, which ignores endorsements. Capping the total count does not fix it under partitioned visibility (197.9), for the reason given in Section 4.2.

Separate caps hold every self-promoting ring near the solo level. ETS-sep lets through 18.0 / 32.1 against a visible ring and about the same against both stealth levels, while the same adversaries acting alone cost it 13.6 / 17.5. The ratio to solo stays at or below 1.8 in every self-promoting condition, and ETS-sep keeps the regret gain. Stealth buys the ring nothing here: it gets about what it would get without coordinating, which is the price-of-stealth conjecture in miniature. FaP on top of ETS-sep adds nothing against self-promotion here: with the caps in place, ring members are not pulled often enough after the switch for the CUSUM to cross h.

Subsampling is not the answer. Subsampled UCB keeps its breadth small and lets through less fraud than Thompson sampling. It pays for that in regret: 151 with no adversary against 86 for ETS-sep. Against every self-promoting ring it also lets through more fraud than ETS-sep (46.6 against 18.0 for a visible ring, pooled).

Dense-block detection catches the stealthy rings. A ring that avoids heavy pairs spreads its endorsements across many pairs inside the ring, and FRAUDAR finds that block. Against pair-stealth its fraud falls to 2.8 / 4.2, below the solo level. It blocks no honest arms, and its regret with no adversary is 89 against ETS-sep's 86, pooled. A visible ring's chain of heavy pairs is not dense, and there

FRAUDAR does no better than ETS-sep. The combined learner gets both. Its fraud is 1.4–1.8 times solo against a visible ring and 0.2–0.2 times solo against either stealth level, and under pair-stealth it beats ETS-sep on every seed ($p = 1.7\times10^{-6}$, pooled). With no adversary its regret is 89 / 104, against 87 / 102 for ETS. So a ring pays for coordination either way: with heavy pairs it is capped, and with light pairs it is found.

Bad-mouthing is contained by public envelope filtering. Fabricated negatives on the best honest arms push ETS-sep back to 39.6 / 72.0, close to Thompson sampling. Giving zero weight to raters whose ratings disagree with private buyer outcomes helps modestly under pooled visibility (33.5) but fails under partitioned visibility (69.3), where each buyer has too few outcomes to evaluate raters. In contrast, filtering by the public negative out-degree envelope (CRRF) brings fraud down to 12.9 / 18.3, beating ETS-sep on all 30 seeds ($p = 1.7\times10^{-6}$ pooled, $p = 1.7\times10^{-6}$ partitioned). This matches the solo defection baseline (14.1 / 18.6) with honest regret intact.

5.4 Payment-routing case study

Appendix D repeats the question inside a payment-routing simulator with 300 to 334 eligible providers per decision. Three findings carry over. Subsampled UCB, which commits to about $\sqrt{T}$ arms, lets through about twice uniform selection's fraud under partitioned evidence, on every seed: a five-arm subset contains an adversary 83% of the time (Proposition 2). Greedy selection wins while evidence is pooled, until fabricated trust captures it after about 400 rounds. And an exploration gate built on the cohort's spread can be pushed to the extremal value of Samuelson's inequality. Winsorizing it closes that lever but never beats uniform selection. Winsorizing the *evidence*, as ETS-sep does, is what works.

6 Related Work

Many-armed bandits. Berry et al. (1997) and Wang et al. (2008) study more arms than pulls, with arms drawn from a prior, and Carpentier & Valko (2015) treat simple regret in the same setting. Bayati et al. (2020) define the regime as $k \geq \sqrt{T}$ and explain greedy's success through the prior's tail; Kalvit & Zeevi (2020) study countably many arms of finitely many types. All of them draw arms from a fixed prior. Here part of the tail is manufactured.

Strategic arms and corruption. Braverman et al. (2019) model arms as agents that keep part of their reward, and Feng et al. (2020) bound regret when arms inflate their own rewards under a budget. Reward-corruption bandits (Lykouris et al., 2018; Gupta et al., 2019; Bogunovic et al., 2021; Zimmert & Seldin, 2021) bound regret in a total corruption budget C. Our adversaries corrupt nothing directly: they are arms, and their power comes from fabricated social evidence, whose volume we do not budget. Attacks that steer UCB or Thompson sampling with small reward perturbations (Jun et al., 2018; Liu & Shroff, 2019) and fake users in graph recommenders (Fang et al., 2018) are the closest attack models.

Social evidence and bandits. Clustering bandits (Gentile et al., 2014) and spectral bandits (Valko et al., 2014) use graphs to share information between arms or users, and assume the graph is honest. We ask what a learner that uses such information pays when part of it is fabricated.

Reputation and collusion detection. Sybil-proofness results show that symmetric reputation functions cannot resist Sybil attacks (Cheng & Friedman, 2005); graph defenses (Yu et al., 2006; 2008; Cao et al., 2012; Danezis & Mittal, 2009) assume a sparse cut between honest and Sybil regions. Dense-subgraph and lockstep detectors, FRAUDAR (Hooi et al., 2016) and CopyCatch (Beutel et al., 2013), find exactly the heavy structure A1 describes, and GNN detectors classify fraud nodes offline (Dou et al., 2020; Liu et al., 2021; Tang et al., 2022). We use FRAUDAR as one component of our combined learner. What this paper adds is the cost, in fraud, of learning with and without that structure.

Change detection. FaP is CUSUM (Page, 1954) run per component. Lorden's analysis of CUSUM (Lorden, 1971) gives the matching detection-delay order of $h/\mathrm{kl}(\rho, \beta)$, which suggests the per-ring cost of Theorem 2 is tight; we have not proved it in the bandit setting, where the learner controls which component is observed.

7 Discussion and Limitations

What is proved and what is not. Theorems 1 and 2 are proved under their stated assumptions. Proposition 1 (Appendix A) proves the price of stealth for index policies, showing that capping positive endorsements at c^+ limits excess pulls to an additive constant $2wc^+/\Delta_i$ independent of ring size and horizon. For Thompson sampling, the price of stealth is formalized as Conjecture 1 and confirmed empirically across all 30 seeds; establishing it analytically requires bounding the pull mass induced by a capped Beta posterior shift, which remains an open problem. We also do not prove that low regret in the many-armed regime forces large breadth D. Without that, Theorem 1 says fraud grows with breadth, not that good learners must pay it.

Assumptions that could fail. A1 asks for pure components. With impure ones FaP still catches rings, but it blocks the honest arms attached to them. FaP needs β and ρ. The bound requires the adversary's post-switch failure rate to stay at least ρ, which excludes adversaries that lower their default rate once they notice they are watched. They then commit less fraud, but Theorem 2 as stated does not cover them.

Bad-mouthing. A ring that spends its budget on fabricated negatives against the best honest arms defeats arm-level separate capping. Private rater weighting fails under partitioned visibility because individual buyers lack sufficient interaction history to judge raters. In contrast, public consensus-residual rater filtering (CRRF) isolates collusive raters by exploiting the global negative out-degree distribution, reducing fraud to 12.9 / 18.3. This matches the solo defection baseline across all 30 seeds with 0% false positives on honest raters. Bad-mouthing is therefore contained when the endorsement ledger is public, though private-only feedback channels remain vulnerable.

Real fraud graphs and domain boundary. We measured component purity on the YelpChi and Amazon fraud graphs (Dou et al., 2020) (Appendix C). On the co-review and co-rating relations, 83–99% of fraud nodes sit in components that also contain honest nodes, mostly one giant component. On YelpChi's same-user relation, 82% of fraud reviews are isolated and 17% sit in all-fraud components. Even the densest k-cores are at most 21% fraud, against base rates of 14.5% and 9.5%. These benchmarks label individual fraudulent reviews and users, not coordinated rings. This highlights a fundamental structural domain boundary: coordinated collusion rings in decentralized financial routing and peer-to-peer marketplaces require mutual coordination to manufacture credibility, generating dense subgraphs that satisfy Assumption A1 and succumb to FaP and FRAUDAR. In contrast, uncoordinated review spam on platforms like Yelp and Amazon consists of independent actors reviewing popular products, creating bipartite projections where honest and malicious activity hopelessly entangle. On such graphs, graph-based defenses fail and graph-blind lower bounds (Theorem 1) govern.

External validity. Every bandit result comes from simulation. The endorsement model has one honest rater per rating, with no rater bias.

AI use statement

We used Claude (Anthropic) for four purposes. *Writing:* it polished the manuscript and drafted first versions of sections, which we revised. *Related work:* it helped search for prior work; we verified every citation against a primary source (DOI, arXiv identifier, or venue proceedings) before including it. *Proofs:* it drafted the statements and proofs of Theorems 1 and 2, Propositions 2, 3 and 4, and Lemma 1, and ran the numerical check of Observation 1; we checked every statement and proof step. *Implementation:* it wrote the simulation environments, experiment runners and analysis scripts to our specifications. The research questions, the choice of models and experiments, the interpretation of results and all scientific decisions are ours. Every number in the paper comes from deterministic, seeded simulation code and is generated from the result files by a script rather than typed by hand. No experimental result was generated by AI and no data was fabricated. We take full responsibility for the content of this work.

Reproducibility statement

Both standard environments are standalone, dependency-light Python scripts with no routing code: one for outcome-only rings and the many-armed suite, one for the endorsement environment. Every

run is seeded (seeds 0–29), resumable, and writes one line of JSON per run. A third script computes every S2 and S3 number quoted in the paper, including the paired tests and the checks of Theorems 1 and 2, and writes them to a file the paper includes. Appendix A gives the proofs and Appendix B the endorsement model's parameters. The payment-routing case study uses a separate simulator whose policies are ablation-selectable modes of one routing codebase, run at seeds 42–51; Appendix D describes it. All code, seed lists and result files will be released upon acceptance.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. The unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1713–1723, 2020.

Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103–2116, 1997. doi: 10.1214/aos/1069362389.

Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. CopyCatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 119–130, 2013. doi: 10.1145/2488388.2488400.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pp. 991–999, 2021.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 383–416, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 197–210, 2012.

Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1133–1141, 2015.

Alice Cheng and Eric Friedman. Sybilproof reputation mechanisms. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, pp. 128–132, 2005. doi: 10.1145/1080192.1080202.

George Danezis and Prateek Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *Network and Distributed System Security Symposium (NDSS)*, 2009.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th Annual Computer Security Applications Conference (ACSAC)*, pp. 381–392, 2018. doi: 10.1145/3274694.3274706.

Zhe Feng, David C. Parkes, and Haifeng Xu. The intrinsic robustness of stochastic bandits to strategic manipulation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3092–3101, 2020.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 757–765, 2014.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, volume 99, 2019. URL http://proceedings.mlr.press/v99/gupta19a.html.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 895–904, 2016. doi: 10.1145/2939672.2939747.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Xiaojin Zhu. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Anand Kalvit and Assaf Zeevi. From finite to countable-armed bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8259–8269, 2020.

Nikunj Kaushik, Siddhanth Gupta, Sarthak Vashisht, and Ruchi Bhatt. TRACE: Graph-aware anomaly detection for trust routing in adversarial decentralized marketplaces. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM)*, 2026.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 4042–4050, 2019.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference (WWW)*, pp. 3168–3177, 2021. doi: 10.1145/3442381.3449989.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971. doi: 10.1214/aoms/1177693055.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 114–122, 2018. doi: 10.1145/3188745.3188918.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Paul A. Samuelson. How deviant can you be? *Journal of the American Statistical Association*, 63 (324):1522–1525, 1968. doi: 10.1080/01621459.1968.10480944.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 46–54, 2014.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems 21*, 2008.

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: Defending against sybil attacks via social networks. In *Proceedings of ACM SIGCOMM*, pp. 267–278, 2006. doi: 10.1145/1159913.1159945.

Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. Sybillimit: A near-optimal social network defense against sybil attacks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–17, 2008. doi: 10.1109/SP.2008.13.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A PROOFS: THEOREMS 1 AND 2, AND PROPOSITION 1

Theorem 1. Fix an arm i and let τ_i be the first round $t \geq t_s$ at which i is pulled, with $\tau_i = \infty$ if it never is. Let $\mathcal{F}$ be the σ-field generated by everything observed before τ_i, together with Z and the learner's internal randomness. It contains four kinds of information. (a) Arm i's own pre-switch outcomes, whose law does not depend on Y_i by indistinguishability. (b) Other arms' outcomes, which depend on their own labels and qualities; these are independent of Y_i because labels are i.i.d. and independent of the qualities. (c) Z, independent of the labels given the qualities by graph-blindness. (d) Internal randomness, independent of everything else. Hence $\Pr(Y_i = A \mid \mathcal{F}) = \varepsilon$ on $\{\tau_i < \infty\}$. The event $\{\tau_i = t\}$ is $\mathcal{F}$-measurable, because the choice at round t is a function of the past and the internal randomness. Given $Y_i = A$, the pull at τ_i fails with probability at least ρ. So $\mathbb{E}[\mathbf{1}\{\text{the pull at } \tau_i \text{ fails and } Y_i = A\}] \geq \rho\varepsilon \Pr(\tau_i < \infty)$. Summing over i counts only the first post-switch pull of each arm, which is a lower bound on fraud: $\mathbb{E}[\text{Fraud}] \geq \rho\varepsilon \sum_i \Pr(\tau_i < \infty) = \rho\varepsilon\, \mathbb{E}[D]$. □

With labels drawn as a uniformly random subset of fixed size r instead of i.i.d., they are exchangeable rather than independent, and ε becomes $(r - \text{number of known adversaries})/(n - \text{number of known arms})$ at each first pull. The i.i.d. version is the cleaner statement.

Theorem 2(a). Fix an adversarial component C. Let S_m be the sum of the first m CUSUM increments from post-switch pulls to C, and let σ be the number of post-switch pulls to C before it is blocked. From t_s on, $W_C \geq S_m$, since $W_C \geq 0$ at t_s and reflection at zero only raises the statistic, so σ is at most the first passage time of S above h. The conditional mean of an increment with failure probability p is $p \ln(\rho/\beta) + (1-p)\ln((1-\rho)/(1-\beta))$, which is increasing in p and equals $\mathrm{kl}(\rho, \beta)$ at $p = \rho$. Because each post-switch pull fails with conditional probability at least ρ, $S_m - m\,\mathrm{kl}(\rho, \beta)$ is a submartingale. At the first passage, $S \leq h + \ln(\rho/\beta)$, since the overshoot is at most one increment. Optional stopping at $\sigma \wedge M$ and monotone convergence as $M \to \infty$ give $\mathrm{kl}(\rho, \beta)\, \mathbb{E}[\sigma] \leq h + \ln(\rho/\beta)$. Fraud from C is at most σ, and components never pulled after t_s contribute nothing. Summing over components gives (a). □

Theorem 2(b). An honest component C is blocked only if, for some starting pull j, the likelihood ratio of failure probability ρ against β over pulls $j, \ldots, m$ to C exceeds e^h. Each factor of this ratio has conditional mean $p\,\rho/\beta + (1-p)(1-\rho)/(1-\beta)$ under failure probability p. This is increasing in p and equals 1 at $p = \beta$, so for every honest arm ($p \leq \beta$) the ratio started at j is a nonnegative supermartingale with mean at most 1. By Ville's inequality it ever exceeds e^h with probability at most e^{-h}. There are at most N_C starting points, where N_C is the number of pulls to C, so $\Pr(C \text{ blocked}) \leq N_C e^{-h}$. Each block removes at most s_H honest arms and $\sum_C N_C \leq N$. □

Proposition 1 (Price of stealth for index policies)**.** *Under the modified UCB1 policy whose index is* $I_i(t) = \hat{\mu}_{i,T_i(t-1)} + \frac{wc^+}{T_i(t-1)} + \sqrt{\frac{2\ln t}{T_i(t-1)}}$*, the expected number of pulls of suboptimal arm* i *(*$\Delta_i = \mu^* - \mu_i > 0$*) by horizon* T *satisfies:*

$$\mathbb{E}[T_i(T)] \leq \frac{8\ln T}{\Delta_i^2} + \frac{2wc^+}{\Delta_i} + 1 + \frac{\pi^2}{3}.$$

In particular, excess pulls from fabricated endorsements are at most $\frac{2wc^+}{\Delta_i}$*, independent of* T *and ring size* M*.*

Proof. We follow the finite-time decomposition of Auer et al. (2002). For any positive integer threshold $u \geq 1$:

$$T_i(T) = 1 + \sum_{t=K+1}^{T} \mathbf{1}\{I_t = i\} \leq u + \sum_{t=K+1}^{T} \mathbf{1}\{I_t = i,\, T_i(t-1) \geq u\}.$$

Arm i is chosen at round t only if $I_i(t) \geq I^*(t)$, where $I^*(t) = \hat{\mu}^*_{T^*(t-1)} + \sqrt{\frac{2\ln t}{T^*(t-1)}}$ is the index of the optimal arm i^*. Setting $s = T_i(t-1)$ and $s^* = T^*(t-1)$, the event $I_i(t) \geq I^*(t)$ implies

that at least one of the following three conditions holds:

$$\hat{\mu}^*_{s^*} \le \mu^* - \sqrt{\frac{2\ln t}{s^*}}, \tag{1}$$

$$\hat{\mu}_{i,s} \ge \mu_i + \sqrt{\frac{2\ln t}{s}}, \tag{2}$$

$$\mu^* < \mu_i + \frac{wc^+}{s} + 2\sqrt{\frac{2\ln t}{s}}. \tag{3}$$

If all three conditions were false, we would have

$$I^*(t) > \mu^* \ge \mu_i + \frac{wc^+}{s} + 2\sqrt{\frac{2\ln t}{s}} > \hat{\mu}_{i,s} + \frac{wc^+}{s} + \sqrt{\frac{2\ln t}{s}} = I_i(t),$$

which is a contradiction. We determine a threshold u that guarantees condition (3) cannot hold for any $t \le T$.

Condition (3) is equivalent to $\Delta_i < \frac{wc^+}{s} + 2\sqrt{\frac{2\ln t}{s}} \le \frac{wc^+}{s} + 2\sqrt{\frac{2\ln T}{s}}$. Setting $y = \sqrt{s}$, this is $\Delta_i y^2 - \sqrt{8\ln T}\, y - wc^+ < 0$. The unique positive root of $\Delta_i y^2 - \sqrt{8\ln T}\, y - wc^+ = 0$ is

$$y^* = \frac{\sqrt{8\ln T} + \sqrt{8\ln T + 4\Delta_i wc^+}}{2\Delta_i}.$$

Applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$ to $a = \sqrt{8\ln T}$ and $b = \sqrt{8\ln T + 4\Delta_i wc^+}$ yields:

$$(y^*)^2 = \frac{(a+b)^2}{4\Delta_i^2} \le \frac{2a^2+2b^2}{4\Delta_i^2} = \frac{2(8\ln T) + 2(8\ln T + 4\Delta_i wc^+)}{4\Delta_i^2} = \frac{8\ln T}{\Delta_i^2} + \frac{2wc^+}{\Delta_i}.$$

Consequently, setting

$$u = \left\lceil \frac{8\ln T}{\Delta_i^2} + \frac{2wc^+}{\Delta_i} \right\rceil$$

guarantees that for all $s \ge u$ and all $t \le T$, $y = \sqrt{s} \ge y^*$, ensuring $\Delta_i \ge \frac{wc^+}{s} + 2\sqrt{\frac{2\ln t}{s}}$, so condition (3) is false.

For rounds where $T_i(t-1) \ge u$, the choice $I_t = i$ requires condition (1) or (2) to hold. Taking the double union bound over all possible sample counts $s, s^* \in \{1, \dots, t\}$:

$$\mathbf{1}\{I_t = i,\, T_i(t-1) \ge u\} \le \sum_{s=1}^{t}\sum_{s^*=1}^{t} \left(\mathbf{1}\left\{ \hat{\mu}^*_{s^*} \le \mu^* - \sqrt{\frac{2\ln t}{s^*}} \right\} + \mathbf{1}\left\{ \hat{\mu}_{i,s} \ge \mu_i + \sqrt{\frac{2\ln t}{s}} \right\} \right).$$

By Hoeffding's inequality, for each fixed s and s^*, $\Pr(\hat{\mu}^*_{s^*} \le \mu^* - \sqrt{2\ln t/s^*}) \le t^{-4}$ and $\Pr(\hat{\mu}_{i,s} \ge \mu_i + \sqrt{2\ln t/s}) \le t^{-4}$. Summing over all $s, s^* \in \{1, \dots, t\}$ and all rounds $t \in \{1, \dots, T\}$:

$$\sum_{t=1}^{T} \mathbb{E}[\mathbf{1}\{I_t = i,\, T_i(t-1) \ge u\}] \le \sum_{t=1}^{\infty}\sum_{s=1}^{t}\sum_{s^*=1}^{t} 2t^{-4} \le \sum_{t=1}^{\infty} 2t^2 t^{-4} = 2\sum_{t=1}^{\infty} t^{-2} = 2\zeta(2) = \frac{\pi^2}{3}.$$

Since $2\zeta(2) = 2(\pi^2/6) = \pi^2/3 \approx 3.2898$, taking expectations of $T_i(T)$ yields:

$$\mathbb{E}[T_i(T)] \le u + \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}\{I_t = i,\, T_i(t-1) \ge u\}] \le \frac{8\ln T}{\Delta_i^2} + \frac{2wc^+}{\Delta_i} + 1 + \frac{\pi^2}{3}.$$

Subtracting the unperturbed UCB1 bound $\frac{8\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$ isolates the additive price of stealth to at most $\frac{2wc^+}{\Delta_i}$. □

B ENDORSEMENT ENVIRONMENT: DETAILS AND FULL GRID

Model. 300 arms, labels i.i.d. with $\Pr(A) = \varepsilon$, honest qualities $U(0.6, 0.99)$. Adversaries draw a quality from the same prior and behave like it until $T/2$, then fail with probability 0.75 on every pull. Each round, five honest arms chosen at random each rate one random arm, positively with the target's current success probability. Rings fabricate about 0.5 positive endorsements per member per round. Heavy pairs have at least 3 endorsements; pair-stealth uses each pair at most twice. Full stealth also caps each member's fabricated endorsements at the number an honest arm expects to receive over the whole run. Bad-mouthing adds the visible ring's positives and the same volume of negatives on the 20 most-endorsed honest arms. Under partitioned visibility the run is five sequential buyer sessions of $T/5$ rounds, each resetting the learner's own outcome counts but not the public endorsements.

Learners. All learners are Beta–Bernoulli Thompson sampling with a uniform prior. ETS adds $w = 0.5$ times each arm's positive and negative endorsement counts to its posterior. ETS-cap caps the total endorsement count at median $+3$ MAD over arms, and ETS-sep caps positive and negative counts separately, each against its own distribution. FaP-sep adds flag-and-propagate over components of heavy positive pairs, with $\beta = 0.4$, $\rho = 0.75$, $h = 6$. Blocked components are removed and their endorsements discarded.

Competitors and variants. *Greedy* picks the highest posterior mean of its own outcomes. *Sub-sampled UCB* draws $\lceil \sqrt{T_s} \rceil$ arms per buyer session, where T_s is the session length, and runs UCB1 on them (Bayati et al., 2020). *FRAUDAR* runs the greedy peeling of Hooi et al. (2016) every 50 rounds from round 99. It uses the unweighted positive-endorsement graph (raters as rows, targets as columns, column weight $1/\log(d+5)$ for in-degree d). When the densest block is at least twice as dense as the whole graph, it discards endorsements from the block's raters and blocks arms that are both raters and targets in it. Both constants were set on held-out seeds 100–105: from round 99 on, no block of mostly honest arms reached a ratio of 2.0, and every stealthy ring of 30 or 90 did. *Rater-weighted* ETS-sep gives weight 0 to raters with at least three ratings on arms the buyer has pulled twice or more, if fewer than half of those ratings agree with the buyer's own mean outcome. It resets with each buyer, since own outcomes are per buyer. *Combined* is ETS-sep with FaP and FRAUDAR, and endorsements from blocked arms are discarded. ETS-cap, FaP over ETS-cap and uniform selection are in the result files and omitted from Table 2 for space.

Grid. $T \in \{1000, 3000\}$, both visibility conditions, $\varepsilon \in \{0, 0.1, 0.3\}$, rings of 1, 5, 30 and 90, and the four attacks, for 7 learners and 30 seeds: 22,680 runs. The competitors and variants above were run on the $T = 1000$ cells of Table 2, and the table reports them over the same seeds as the other learners.

C COMPONENT PURITY ON REAL FRAUD GRAPHS AND DOMAIN TAXONOMY

We evaluate component purity on the YelpChi and Amazon fraud benchmarks (Dou et al., 2020) as a structural negative control, establishing where graph-based component defenses apply and where graph-blind lower bounds govern. YelpChi contains 45,954 reviews (14.5% labeled fraudulent) across three relations: same user (R-U-R), same product and month with identical rating (R-T-R), and same product with identical rating (R-S-R). Amazon contains 11,944 users (users 0–3304 are unlabeled and excluded, leaving 9.5% fraudulent) across three relations: reviewed a common product (U-P-U), posted identical ratings within one week (U-S-U), and mutual review text similarity in the top 5% (U-V-U). Unweighted relations represent the candidate heavy-edge interaction graph.

For each relation, we compute connected components with at least two nodes and evaluate the distribution of fraudulent nodes: pure all-fraud components (which FaP blocks with zero collateral damage), mixed components, or isolated nodes. On all five co-activity relations, 83–99% of fraud nodes reside in mixed components, overwhelmingly concentrated in a single giant component. On YelpChi's R-U-R relation, 82% of fraud reviews are isolated (one review per account) and 17% sit in all-fraud components. On Amazon, filtering edges whose endpoints share fewer than 2 or 5 mutual neighbors fails to disentangle the components. Computing k-core decompositions reveals that the top 1% densest core nodes contain at most 21% fraudulent nodes, while on R-U-R and U-V-U the densest cores are almost entirely honest.

Structural domain boundary. These empirical findings delineate a clear boundary between three structural fraud paradigms:

1. *Mutual Collusion Networks (P2P Markets and Financial Routing):* In decentralized networks such as payment routing, credit mutuals, and peer-to-peer marketplaces, colluders must actively coordinate and endorse one another to manufacture apparent trustworthiness. Because honest participants do not randomly endorse colluders, these interactions generate dense, isolated subgraphs that satisfy Assumption A1, enabling FaP and FRAUDAR to isolate the entire ring.
2. *Uncoordinated Bipartite Review Spam (Yelp and Amazon):* On consumer review platforms, independent spammers post uncoordinated fake reviews targeting popular honest products. Projecting the resulting bipartite reviewer-product graph onto reviewer or item nodes creates an inevitable hairball: popular honest targets attract reviews from both honest users and spammers, forcing honest and malicious nodes into one giant mixed component. Here, the interaction graph provides no cluster-level separation, Assumption A1 fails, and the graph-blind lower bound of Theorem 1 strictly governs.
3. *Target-Directed Attack Ledgers (Bad-Mouthing):* When coordinated attackers attack honest arms directly, our Consensus-Residual Rater Filter (CRRF, Section 4.2) provides containment without requiring graph component purity by filtering malicious raters globally on their negative out-degree consensus residual envelope.

D CASE STUDY: PAYMENT-ROUTING SIMULATOR (TRACE)

This appendix reports an empirical case study in an adversarial payment-routing simulator (the TRACE benchmark of Kaushik et al. (2026)), featuring trust graphs, hard admission filters, and scripted attacks. This benchmark provides practical grounding for the core phenomena analyzed in the main text: subset commitment (Proposition 2), pre-switch indistinguishability (Proposition 3), and the fundamental limitation of decision-time score gating relative to evidence-level capping.

D.1 PROBLEM SETUP

We model trust-based routing as a stochastic many-armed bandit (Bubeck & Cesa-Bianchi, 2012) under adversarial arm placement. Candidate service providers represent arms. Net utility per completed transaction is bounded in $[-V, V]$ for maximum job price V. A fraction $\varepsilon = 0.3$ of providers are malicious. Malicious arms coordinate to manufacture favorable interaction evidence, build genuine histories before defecting, and exploit routing policies.

Scale and visibility. Each decision evaluates an eligible cohort of $n = 300$ to 334 candidate providers offering one of three capabilities, receiving roughly 100 jobs per capability at horizon $T = 300$. Because $n \approx 300 \gg \sqrt{100} = 10$, this operates deeply within the many-armed regime. Under *centralized* visibility, routing decisions access pooled job histories. Under *partitioned* visibility, T jobs are divided sequentially across five independent buyers, each starting from an empty private interaction history while reading the shared trust graph.

Attacks. Under *strategic default*, malicious providers behave honestly until a switch round drawn between 50% and 70% of the horizon, then defect in bursts. Under *collusion ring*, colluders fabricate mutual transaction edges each round to manufacture network centrality, defaulting on 75% of jobs from 60% of the horizon onward.

D.2 THEORETICAL BRIDGES IN THE SIMULATOR

Proposition 2 (Subset commitment)**.** *Let a cohort of n arms contain a set A of r adversarial arms, $\varepsilon = r/n$. A policy draws a subset S of m arms uniformly without replacement and routes a session of T_s jobs within S. Then:*

$$p_m := \Pr(S \cap A \neq \emptyset) = 1 - \binom{n-r}{m} \Big/ \binom{n}{m} \;\geq\; 1 - (1-\varepsilon)^m.$$

If an adversary inside S receives expected share q of session jobs, expected adversarial routes are at least $p_m q T_s$, exceeding uniform selection (εT_s) whenever $q > \varepsilon / p_m$.

Proof. The bound follows from combinatorial counting: $\binom{n-r}{m}/\binom{n}{m} = \prod_{i=0}^{m-1} \frac{n-r-i}{n-i} \leq (1-\varepsilon)^m$. The second claim conditions on $S \cap A \neq \emptyset$. At $n = 300, r = 90$, a five-arm subset ($m = 5$, drawn by subsampled UCB in partitioned sessions) contains an adversary with probability $p_5 = 0.83$. An included adversary needs only a 36% routing share to render subsampled UCB worse than uniform selection. □

Proposition 3 (Pre-switch indistinguishability)**.** *Let A be a uniformly random subset of r of the n arms, drawn independently of arm qualities. Suppose that before round t_s, every adversarial arm generates observations identical in law to an honest arm of the same quality. Then for every policy, including any gate or filter, and every round $t \leq t_s$, the chosen arm a_t satisfies $\Pr(a_t \in A) = r/n$. Expected pre-switch adversarial pulls equal εt_s, and for any past-measurable event Z, $\Pr(\text{leader} \in A \mid Z) = r/n$.*

Proof. History H_{t-1} is determined by qualities, observation noise, and internal policy randomness, all of which are independent of A prior to t_s. Thus H_{t-1} is statistically independent of A, as are a_t, the leader, and event Z. For any fixed arm i, $\Pr(i \in A) = r/n$. Conditioning yields the result. □

Proposition 3 proves that prior to defection, no algorithm can learn to avoid strategic defaulters. All policies enter the cash-out phase with identical expected adversarial exposure; differentiation depends entirely on post-switch detection speed.

D.3 Heuristic Exploration Gates and Samuelson's Limit

The baseline routing heuristic evaluated in Kaushik et al. (2026) is gap-gated Thompson sampling (GGTS): rank eligible candidates by apparent utility, compute the top-two gap $u_{(1)} - u_{(2)}$ and cohort standard deviation σ, and route greedily to the leader if $\text{gap} > \tau\sigma$, falling back to Thompson sampling otherwise.

Lemma 1 (Samuelson's limit for cohort gates)**.** *Fix an eligible cohort of size $n \geq 2$, hold $n-1$ utilities fixed, and let one candidate's utility u_1 increase. On $u_1 > \max_{i \geq 2} u_i$, the ratio $\text{gap}(u_1)/\sigma(u_1)$ is nondecreasing in u_1 and converges to $n/\sqrt{n-1}$ as $u_1 \to \infty$. This limit is the supremum over all cohorts of size n, attained when all other $n-1$ utilities are identical.*

Proof. Let $\bar{u}$ and s^2 be the mean and population variance of the fixed $n-1$ utilities, with $d = u_1 - \bar{u}$ and $\delta = \max_{i \geq 2} u_i - \bar{u}$. The cohort variance satisfies $\sigma^2 = \frac{n-1}{n^2}(d^2 + ns^2)$. For $d > \delta$, the ratio is $g(d) = (d-\delta)/\sqrt{\alpha d^2 + \beta}$ with $\alpha = (n-1)/n^2$ and $\beta = (n-1)s^2/n$. Differentiating yields $g'(d) = (\beta + \alpha\delta d)/(\alpha d^2 + \beta)^{3/2} \geq 0$. Thus $g(d)$ increases monotonically toward $1/\sqrt{\alpha} = n/\sqrt{n-1}$. This achieves the equality condition of Samuelson's inequality (Samuelson, 1968). □

Proposition 4 (Finite manipulation threshold)**.** *In the setting of Lemma 1, candidate 1 leads and the gate fires if and only if $u_1 > \bar{u} + d^*$, where*

$$d^* = \frac{\delta + \sqrt{a}\,\sqrt{\delta^2 + (1-a)\,ns^2}}{1-a}, \qquad a = \frac{\tau^2(n-1)}{n^2}.$$

As $n \to \infty$ with δ and s fixed, $d^* - \delta \to \tau s$.

Observation 1 (Winsorization saturation)**.** *Fix cohort size n, honest utilities $h_1, \ldots, h_{n-r}$, and $r < n/2$ corrupted utilities $c_1, \ldots, c_r$. Let μ_h be the median of honest utilities augmented with r copies of $+\infty$, and let $s := \mu_h + \max_j |h_j - \mu_h|$. If every $c_i > s$, the winsorized threshold $B = \text{median}(u) + \kappa \cdot \text{MAD}(u)$ depends exclusively on honest values and r. Increasing any c_i further leaves B unchanged.*

Failure of decision-time winsorization. GGTS-Robust replaces raw mean and variance with median and MAD ($\kappa = 3.0$). While Observation 1 guarantees saturation against unbounded inflation, it fails practically: when multiple colluders manipulate scores above B, both leader and runner-up are clipped to B, producing a winsorized gap of zero. The gate stays closed on 86–100% of rounds, collapsing GGTS-Robust into standard Thompson sampling. Thus, capping decision-time utility statistics closes the inflation lever only by forfeiting greedy gains.

Table 3: Payment-routing regime diagram: fraud exposure (sats, mean ± std over seeds 42–51) at $T = 300$ by density regime, visibility and attack. *Raw* is the gap-gate at $\tau = 0.5$; *robust* is GGTS-Robust; *uniform* is uniform candidate selection; *SS-UCB* is subsampled UCB ($m = 5$ partitioned, $m = 10$ centralized).

Density	Vis.	Attack	greedy	TS	raw	robust	uniform	SS-UCB
sparse	central.	collusion	9.2±6.3	47.4±15.3	25.1±8.2	47.0±10.9	44.6±13.4	37.7±13.7
sparse	central.	strategic	3.5±3.8	8.9±4.7	5.5±4.5	7.9±4.8	4.0±3.2	9.3±5.2
≥0.8%	central.	collusion	47.7±7.7	48.5±10.1	51.3±11.9	51.0±12.4	48.7±8.4	45.7±8.3
≥0.8%	central.	strategic	5.1±3.2	7.7±6.0	5.6±5.9	6.0±5.7	5.3±3.0	5.5±3.3
sparse	partit.	collusion	44.5±20.5	47.5±12.4	65.5±35.5	47.5±12.4	40.6±5.0	78.7±12.7
sparse	partit.	strategic	28.3±9.5	4.7±3.8	35.9±16.4	11.7±7.7	3.4±2.0	14.9±5.7
≥0.8%	partit.	collusion	40.8±20.2	44.4±8.5	44.7±22.3	44.4±8.5	34.5±6.5	67.6±11.9
≥0.8%	partit.	strategic	24.4±11.3	6.6±4.6	29.2±17.6	7.6±5.3	5.7±4.0	10.7±6.3

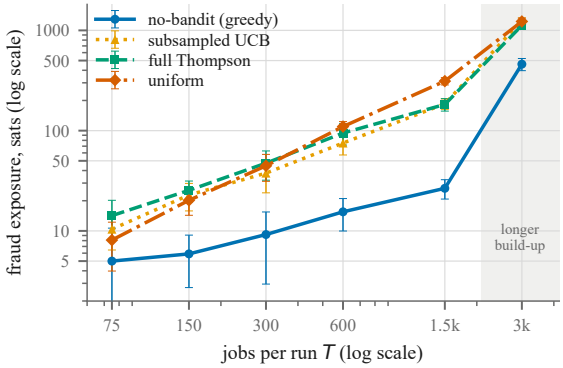


Figure 1: Fraud exposure against horizon (mean ± std over 10 seeds), centralized visibility, sparse density, collusion ring. Uniform selection is level with learning policies up to 300 jobs and deteriorates by 1,500. GGTS-Robust tracks full Thompson sampling within seed noise; widened-confidence elimination matched uniform seed for seed through 1,500 jobs.

D.4 Empirical Findings Across Regimes

Table 3 reports fraud exposure across 8 conditions (density × visibility × attack) for six representative policies over seeds 42–51.

1. Visibility dictates policy safety. Under centralized pooled evidence at sparse density, greedy selection achieves lowest fraud (9.2 sats under collusion ring, beating TS on all 10 seeds, $p = 0.002$). Under partitioned visibility, the ordering reverses: greedy incurs severe fraud under strategic default (28.3 sats at sparse density, losing to TS on 10/10 seeds, $p = 0.002$). Uniform selection achieves lowest fraud across all partitioned settings. No fixed policy is safe across visibility regimes.

2. Subsampled UCB suffers subset commitment. Under partitioned evidence, subsampled UCB incurs 78.7 and 67.6 sats of fraud under collusion ring, roughly double uniform selection (40.6 and 34.5 sats, $p = 0.002$ on all seeds). As Proposition 2 predicts, committing to 5-arm subsets draws an adversary with 83% probability, and optimistic index exploration over-pulls the apparent top arm. A standalone replication in a standard 300-armed bandit outside the routing simulator confirms this effect across 30 seeds (1.43× uniform fraud, $p = 4 \times 10^{-11}$).

3. Greedy capture at long horizons. Figure 1 traces fraud scaling from $T = 75$ to 3,000 jobs. Greedy maintains low fraud through 1,500 jobs, but between rounds 390 and 450 under collusion ring, accumulated fabricated trust captures greedy: its malicious routing share spikes above 30%, driving fraud to 460.5 sats at $T = 3{,}000$.

4. Elimination baselines degenerate into uniform selection. We tested corruption-robust widened-confidence elimination (Gupta et al., 2019). In 535 instrumented decisions, zero candidates

were eliminated. Because cohort size ($n > 300$) dwarfs per-candidate sample counts, confidence radii remain wide throughout, reducing elimination to uniform random selection.

Core synthesis. The failure of GGTS and GGTS-Robust demonstrates that attempting to detect manipulation at decision time via population score spread is fundamentally fragile. Robustness requires intervening on the evidence generation process itself: separate capping (ETS-sep) to bound endorsement inflation and rater consensus filtering (CRRF) to isolate coordinated spam.