CORRUPTED BLOCKS ARE WORSE THAN NONE: AGENTTRACE, A MEASUREMENT STUDY OF LATENT-SOURCE RECOVERY AMONG IDENTIFIER-LEAKING CLIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering which events in an anonymized stream share a source is a well-posed learning problem that lacks scorable data without privileged identity, simulation or leakage: the predicted quantity is the field removed. AGENTTRACE takes labels from inside the public stream: structured client identifiers across 182,119,794 order events and 23 asset-day sessions resolve to 153 actors, withheld from every model and corroborated, unevenly across identifier families, by an identifier-blind relation. The capture cannot be redistributed; the reusable object is the protocol. The labels cover 38.99% of events and are not random: a behavior-only classifier reads labellability at 0.8790 mean AUROC over 17 sessions, under a rule fixed in advance, so this study measures recovery among identifier-leaking clients. Inside that population the signal exceeds identifier-family classification: with family held constant, version-1 UUID actors separate at 0.9688 and 0.9981 balanced accuracy. The central measurement is a factorial, none of it preregistered, that fixes featurizer, algorithm, oracle K and metric and varies only the imposed boundary source. Over 16 sessions at $W = 32$, oracle boundaries beat random size-matched blocks by $+0.1982$ mean order-level ARI, positive in all 16, but beat not blocking at all by only $+0.0233$, a gain that does not survive asset-level inference and reverses in all four ETH sessions. Random size-matched and contiguous blocks both score below not blocking in all 16, by -0.1749 and -0.1714 mean ARI. The fingerprint survives a change of day and of month, at 7.97 to 31.23 times chance across ordered day pairs, and transfers far more weakly across assets, 2.4 times chance pooled and 4.0 once the one build-keyed family is excluded. A label-free contrastive encoder shows no consistent held-out-day gain at oracle K, yet avoids the collapse hand features suffer under density-based clustering. Neither timestamp coarsening nor jitter nor size discretization closes the behavioral channel at any magnitude we could build against a refit hand-feature attacker; jitter of at least 1 ms deletes the timing-metadata side channel and a 1% size grid the size-precision fingerprint at 0.3% distortion.

1 INTRODUCTION

An anonymized asynchronous event stream is a sequence of timestamped records whose source field has been removed. Recovering which records share a source, the latent actor partition, is a well-posed learning problem that arises wherever such a stream is published: server logs stripped of user identifiers, encrypted traffic to a shared service, sensor events from a device several people use. What has to be predicted is exactly what was deleted, so a prediction has nothing to be scored against. The literature works around that by taking identity from a privileged source (Tumminello et al., 2012; Cont et al., 2023; Sato & Kanazawa, 2023a; Kirilenko et al., 2017) or by simulating the stream so labels are assigned at generation time; a study of trading-agent recovery gives the reason: "we do not have access to ground truth when working with real-world data" (Wilinski et al., 2025).

Our substrate supplies labels from inside the published stream: a public feed of individual order submissions from one cryptocurrency exchange, 182,119,794 of them over 23 *sessions*, each one

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.

traded instrument on one date, collected before the venue stopped rebroadcasting the field the labels rest on. Some clients populate a free-text client identifier with structure, and six such layouts resolve to 153 *actors*, an actor being an inferred client or process namespace, covering 38.99% of submissions. The partition is read out of the public stream, withheld from every model, and corroborated, unevenly across identifier families (layout classes), by a relation that reads no identifier field.

The population those labels cover is not the stream, and we measure the gap. A classifier given 20 behavioral features and no identifier-derived input predicts whether an order is labellable at a mean AUROC of 0.8790 across 17 scored sessions. The mapping from that number onto a scope decision was fixed before the diagnostic existed: a high AUROC narrows the claim to recovery among identifier-leaking clients, and whether recovery extends to clients whose identifier hygiene is sound is not measurable with this instrument. What survives the narrowing is more than identifier-family classification: identifier families are themselves separable from behavior at 0.9717 balanced accuracy against a chance of 0.100, yet with family held constant actors still separate at six to seven times chance in the largest family and above 0.96 in the version-1 UUID (node-field identifier) family.

AGENTTRACE is the measurement study and evaluation protocol this makes possible. A *trace* is W events consecutive within one actor's own sequence inside one session, $W = 32$ in the primary configuration, and a *segmentation oracle* hands a method that grouping instead of requiring it to be inferred. Given segmentation, a gradient-boosted classifier separates actors at 0.7064 balanced accuracy against a chance of 0.0339.

What segmentation is worth had been measured as the difference between two whole pipelines, which confounds the boundary source with the featurizer, the dimensionality and the clustering algorithm. A factorial holds featurizer, algorithm, metric and oracle K, the true actor count supplied to every arm alike, fixed while varying the imposed boundary source. Over 16 sessions, correct boundaries beat random blocks of the same sizes by $+0.1982$ mean order-level adjusted Rand index (ARI), positive in all 16. Blocks of those sizes with the wrong membership do harm: random and contiguous blocks fall below per-order clustering in every session, by -0.1749 and -0.1714. Against that same baseline correct boundaries gain only $+0.0233$, which does not survive treating the four assets rather than the 16 sessions as the unit of inference, and which reverses on ETH.

Two further measurements bound the fingerprint: it survives a change of day and of month, at 7.97 to 31.23 times chance across ordered day pairs and about sixteen times across a month gap, and transfers far more weakly across assets, 2.4 times chance pooled and 4.0 once the one family whose label unit is a software build is excluded from both sides. A label-free contrastive set encoder, attention-based or not (Lee et al., 2019; Zaheer et al., 2017; Chen et al., 2020), gains nothing consistent in oracle-K accuracy once a day is withheld from its objective, but under one protocol for choosing the number of actors both instances avoid the collapse hand-crafted features suffer; the count stays poorly estimated.

Two contributions carry the paper. One is the label instrument: a corpus whose actor partition is read out of the public stream, audited family by family, and scoped by a diagnostic that fixes the population it speaks for. The corpus cannot be redistributed, so the reusable object is the protocol, released as code, a SHA-256 capture manifest, hashed labels and a quantized smoke-test sample. The other is that factorial, with a registered leave-one-family-out replication. Supporting analyses cover channel decomposition and leakage audits, transfer across day, month and asset, three output-side defenses that each act as a step function on one feature class, and two evaluation-protocol findings that a corpus of real absences exposes and a synthetic one hides.

2 RELATED WORK

Actor identification in market data. Work adjacent to our substrate obtains identity rather than recovering it. Tumminello et al. (2012) cluster investors whose identities a national registry supplies, so the actor partition is an input to the method, and that registry carries the work that follows (Musciotto et al., 2018; Baltakys et al., 2018); elsewhere a broker's clients, audit-trail accounts and venue-flagged high-frequency traders are grouped into behavioral types with the account label in hand (Cont et al., 2023; Kirilenko et al., 2017; Hagströmer & Nordén, 2013; Menkveld, 2013; Wright et al., 2022). Nearest of all, Sato & Kanazawa (2023a;b) measure per-account order splitting

with exchange account identifiers available, establishing that an account leaves persistent structure in the flow; the identifiers are never withheld, so recoverability is never tested. Without labels the question is redefined into type detection or downstream-validated archetypes (Zhang et al., 2025; Goudarzi & Bazzana, 2023; Nechepurenko, 2026), de-anonymizing anonymous orders is named as a threat with no dataset and no measurement (Mavroudis, 2019), and anonymity is studied as a policy switch priced in liquidity (Foucault et al., 2007; Comerton-Forde et al., 2005; Comerton-Forde & Tang, 2009).

Re-identification and behavioral fingerprinting. We follow the form of Narayanan & Shmatikov (2008; 2009): take a release believed anonymous, show identity survives, and quantify what recovering it costs. Mobility and credit-card studies fix the form of the budgets, identifiability against observation length and against resolution coarsening (de Montjoye et al., 2013; 2015). Re-identification from an incomplete sample is our position exactly (Rocher et al., 2019). Address clustering on a partially tagged public ledger is the closest methodological cousin outside this domain (Meiklejohn et al., 2013). Behavior alone re-identifies people through browsing traces (Oliveira et al., 2025), writing style (Abbasi & Chen, 2008; Narayanan et al., 2012), keystroke and touch timing (Monrose & Rubin, 2000; Killourhy & Maxion, 2009; Frank et al., 2013) and browser configuration (Eckersley, 2010). Our labels being in-band, no linkage against an auxiliary dataset stands between feed and score. Crichton et al. (2025) are skeptical that behavior-based identification holds up at scale, and our measurements agree in part: behavior separates actors sharply once events have been grouped, while recovering the grouping without supervision is weak.

Traffic fingerprinting, entity resolution, open sets and stream encoders. Traffic fingerprinting supplies the protocol grammar we adopt: open-world scoring at realistic base rates (Panchenko et al., 2016; Cherubin et al., 2022), stress tests breaking laboratory assumptions (Juarez et al., 2014), the ladder from hand-engineered features to deep sequence models (Sirinam et al., 2018; Rimmer et al., 2018; Bhat et al., 2019; Sirinam et al., 2019), an attack reported alongside what stopping it costs against an adaptive attacker (Wang et al., 2014), and device-instance identification at small K (Sivanathan et al., 2019; Miettinen et al., 2017). Author disambiguation is the closest well-posed instance of T1a's abstract form, with the same blocking problem, imbalance and metric conventions (Subramanian et al., 2021; Fellegi & Sunter, 1969); deciding pairs independently is weaker than propagating co-occurrence constraints into a global partition (Bhattacharya & Getoor, 2007; Mudgal et al., 2018; Li et al., 2020), which is what our coincidence baseline attempts and fails to achieve at the order level. Unknown classes at test time, enrollment from k examples and gallery retrieval are standard (Scheirer et al., 2013; Bendale & Boult, 2016; Vaze et al., 2022b; Snell et al., 2017; Schroff et al., 2015; Snyder et al., 2018; Zheng et al., 2015; Ye et al., 2022), pipeline-derived labels are accepted when the derivation is documented (Nagrani et al., 2017), and generalized category discovery is the nearest formalism (Vaze et al., 2022a), from which ours differs in having no labeled seed at inference, an unknown and unbalanced K, and channel attribution as the question. Neural temporal point processes with their transformer and state-space successors (Du et al., 2016; Mei & Eisner, 2017; Zuo et al., 2020; Shchur et al., 2020; 2021; Rubanova et al., 2019; Gu et al., 2022; Gu & Dao, 2023; Xue et al., 2024), permutation-invariant set encoders (Lee et al., 2019; Zaheer et al., 2017; Desplanques et al., 2020), contrastive self-supervision (Chen et al., 2020; He et al., 2020; van den Oord et al., 2018) and cluster-then-learn (Caron et al., 2018; Xie et al., 2016) are this modality's encoder families. These literatures address closely related formal tasks; we found no prior benchmark combining this public-feed label construction with order-level latent-source scoring. We evaluate three learned encoders among them: two invariant set encoders, attention-based and statistics-pooling, and one order-aware recurrent encoder (Section 5).

3 DATA

The corpus is useful for one reason: some clients embed structure in a free-text field the venue rebroadcasts.

Corpus and labels. We parse 465 capture files in place, 182,119,794 of whose lines are `received` events, the message the venue emits when a client submits an order. The captures cover 23 sessions on BTC, ETH, SOL, AVAX and SHIB across 2022-04, 2022-05, 2023-03 and 2023-08, 18 carrying the client field and 5 not. Of the `received` events, 160,185,756 carry a client-supplied

identifier, assigned by one ordered rule chain to exactly one of ten layouts, so the families partition the events. Six yield actor labels, among them version-1 UUIDs keyed on the node field (Leach et al., 2005) and a constant-prefix layout, together covering 38.99% of all `received` events (Appendix E). The constant-prefix family, close to half the actor population, is discovered statistically rather than by hand, under a recurrence rule that admits 74 of 83 candidate heads.

Actor population. Taking version nibbles at face value manufactures phantom single-event actors, which an embedded-clock test separates without exception, and a stable node value establishes a UUID node namespace rather than a physical interface address, RFC 9562 permitting randomly derived nodes (Davis et al., 2024) (Appendix E). The corpus contains 153 actors and no long tail: the smallest carries 38 orders, and the count is insensitive to the density floor. Generalization probes occur naturally, with 42 actors in all four primary assets and 35 in exactly one, 58 on exactly one day, and 19 recurring across the roughly one-month gap between capture eras. Some operators run several identical processes; two rules frozen before any model was fit, one on identifier structure and one on activity vectors, confirm four such groups of seven, collapsing 28 actors and taking K from 153 to 129. Neither reads activation time, so the synchronized activation in Appendix F is independent corroboration.

A label-validity certificate that uses no identifier, and where it runs out. Among pairs of labeled events sharing an exact microsecond timestamp within a session, 0.9994 of 2,139,321 pairs are same-actor pairs (Clopper–Pearson 95% interval $[0.99937, 0.99943]$) against a per-session chance purity of 0.156–0.275, and per-session values run 0.9987–1.0000 across all 18 labeled sessions. The relation covers 4.86% of labeled events, so it certifies rather than labels. That pooled figure is dominated by families whose chance baseline is already near one, so the certificate is always shown against each family's own baseline (Appendix Table S1); two of the six families contribute almost no evidence. The mechanism is a hypothesis: a batched client write stamped once by the matching engine would produce this, while the competing explanation of coarse capture-side granularity predicts purity at the chance level, and the feed does not adjudicate between them.

What the labels are not. An actor is an identifier-derived source namespace, not an economic entity, and its grain depends on the family: a version-1 node value marks a UUID node namespace, a counter layout a process or installation, and the constant-prefix layout a software build, so the 153-way partition mixes at least three grains of client-software identity and no result here is a statement about one uniform construct. Where one operator runs parallel instances the instance-level task is harder than entity re-identification, which is why we report both partitions. And 61.01% of `received` events carry no recoverable label, because the session lacks the client field or the identifier has no recoverable structure, as when a client randomizes it properly, so the labeled remainder is selected on identifier hygiene; that selection is measured rather than conceded (Section 5) (Rocher et al., 2019).

4 Evaluation protocol

Tasks, and which of them this paper reports. Two tasks are primary. T1a partitions a balanced set of traces with identities hidden and the segmentation oracle given; T1b partitions a contiguous slice of the labeled stream (unlabeled orders removed for every arm) at the level of the individual event, with no segmentation given. T1a hands the model the fact that a trace's 32 events share a source, the standard authorship-clustering formulation, so high T1a with low T1b marks a capability gap. Six further tasks separate the ways a fingerprint can fail: closed-set classification, open-set admission, held-out-day and held-out-asset retrieval, enrollment from k traces, observation and defense budgets, and identifier-family attribution (T2–T8, specified with precedents in Appendix G). This paper reports T1a, T1b, T3, T4, T5, the supervised closed-set experiment, a reduced form of T8, a selection diagnostic, and the reported components of T7.

Concrete instantiation. A supervised closed-set experiment scores traces on BTC and ETH 2022-05-16 with a gradient-boosted tree classifier in seven variants that isolate the channels. T1a clusters traces in all 16 primary sessions and T1b partitions the middle 500,000 labeled orders of a held-out session at the level of the individual order. The segmentation factorial rescores 16 such 500,000-order slices, four assets on four consecutive days, under four boundary sources at

$W \in \{2, 4, 8, 16, 32\}$, every arm sharing an order set, a featurizer, a standardization, an algorithm, an oracle K and a metric. Retrieval follows the person re-identification protocol across a held-out day, the month gap and a held-out asset pair with the day fixed (Appendix G). Open-set admission scores three unknown sets, including one matched to each session's natural pool (Vaze et al., 2022b). Three further runs support the scope claims: a reduced form of T8, a selection diagnostic, and family-stratified label validity.

Channels: four registered, three surviving. The protocol decomposes recovery into four channels, one of which the experiment testing it removes. R is how much and how fast. S is when: session-relative activation windows and time of day. B is how: the rate- and schedule-free shape of sizes, prices, lifetimes, cancellations and inter-event gaps. M was sub-millisecond metadata and arrival coincidence, removed by the falsification in Section 5. Each is isolated by a variant of one identical protocol and reported as a total and a marginal (Appendix G).

Leaks and channels are not the same thing. The distinction governing every regime is between information available to us and information available to anyone watching the feed: a *leak* is ours alone and forbidden everywhere, removed by construction rather than post-hoc checking (Appendix M enumerates the forbidden fields and the full read-set contract). A *channel* is what the venue broadcasts to everyone, including timestamps at native resolution, and exploiting one is the attack, not a shortcut around it. Hence paired regimes: one quantizes timestamps to 1 ms and forbids sub-quantum, self-coincidence and cross-event co-occurrence features; the other runs at native microseconds and admits a coincidence union-find as a first-class method.

Audits, populations and preregistration. Leakage is the standard failure of benchmarks built on found data (Kapoor & Narayanan, 2023). The suite gates itself under two rules adopted after our own pipeline broke them (Reproducibility Statement, Appendix A). The scored population is actors with at least 256 events in the session, $8W$ at $W = 32$ and the smallest value yielding four training and four evaluation non-overlapping traces (Appendix H). All of this is fixed in a preregistered protocol, amendments appended rather than edited in, including two metadata-channel predictions, each with a mechanism, a threshold and a falsifier, registered before the experiments that decide them (Section 5, Appendix K).

Uncertainty conventions. Supervised splits are temporal within a session with a guard band (Appendix H), and clustering is reported per session and never pooled. Uncertainty is reported at the level each quantity supports: classification and retrieval numbers carry a BCa bootstrap over 1,000 resamples, a permutation null and the session's exact chance baseline. The factorial's per-session contrasts carry a paired percentile bootstrap over orders with both clusterings held fixed, conditional metric uncertainty rather than uncertainty over sessions or actors; cross-session statements are made over the session and asset distributions instead. The learned encoder's spread is over three trained seeds only, and the permutation nulls put an add-one floor of $p < 0.005$ on classification and $p < 0.001$ on retrieval p-values (Appendix G). Differences between regimes are read against the Bland–Altman mean of the two arms rather than one of them, which withdraws one conclusion (Appendix I).

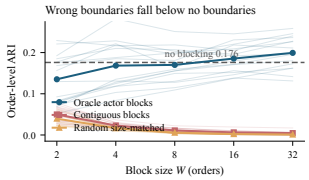


Figure 1: Mean order-level ARI against block size W: three imposed-boundary arms and the no-blocking line, 16 sessions.

5 Results

Identity is recoverable from behavior once segmentation is given. With segmentation supplied, a gradient-boosted classifier separates 29 BTC and 30 ETH actors at 0.7064 mean balanced accuracy against a chance of 0.0339, and the rate- and schedule-free representation alone carries 0.7137 on those two sessions. Both stay far above chance in all 16 sessions ($p < 0.005$; weakest, full set, SOL 2022-05-17 at 0.5825–0.5938; Appendix S, Table S8).

Corrupted blocks are worse than none. The factorial runs over 16 sessions, four assets across four consecutive days (Figure 1; per-asset values in Appendix N). None of its contrasts was reg-

istered in advance, so each stands on how consistently it holds. Its widest contrast sets oracle boundaries against random size-matched blocks, sharing the featurizer, algorithm and block sizes and differing only in membership: at $W = 32$ it is $+0.1982$ mean order-level ARI, from $+0.1308$ to $+0.2574$, positive in all 16 sessions. Nor is the oracle-to-contiguous gap an artifact of the hand featurizer: on the same orders one frozen, non-adaptive Set Transformer, each session scored by the checkpoint that held its day out, reaches -0.0001 to 0.0142 order-level ARI under contiguous blocks, 0.1284 to 0.2452 below its oracle-block score, and the registered rule holds in all 16 sessions and four panels (three seeds, no learned $W = 1$ arm, two launch batches; Table S27). Corrupting the boundaries costs more than never drawing them: random size-matched blocks score -0.1749 against per-order clustering and contiguous blocks -0.1714, each negative in all 16, so two distinct corruption constructions agree.

The second contrast, against no blocking at all, is the alternative available to a method with no oracle to corrupt, and it is not established. The mean gain is $+0.0233$, positive in 12 of 16 sessions, but the sessions are not 16 independent draws: 94.2% of their variance lies between the four assets, four consecutive days of one asset behaving as near-replicates, so the effective sample is four. Clustered by asset the one-sided 95% lower bound is -0.0351 at $p = 0.209$; clustering by session returns $+0.0047$ at $p = 0.030$, printed only to mark the error of treating clustered observations as independent. That choice of unit was made after we saw the assets separate, so it is post-hoc, moving the result against the claim we had been making.

The assets separate cleanly: BTC, AVAX and SOL gain between $+0.0306$ and $+0.0685$ with no reversal in twelve sessions, while ETH loses -0.0472 and reverses in all four of its own. A harder no-blocking baseline does not account for it: SOL, ETH and AVAX have similar per-order ARI while only ETH reverses, and the gain does not correlate with baseline recoverability read against the Bland–Altman mean (Appendix I). The gain is confined to the larger windows, $W = 32$ being the primary window fixed before any of these results existed.

No single family drives the factorial. A leave-one-family-out replication, registered before its exclusion conditions were computed and run under 448 leakage gates, repeats the factorial under four exclusion conditions and each identifier family (Appendix O): the destructive contrasts hold in every one, 0 of 16 positive throughout, and the oracle-versus-no-blocking gain stays insignificant at the asset level under either single-family exclusion, so no single family's composition or validity drives these conclusions, not that the proxy labels are established as valid. The ETH reversal does not localize to any family, so its driver remains unidentified.

P01 confirmed, P02 falsified. Two predictions about the sub-millisecond timing channel were registered before their deciding data existed. P01 (the channel adds at most 0.02 once segmentation is given) passes at $\Delta_M = -0.0012$. P02 (it would dominate when segmentation is unknown) is falsified at $\Delta_M = -0.1174$, opposite in sign and about four times its falsifier threshold: union-find over exact-microsecond equality returns 486,106 clusters over 500,000 orders (Appendix D), so near-perfect pairwise precision at 4.86% coverage does not compose into a partition. The registered consequence was executed and Channel M left the decomposition (Appendix Q).

Unsupervised recovery is real and weak, and supervision pays late. Across the 16 primary sessions, instance-level ARI averages 0.2215, from 0.1532 to 0.2736, reported per session and never pooled. Oracle K is supervision, held constant across every arm compared, and ARI is chance-corrected, so the behavior supporting 0.7064 under supervision supports a partition well short of ground truth without it. A classifier's predicted labels are themselves a partition, so the two sit on one axis that rises non-monotonically: on the one session it is measured, one prototype per actor scores below unsupervised clustering and does not overtake it until eight labels per actor (Appendix L). Replacing oracle K with an estimate costs 24.7% of mean ARI under eigengap and 87.2% under density-based HDBSCAN.

Learned invariant set encoders avoid the hand-feature collapse; the oracle-K gain is transductive. A contrastive Set Transformer trained without labels on 104,742 traces and frozen reaches 0.2570 mean oracle-K ARI over three seeds against 0.2215 for hand-crafted features, training and scoring on the same sixteen sessions (Appendix V). Retrained with each of four days held out in turn, the oracle-K gap against hand features is $+0.0117$, $+0.0138$, $+0.0202$ and -0.0060: three

folds positive, one negative, mean $+0.0099$ against a fold-to-fold SD of 0.0112. On 05-16 all three seeds beat hand and the gap exceeds that fold's seed range, alone among the four. Days are the unit this design generalizes over, and across days the gap does not separate from zero: the withdrawal stands. HDBSCAN collapses on hand features on all four held-out days (0.0212, 0.0351, 0.0216, 0.0355) and not on the frozen embedding (0.2029, 0.2499, 0.2419, 0.2444), interpolating folds at or above extrapolating ones as preregistered; eigengap supports no such reading and is unstable for both. A parameter-matched statistics-pooling control (Deep Sets family, 680,192 against 705,024 Set Transformer parameters) reproduces the HDBSCAN repair on both adjudication folds (0.2031 against 0.2029, 0.2537 against 0.2444; hand 0.0212/0.0355, R1 fixed in advance) and the oracle-K withdrawal (0.2098/0.1828 against hand 0.2286/0.2158, no advantage), so the repair is a property of learned invariant representations, not attention – the Set Transformer is one instance of the class. A parameter-matched order-aware encoder (707,952 against 705,024 parameters) also avoids the collapse, at 0.1909/0.2099 on 05-14/05-17 against 0.2029/0.2444 (hand 0.0212/0.0355): inside the preregistered equivalence margin on one fold and below it on the other, so neither equivalence nor a direction is claimed; its order-shuffled twin (0.1868/0.2326) stays within that margin on both folds, in opposite directions, so for this architecture presentation order carries no recoverable signal under this protocol (Appendix V). What the class separates is actors, not families: within-family ARI is 0.5410 for `uuid_v1`, 0.2515 for `const_prefix16`, 0.2306 for `zero_mid_node` and 0.0657 for `date_counter`, against a pooled 0.2570, full-corpus embedding; held-out checkpoints reproduce the pattern in the same family order (0.5000, 0.2113, 0.2431, 0.0574), with the two middle families exchanging places, so the held-out advantage is also actor recovery.

The fingerprint survives a day and a month, and is substantially asset-specific. Retrieval scores only queries whose actor has a gallery entry, 202,186 of 314,226, so Rank-1 here is among retrievable queries, not an end-to-end re-identification rate. Over all 48 ordered day pairs Rank-1 runs 7.97 to 31.23 times chance, mean 17.4, with 47 of 48 above the registered eightfold threshold, so day transfer is a property of the axis, not of the chosen pair (Appendix W). Across the month gap the same protocol returns 0.3581, about sixteen times chance (Appendix Table S12). Across assets it transfers far more weakly: with the day fixed so only the asset axis varies, a gallery of 66 BTC and ETH actors matches SOL and AVAX queries at 0.0366, about 2.4 times chance – a specific failure rather than a collapse of the representation, since clustering those query traces among themselves returns 0.3338 ARI, above the 0.2215 within-session average. A long-horizon warning travels with it: retaining the ten price-scale features across a month scores two orders of magnitude below chance, price level acting as a clock.

Part of the cross-asset failure belongs to the label unit rather than to behavior. The same analysis rescores retrieval (Appendix O). Cross-asset Rank-1 is 2.42 times chance pooled, 4.01 with `const_prefix16` (the one family whose label unit is a software build, not a single actor) excluded from gallery and query alike, and 3.47 restricting both sides to corroborated-unit `uuid_v1` actors. The registered rule softens the reading above four times chance; 4.01 clears it by 0.01 with a binomial interval straddling the boundary, so the threshold settles nothing: on the identical 5,265 non-build-family queries, removing the 24 build-level gallery entries takes Rank-1 hits from 8 to 503, and that family carries essentially all of the pooled cross-asset Rank-1, earned by matching the build rather than the actor. Day transfer is not carried that way: under the same exclusion it holds at 10.1 to 12.8 times chance on all four assets, above the eightfold registered as survival.

On this corpus, random actor holdouts are easier than naturally unseen actors. Open-set admission is usually evaluated by calling a random subset of known classes unknown; this corpus supplies actors genuinely unseen in a session's training window, and the two disagree. Max-logit scoring on the closed-set classifier (Vaze et al., 2022b) reaches a mean AUROC of 0.7368 on the synthetic 30% holdout but 0.3599 on natural probes, below chance in 8 of 11 scorable sessions; orientation does not explain it, since the variants share one code path and a flipped convention would invert both. An energy score reproduces the gap ($+0.3904$ against max-logit's $+0.3770$); Mahalanobis settles nothing (Appendix P). Two attempts to attribute it to composition fail: restricting both pools to the single family supplying 82.7% of the natural one leaves the disagreement intact (0.4107 against 0.7201), and matching the holdout to each session's natural pool in size and family composition leaves it at 0.3700 against 0.6437 on six of eleven sessions, a narrow base, positive under every leave-one-session-out drop ($+0.15$ to $+0.37$) yet nearly halved without BTC 2022-05-

14 – both reversals are sessions whose natural pool holds two actors. A matched holdout is carved from the known set while natural unknowns never entered it, and naturalness and cohort structure stay unseparated: one session's natural pool is a sibling group that activates together the next day, an artifact no i.i.d.-actor or single-session simulation would surface (Appendix P).

The certificate is uneven across families, and the signal is not merely client software. Within each family, against that family's own chance baseline, coincidence lifts pair purity by $9.32\times$ and $7.66\times$ in the two counter-keyed families but only $1.01\times$ and $1.00\times$ in the two supplying almost no evidence, so the pooled figure must never stand alone. Families separate from behavior at 0.9717 balanced accuracy, raising the objection that a method might read the SDK; holding family constant removes that, and identity survives at 0.9688 and 0.9981 inside `uuid_v1` against 0.5107 and 0.4274 inside `const_prefix16` (Appendix R), a factor-of-two spread.

Coarsening the timestamps does not close the channel, and partly re-manufactures the channel it removes. We replay the two labeled 2022-05-16 sessions with every arrival floored to a quantum q and every feature re-derived, a counterfactual simulation on a stream we hold, scoring one attacker fitted on clean traces and one retrained from scratch on the defended traces (Appendix Y). Across q from 1 μs to 100 ms, both attackers, sessions and timestamp regimes, clustering ARI stays 0.18–0.21 over a ten-seed ensemble against a permutation null whose 95th percentile is 0.0002, and balanced accuracy stays 0.68–0.74 against a chance of 0.0339, at a cost of 57.7 ms RMS displacement, a utility-loss lower bound. Adaptivity buys nothing measurable and no degradation is licensed for the adaptive attacker, right-censored at $q = 1$ s where two dispersion features are undefined under either regime. At 16 sessions ($q = 100$ ms) the adaptive attacker stays undegraded (0/16) while the non-adaptive one degrades in 5/16, severely on all four AVAX ($0.259 \rightarrow 0.076$), and refitting the scaler restores it: the program's clearest overstatement. Flooring arrivals also manufactures exact-timestamp ties, read as the metadata channel the falsification removed: at $q = 100$ ms those columns move 1.9–3.9 clean SD, reaching the partition on ETH under both arms (10–100 ms, undetectable on BTC) and on AVAX under the adaptive arm (2/4, thin-book) (Appendix Y, Figure S4, Tables S14–S16).

Jitter distorts the signal for real and still does not close the channel. Symmetric jitter up to $\delta = \pm 1$ s scrambles half of adjacent orderings at the ceiling and degrades the one-second message-rate series by 0.36–0.38 clean SD at 577 ms RMS displacement, yet clustering ARI stays 0.17–0.19 against the null and balanced accuracy 0.68–0.74 under the retrained attacker at every δ up to 1 s, never below 0.66 under any arm (Appendix Y, Table S18). On BTC only the non-adaptive classifier degrades at $\delta \geq 100$ ms (0.7373 to 0.6634) and retraining recovers most of it (0.7039–0.7164), overstating protection; ETH shows none. The directional ARI decline (0.19 to 0.17) never clears clean's range and is reported, not claimed. At $\delta = 100$ ms the null result replicates across all 16 primary sessions and both standardizer fittings: 0 of 32 cells claimable (Appendix Y). Jitter of at least 1 ms deletes the exact-timestamp coincidence features, the mirror of coarsening's tie manufacture.

Few observations suffice, and uniqueness lives in attributes rather than timing. Rank-1 retrieval against a held-out-day gallery rises monotonically with the observation budget in all eight asset-by-regime curves, exceeding 0.5 within 4–32 orders in three of four assets (not reached in BTC at $W' = 32$); Rank-1 ≥ 0.9 is reached nowhere (Appendix Y, Table S19). One-second timing alone identifies almost no one at small budgets in deep books (≤ 0.0002 at one order, BTC/ETH), but the joint tuple of second, side, order type and exact size makes 49–79% of labeled actors unique from a single order and above 90% by 16 orders in every session, an oracle-matcher upper bound, not an end-to-end rate or a private-size-grid artifact.

Coarsening size deletes a precision fingerprint, the program's one claimable degradation, and it plateaus. Rounding size onto a 1% geometric grid (mean relative error 0.29%) is the only defense whose defended ten-seed clustering-ARI range lies wholly below the clean range, all four panels, both attackers: a quarter to a third of ARI, $\sim$0.19 to 0.12–0.16; coarsening 50 times further buys nothing more. The retrained classifier is essentially unchanged (-2.3 to $+1.1$ points) and adaptive ARI never enters the null range: the effectiveness criterion is met nowhere. Single-order uniqueness falls only gradually (0.5–2 points at 1%, roughly halved at 50%); eight-order uniqueness

stays essentially untouched (0.78–1.00). The degradation replicates across all 16 primary sessions and both standardizer fittings with zero exceptions (32 of 32 cells), surviving the asset heterogeneity that undercut session-level inference elsewhere (Appendix Y).

The one defense that degrades the hand-feature attack does not degrade a learned estimated-K one. Under an attacker denied the true actor count – a frozen held-out Set Transformer embedding, estimated-K HDBSCAN, non-adaptive, segmentation still given – size discretization at $g = 1\%$ sits inside clean's own seed range on both sessions (0.1585/0.1577; 0.1692/0.1711), and coarsening and jitter stay within ± 0.009 of clean (largest movement $+0.0084$). The mechanism is attacker-specific: discretization deletes a precision/roundness class the encoder never reads (its per-event inputs carry only a log-size term a 1% grid moves by at most 0.5%), so it defends feature-engineered attacks only (Appendix Y).

The labellable population is behaviorally distinct from the rest of the stream. Whether the scorable 38.99% represents the stream is decidable from behavior, and the rule mapping the answer onto the paper's scope was fixed first. A gradient-boosted classifier given 20 order-level behavioral features and no identifier-derived input predicts labellability at a mean AUROC of 0.8790 across the 17 scored sessions, each at $p < 0.005$, label-shuffle controls 0.48–0.52 (Appendix U, Figure S3). The labeled subset is not a random sample of the stream, and every recovery number here is a statement about identifier-leaking clients.

6 Discussion and Conclusion

Boundaries are a fragility, not a constraint. Pooling orders into blocks reduces variance and raises the clustered unit's dimensionality, and in the random and contiguous constructions tested here buys nothing. Every order in a block inherits the block's one cluster, so an incorrect block has no downstream fix: wrong blocks land far below per-order clustering in every session under both corruption schemes, while correct blocks gain little on average and inconsistently across assets. Cardinality is the other half: density-based selection collapses on hand features and not on the learned embedding, specific to that estimator, and the near-equal mean count is cancellation, not recovery.

None of the three defenses closes the behavioral channel, and each acts as a step function. The three share a mechanism: each deletes or manufactures a discrete feature class rather than turning a continuous dial, so none traces a graded privacy–utility frontier. Under the oracle-K, segmentation-given threat model built here, the behavioral channel survives all three. Coarsening had the least to remove: the registered falsification had already retired the sub-millisecond channel. What jitter and the size grid delete is narrower than anonymity. The grid's protection does not reach a learned-embedding attacker clustering at an estimated actor count, so the negative result does not rest on handing the attacker the true count. Output-side perturbation deletes fingerprint features, not identity: each success is confined to the attack that read the deleted class. The residual exposure is the behavioral structure itself.

Limits. The population limit is the tightest: labellability is predictable from behavior at 0.8790 AUROC, so the actors we score are those whose client software leaks a structured identifier, and nothing here estimates recoverability for clients with sound identifier hygiene in either direction. Two things keep that narrowing from emptying the result: the reading of the diagnostic was fixed before it ran, and within the population the signal is more than identifier-family classification. Behavior also identifies an actor across days and months far more readily than across assets, so every recovery number here is scoped twice. Other limits are of sample and of kind (Appendix T).

Conclusion. Corrupted blocks are worse than none, and the value of correct ones is heterogeneous: their gain over no blocking is not established under asset-level inference and reverses on ETH for unidentified reasons. The same behavior transfers across days far better than across assets, part of that shortfall a build-level label rather than behavior. The open problem is boundary inference at the order level, where a method must beat the no-blocking arm.

ETHICS STATEMENT

The corpus is a public broadcast feed. Every event analyzed here was transmitted by the venue to any connected client without authentication, and no access control, rate limit or contractual restriction was circumvented to obtain it. We follow the Menlo framework for research on found network data (Kenneally & Dittrich, 2012): the analysis is observational and no result is reported at the level of an identified party.

The labels are pseudonymous and machine-level. A version-1 node value identifies a stable UUID node namespace and a counter namespace identifies a process or installation, so an actor in this paper is a piece of software. We make no attempt to attribute any actor to a firm or an individual and we hold no auxiliary data that would allow it, but a machine identifier is not thereby free of human consequence: a persistent behavioral fingerprint can expose an operator's strategy or infrastructure to anyone who can link it to a name, and we treat the actors as data subjects for release purposes even though they are processes.

The channel used to construct labels is already closed: the venue stopped broadcasting the client identifier between our two capture eras (zero occurrences across the 2023 captures), so the identifier structure this paper rests on cannot be read from the live feed. The channels we measure are not closed by that change; behavioral shape and arrival coincidence both remain, the latter at 4.51–4.73% of received events in the three post-mitigation sessions (Appendix X). Our defense analysis is a counterfactual replay on the labeled population, not an evaluation of a deployed defense, and its lesson for a venue is procedural: a timing perturbation must be evaluated against an attacker allowed to refit on the defended stream, because an attacker frozen on clean data looked defended twice in our replay while adaptation recovered much of the loss (Section 5, Appendix Y). At the magnitudes we could build, none of coarsening, jitter or size gridding protects against the hand-feature attacks measured here, and the only learned attacker we tested was frozen and non-adaptive, so nothing here measures a defense against an attacker that adapts its representation. Two perturbations still address narrower harms: jitter of at least 1 ms deletes the exact-timestamp coincidence channel, and a 1% size grid deletes the size-precision fingerprint that feature-engineered attacks read, at 0.29% mean relative size error. Neither buys behavioral anonymity or carries a formal guarantee such as differential privacy (Dwork et al., 2006), and we do not present them as one. No coordinated-disclosure exchange with the venue is documented as of submission; the construction channel is already withdrawn from the live feed, and the residual channels are properties of a public feed's timestamp resolution rather than of a defect a single operator can patch.

The release rule is concrete. We release the derivation and evaluation code, a manifest of the consumed capture files with a SHA-256 digest for each, hashed actor labels, and a small quantized sample for smoke testing. We do not release derived behavioral features, even with labels hashed, because exact sizes, prices and timestamps permit a re-join against the public raw feed and hence re-derivation of the identifier field for the capture window. Withholding raw captures and joinable features reduces linkage risk but does not remove it: the released methods could still inform attacks on feeds that continue to broadcast structured identifiers.

REPRODUCIBILITY STATEMENT

The analysis follows a preregistered protocol (Appendix K). All random seeds are fixed and listed in Appendix H. Exact rerunning requires the historical captures, which cannot be redistributed. Code and instructions are provided in the supplementary material.

AI USE STATEMENT

Generative AI tools were used to assist with retrieving and discovering related literature, code implementation, writing and polishing. The research ideas originated with the authors. The authors reviewed and verified AI-assisted content and take full responsibility for the paper's accuracy, originality, and conclusions.

REFERENCES

Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to Identity-Level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2): 1–29, 2008. doi: 10.1145/1344411.1344413.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2008. doi: 10.1007/s10791-008-9066-8.

Kęstutis Baltakys, Juho Kanniainen, and Frank Emmert-Streib. Multilayer aggregation with statistical validation: Application to investor networks. *Scientific Reports*, 8(1), 2018. doi: 10.1038/s41598-018-26575-2.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572. IEEE, 2016. doi: 10.1109/CVPR.2016.173.

Sanjit Bhat, David Lu, Albert Kwon, and Srinivas Devadas. Var-CNN: a Data-Efficient website fingerprinting attack based on deep learning. *Proceedings on Privacy Enhancing Technologies*, 2019(4):292–310, 2019. doi: 10.2478/popets-2019-0070.

Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1):5, 2007. doi: 10.1145/1217299.1217304.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, 2020.

Giovanni Cherubin, Rob Jansen, and Carmela Troncoso. Online website fingerprinting: Evaluating website fingerprinting attacks on Tor in the real world. In *Proceedings of the 31st USENIX Security Symposium*. USENIX Association, 2022. URL `https://www.usenix.org/conference/usenixsecurity22/presentation/cherubin`.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

Carole Comerton-Forde and Kar Mei Tang. Anonymity, liquidity and fragmentation. *Journal of Financial Markets*, 12(3):337–367, 2009. doi: 10.1016/j.finmar.2008.12.001.

Carole Comerton-Forde, Alex Frino, and Vito Mollica. The impact of limit order anonymity on liquidity: Evidence from paris, tokyo and korea. *Journal of Economics and Business*, 57(6): 528–540, 2005. doi: 10.1016/j.jeconbus.2005.05.001.

Rama Cont, Mihai Cucuringu, Vacslav Glukhov, and Felix Prenzel. Analysis and modeling of client order flow in limit order markets. *Quantitative Finance*, 23(2):187–205, 2023. doi: 10.1080/14697688.2022.2150282.

Kyle Crichton, Lorrie Faith Cranor, and Nicolas Christin. Rethinking fingerprinting: An assessment of behavior-based methods at scale and implications for web tracking. *Proceedings on Privacy Enhancing Technologies*, 2025(4):795–811, 2025. doi: 10.56553/popets-2025-0158.

K. Davis, B. Peabody, and P. Leach. Universally unique IDentifiers (UUIDs). Report, RFC Editor, 2024.

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1), 2013. doi: 10.1038/srep01376.

Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex Sandy Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221): 536–539, 2015. doi: 10.1126/science.1256297.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pp. 3830–3834. ISCA, 2020. doi: 10.21437/Interspeech.2020-2650.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564. ACM, 2016. doi: 10.1145/2939672.2939875.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Lecture Notes in Computer Science*, pp. 265–284. Springer Berlin Heidelberg, 2006. doi: 10.1007/11681878_14.

Peter Eckersley. How unique is your web browser? In *Lecture Notes in Computer Science*, pp. 1–18. Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-14527-8_1.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

Thierry Foucault, Sophie Moinas, and Erik Theissen. Does anonymity matter in electronic limit order markets? *Review of Financial Studies*, 20(5):1707–1747, 2007. doi: 10.1093/rfs/hhm027.

Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE Transactions on Information Forensics and Security*, 8(1):136–148, 2013. doi: 10.1109/TIFS.2012.2225048.

Mostafa Goudarzi and Flavio Bazzana. Identification of high-frequency trading: A machine learning approach. *Research in International Business and Finance*, 66:102078, 2023. doi: 10.1016/j.ribaf.2023.102078.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Björn Hagströmer and Lars Nordén. The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770, 2013. doi: 10.1016/j.finmar.2013.05.009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.

Marc Juarez, Sadia Afroz, Gunes Acar, Claudia Diaz, and Rachel Greenstadt. A critical evaluation of website fingerprinting attacks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 263–274. ACM, 2014. doi: 10.1145/2660267.2660368.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Erin Kenneally and David Dittrich. The menlo report: Ethical principles guiding information and communication technology research. Report, Elsevier BV, 2012.

Kevin S. Killourhy and Roy A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pp. 125–134. IEEE, 2009. doi: 10.1109/DSN.2009.5270346.

Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-Frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017. doi: 10.1111/jofi.12498.

T. Kohno, A. Broido, and K.C. Claffy. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing*, 2(2):93–108, 2005. doi: 10.1109/TDSC.2005.26.

P. Leach, M. Mealling, and R. Salz. A universally unique IDentifier (UUID) URN namespace. Report, RFC Editor, 2005.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, 2019.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60, 2020. doi: 10.14778/3421424.3421431.

Vasilios Mavroudis. Market manipulation as a security problem, 2019. arXiv:1903.12458.

Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. A fistful of bitcoins. In *Proceedings of the 2013 conference on Internet measurement conference*, pp. 127–140. ACM, 2013. doi: 10.1145/2504730.2504747.

Albert J. Menkveld. High frequency trading and the new market makers. *Journal of Financial Markets*, 16(4):712–740, 2013. doi: 10.1016/j.finmar.2013.06.006.

Markus Miettinen, Samuel Marchal, Ibbad Hafeez, N. Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. IoT SENTINEL: automated Device-Type identification for security enforcement in IoT. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2177–2184. IEEE, 2017. doi: 10.1109/ICDCS.2017.283.

Fabian Monrose and Aviel D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4):351–359, 2000. doi: 10.1016/S0167-739X(99)00059-X.

Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. Deep learning for entity matching. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 19–34. ACM, 2018. doi: 10.1145/3183713.3196926.

Federico Musciotto, Luca Marotta, Jyrki Piilo, and Rosario N. Mantegna. Long-term ecology of investors in a financial market. *Palgrave Communications*, 4(1), 2018. doi: 10.1057/s41599-018-0145-1.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a Large-Scale speaker identification dataset. In *Interspeech 2017*, pp. 2616–2620. ISCA, 2017. doi: 10.21437/Interspeech.2017-950.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008. doi: 10.1109/SP.2008.33.

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187. IEEE, 2009. doi: 10.1109/SP.2009.22.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of Internet-Scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pp. 300–314. IEEE, 2012. doi: 10.1109/SP.2012.46.

Milad Nasr, Alireza Bahramali, and Amir Houmansadr. DeepCorr: strong flow correlation attacks on tor using deep learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1962–1976. ACM, 2018. doi: 10.1145/3243734.3243824.

Maksym Nechepurenko. Fill-side behavioral concentration on Polymarket: Identification limits under record-level attribution, 2026. arXiv:2605.11640.

Marcos Oliveira, Junran Yang, Daniel Griffiths, Denis Bonnay, and Juhi Kulshrestha. Browsing behavior exposes identities on the web. *Scientific Reports*, 15(1), 2025. doi: 10.1038/s41598-025-19950-3.

Andriy Panchenko, Fabian Lanze, Andreas Zinnen, Martin Henze, Jan Pennekamp, Klaus Wehrle, and Thomas Engel. Website fingerprinting at internet scale. In *Proceedings 2016 Network and Distributed System Security Symposium*. Internet Society, 2016. doi: 10.14722/ndss.2016.23477.

Vera Rimmer, Davy Preuveneers, Marc Juarez, Tom Van Goethem, and Wouter Joosen. Automated website fingerprinting through deep learning. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018. doi: 10.14722/ndss.2018.23105.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-019-10933-3.

Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, 17(134):1–32, 2016.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Yuki Sato and Kiyoshi Kanazawa. Inferring microscopic financial information from the long memory in Market-Order flow: A quantitative test of the Lillo-Mike-Farmer model. *Physical Review Letters*, 131(19), 2023a. doi: 10.1103/PhysRevLett.131.197401.

Yuki Sato and Kiyoshi Kanazawa. Quantitative statistical analysis of order-splitting behavior of individual trading accounts in the japanese stock market over nine years. *Physical Review Research*, 5(4), 2023b. doi: 10.1103/PhysRevResearch.5.043131.

W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823. IEEE, 2015. doi: 10.1109/CVPR.2015.7298682.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations (ICLR)*, 2020.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Payap Sirinam, Mohsen Imani, Marc Juarez, and Matthew Wright. Deep fingerprinting. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1928–1943. ACM, 2018. doi: 10.1145/3243734.3243768.

Payap Sirinam, Nate Mathews, Mohammad Saidur Rahman, and Matthew Wright. Triplet fingerprinting: More practical and portable website fingerprinting with n-shot learning. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1131–1148. ACM, 2019. doi: 10.1145/3319535.3354217.

Arunan Sivanathan, Hassan Habibi Gharakheili, Franco Loi, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. Classifying IoT devices in smart environments using network traffic characteristics. *IEEE Transactions on Mobile Computing*, 18(8):1745–1759, 2019. doi: 10.1109/TMC.2018.2866249.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4077–4087, 2017.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333. IEEE, 2018. doi: 10.1109/ICASSP.2018.8461375.

Shivashankar Subramanian, Daniel King, Doug Downey, and Sergey Feldman. S2AND: A benchmark and evaluation system for author name disambiguation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021.

Michele Tumminello, Fabrizio Lillo, Jyrki Piilo, and Rosario N Mantegna. Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, 14(1): 013041, 2012. doi: 10.1088/1367-2630/14/1/013041.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. arXiv:1807.03748.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022b.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. URL https://www.jmlr.org/papers/v11/vinh10a.html.

Tao Wang, Xiang Cai, Rishab Nithyanand, Rob Johnson, and Ian Goldberg. Effective attacks and provable defenses for website fingerprinting. In *Proceedings of the 23rd USENIX Security Symposium*. USENIX Association, 2014. URL https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/wang_tao.

Mateusz Wilinski, Anubha Goel, Alexandros Iosifidis, and Juho Kanniainen. Classifying and clustering trading agents, 2025. arXiv:2505.21662.

Isaac D. Wright, Matthew Reimherr, and John Liechty. A machine learning approach to classification for traders in financial markets. *Stat*, 11(1), 2022. doi: 10.1002/sta4.465.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, 2016.

Siqiao Xue, Xiaoming Shi, Zhixuan Chu, Yan Wang, Hongyan Hao, Fan Zhou, Caigao Jiang, Chen Pan, James Y. Zhang, Qingsong Wen, Jun Zhou, and Hongyuan Mei. EasyTPP: Towards open benchmarking temporal point processes. In *International Conference on Learning Representations (ICLR)*, 2024.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person Re-Identification: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Yichi Zhang, Mihai Cucuringu, Alexander Y. Shestopaloff, and Stefan Zohren. ClusterLOB: Enhancing trading strategies by clustering orders in limit order books, 2025. arXiv:2504.20349.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124. IEEE, 2015. doi: 10.1109/ICCV.2015.133.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 11692–11702, 2020.

A THE AUDIT GATES, THEIR DEFECTS AND THEIR REPAIRS

Blocking gates assert construction: read-sets, split membership, manifest contents and wiring, all deterministic functions of how a run was built. A check whose failure condition is indistinguishable from a scientific finding is a diagnostic instead, always reported and never suppressive. Two of our own gates violated that rule. The regime-separation gate required the two regimes to differ by at least 0.01 on every task, which is the hypothesis under test wearing a gate's clothing. A residual-resolution probe, classifying actors from the sub-quantum remainder $t \bmod q$, has the same defect, since its firing would mean a per-actor clock fingerprint had been found.

The first was replaced by a construction-based successor asserting that the native-resolution feature manifest is a strict superset of the quantized one, that the difference set is exactly the declared metadata features, that at least one has non-zero variance in the evaluation data, and, as a reported diagnostic rather than a gate, that at least one takes a split in the fitted model. The split-gain clause was demoted for the same reason as its predecessor: it fails precisely when the model finds every metadata feature useless, which is the registered prediction P01 confirmed. The variance clause was corrected from a universal to an existential quantifier, a loosening that makes the two largest negative excursions in Δ_M count toward P01's verdict rather than being excluded from it. The residual-resolution probe was reclassified as an escalating diagnostic although it passed, which is what shows the rule is not applied in the direction of a preferred result; it reads 0.0261 on BTC against a threshold of 0.0571 and 0.0199 on ETH against 0.0556.

The second rule is that every blocking gate declares the runs on which it must evaluate substantively, and an assertion verifies that the set was covered. We adopted it because a gate enforcing the timestamp mask matched a variant name against a caller-supplied composite key of the form `session/variant`, never matched, returned not-applicable on all ten variant-runs of a scored experiment, and declared itself non-blocking on that branch. The adjacent read-set gate held the identifier fields in its forbidden set and not the timestamp, so it did not cover the gap. The mask had probably been applied, and probably is not verification, so the run was voided in full. The repair deleted the name matching, folded the check into the read-set gate as a regime-conditional forbidden set with the regime passed as an explicit argument, and added the coverage assertion with a tripwire that halts on any not-applicable result inside a declared coverage set. A referee is entitled to ask how we know our audits ran, and the only answer worth giving is a mechanism.

B NUMERICAL VALIDITY AS A SEPARATE AUDIT PRECONDITION

Leakage and construction gates assert provenance: which raw fields a feature reads, whether splits are disjoint, whether one feature manifest is a superset of another. None of those is a statement about the values an admitted feature takes, so all of them passed on a feature matrix in which 20 of 61 columns in the quantized manifest and 20 of 67 in the native one, about 30% either way, were uninitialized memory. The cause was a shape mismatch between per-trace and per-order inputs: only the first n_{traces} entries of each output were written, leaving 31 of every 32 entries uninitialized. The defect surfaced not through the audit but through non-reproducibility, the same matrix built twice from identical inputs inside one process returning different values, together with a raw feature value of -5.88×10^{305}.

Two assertions now run as preconditions of a scored run. Every feature matrix is built twice and asserted bitwise identical, which closes the non-deterministic case; and each feature is asserted finite and within declared physical bounds, because a wrong-but-stable array passes the determinism check. The affected experiments, E1 and T1a with everything derived from them, were rebuilt and rerun in full, and the re-run reports old against new values with an explicit statement of whether any preregistered verdict changed.

C THE SUPERVISED TRACE TASK AND THE CHANNEL DECOMPOSITION

variant	BTC (K=29)	ETH (K=30)	mean
all	0.7367	0.6761	0.7064
rate-free	0.7255	0.6734	0.6995
schedule-free	0.7254	0.6769	0.7012
rate- & schedule-free	0.7325	0.6948	0.7137
rate-only	0.1178	0.1670	0.1424
FULL, native μs	0.7359	0.6917	0.7138
labels shuffled	0.0217	0.0240	0.0228
chance	0.0345	0.0333	0.0339

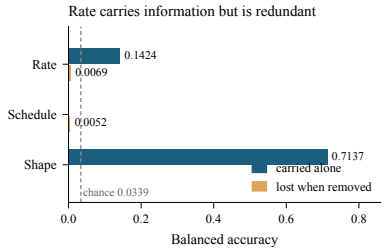


Figure S1: With segmentation given, behavioral shape carries the identity. Left: closed-set trace classification, balanced accuracy, BTC and ETH 2022-05-16, temporal 70/30 split, sibling-collapsed labels. Right: the decomposition those variants yield, a total being what a channel carries alone and a marginal what its removal costs.

D ORDER-LEVEL ERROR DECOMPOSITION

Order-level errors on the T1b slice (BTC 2022-05-16, 500,000 orders, 39 actors) follow the diarization-style decomposition, with correct as the remainder. Without segmentation, k-means at oracle K is 26.4% correct with 41.9% confusion, 31.8% fragmentation and no missed orders. HDBSCAN is 3.2% correct with 15.6% confusion, 61.3% fragmentation and 19.9% miss. The co-incidence union-find is 0.02% correct with 0.002% confusion and 99.98% fragmentation. With the segmentation oracle given, the behavioral arm is 43.4% correct with 17.3% confusion, 39.2% fragmentation and 0.1% miss, and the native-microsecond arm 43.3% correct with 15.5% confusion, 41.0% fragmentation and 0.1% miss. The two regimes therefore trade confusion for fragmentation rather than differing in how much they recover.

E IDENTIFIER FAMILIES AND THE IMPOSTOR SEPARATION

Table S1: Identifier families: label validity and within-family actor identification. Pair purity at a 1 μs quantum is 1.000 in all six labellable families, so the informative quantity is the lift over each family's own chance baseline, pooled over the 18 identifier-bearing sessions. Merge-flagged is the share of actor-sessions raising a block-level bimodality flag, with `uuid_v1` as the false-positive baseline and the battery detecting 38% of 50 synthetic known-merges. Within-family accuracy holds family fixed, BTC and ETH 2022-05-16.

		purity	merge-	within-family actor task			
family	actors	lift	flagged	K	chance	BTC	ETH
`const_prefix16`	74	9.32×	10.1%	14	0.0714	0.5107	0.4274
`date_counter`	10	7.66×	29.1%	—	—	—	—
`uuid_v1`	24	1.98×	11.7%	8 / 9	0.1250 / 0.1111	0.9688	0.9981
`zero_mid_node`	42	1.73×	2.1%	—	—	—	—
`zero_counter`	2	1.01×	23.5%	2	0.5000	1.0000†	1.0000†
`ascii_filler`	1	1.00×	42.9%	—	—	—	—

† Two actors, of which the minority contributes 5–21 test blocks; the bootstrap interval is degenerate and the figure should not be read as a reliable 1.0.

A 16-hex head is admitted to the constant-prefix family only if it recurs at least 4 times inside one part file and at least 64 times corpus-wide; 83 candidates are tested and 74 admitted, and an 8-hex variant of the rule adds none.

One ordered rule chain assigns each identifier-bearing event to exactly one of ten layouts, first match wins, so the families partition the events and their counts sum to the total by construction; agreement between families is therefore not a checkable quantity. Of the 160,185,756 events carrying the field, `uuid_v4` holds 44.82% and `random_hex` 9.93%, neither of which is labellable. The six labellable families are `uuid_v1` (17.91%, 24 actors), `zero_counter` (13.85%, 2), `const_prefix16` (5.70%, 74), `zero_mid_node` (3.44%, 42), `date_counter` (1.90%, 10) and `ascii_filler` (1.53%, 1). The 8-hex variant of the constant-prefix rule added no actors and was not adopted.

The node-bit provenance test behind the impostor separation reads as follows. The unlabeled random population sits at 46.7–49.1% on both the multicast and the locally administered bit in every asset. The BTC and ETH populations that fail the embedded-clock test match it at 47.5–48.4%, while SOL and AVAX deviate at 20.5% and 27.2% multicast and are a mixture, so there the one-event-per-node structure carries the argument on its own. Real version-1 actors differ from both: their bit fractions are per-node constants.

F SIBLING GROUPS AND ACTIVATION SYNCHRONY

Activation time is used by neither construction rule, so it tests the sibling construct rather than restating it. Within a confirmed group the standard deviation of first-activity time is 0.10–3.96 s, against null medians of 18,402–32,205 s drawn from random same-size subsets of the same session's actors at the same density floor; all 14 confirmed group-sessions are left-tail significant at $p < 0.05$ over 2,000 permutations with seed 20260803. Across the seven confirmed group-days present in both BTC and ETH, the two markets' group activation means separate by 0.26–15.53 s with a median of 0.90 s, five of the seven within two seconds and the remaining two at 5.14 s and 15.53 s. A further 83 of 587 suspected-pair tests are significant as well, which is the price of refusing to collapse on one rule alone: real siblings remain uncollapsed and the instance-level task is harder than entity re-identification.

G THE TASK SUITE AND PROTOCOL DETAILS

Beyond the two primary tasks, the suite specifies six. T2 is closed-set classification of traces under a temporal within-session split. T3 admits or rejects actors absent from the enrolled set, scored against both a random holdout of known actors and the actors genuinely unseen in a session's training window (Scheirer et al., 2013; Vaze et al., 2022b). T4 and T5 retrieve a query trace against a gallery built on a held-out day and on held-out assets, with the roughly one-month capture gap as the long-horizon case, one prototype per gallery actor and every query trace ranked by Euclidean distance on gallery-frozen standardized features. T6 enrolls a novel actor from k traces (Snell et al., 2017). T7 reports the observation budget, how much of an actor's activity an observer needs, and the defense budget, how much timestamp coarsening or noise removes the signal, the latter against a retrained as well as a fixed attacker (de Montjoye et al., 2013; Wang et al., 2014; Cherubin et al., 2022). T8 attributes identifier family rather than actor over every event carrying the field.

The seven variants of the supervised closed-set experiment are all features, rate-free, schedule-free, rate- and schedule-free, rate-only, the native-microsecond regime, and a label-shuffle control. T1a runs under three cluster-number protocols: oracle K, the spectral eigengap, and density-based estimation. Traces are sampled without replacement to at most 200 per actor-session under seed 20260803, and the recorded reproduction metadata include the seed, feature manifest, gate results, and corresponding metrics.

H POPULATIONS, SPLITS AND THE LABEL SPACE

Supervised splits are temporal within a session: the first 70% of events train and the last 30% evaluate, with a guard band of at least an hour between them. The closed-set label-space rule is

a definitional repair rather than an exclusion. A class absent from the training window has recall identically zero by construction rather than by failure, so including it in a macro-recall metric applies a multiplicative deflation of $K_{\text{learnable}}/K_{\text{total}}$ that measures nothing. The rule is applied identically in every session, regime and variant, and the actors it removes are enumerated in the released metrics with their order counts, their first and last activity times and the split boundary, so a reader can verify that the exclusion is structural. They are additionally routed to the open-set unknown pool, which is where an actor with no training presence belongs, and retained in an unrestricted variant of every closed-set metric that is reported as a lower bound.

Confirmed sibling groups are assigned to training or to held-out atomically, since an unknown actor that is a near-twin of a known one makes open-set AUROC meaningless. For closed-set tasks the group is admitted to the scored label space only if every member satisfies both presence conditions, and otherwise the whole group is routed to the unknown pool.

In the supervised experiment the rule admits 35 BTC instances collapsing to 29 and 36 ETH instances collapsing to 30, excluding 23 and 16 at a split boundary of 16.8 and 17.8 h. The primary figure is flat across training-presence thresholds of 1, 2, 4 and 8 traces, at 0.737, 0.737, 0.737 and 0.713 on BTC.

The density floor is reported at 128, 256, 512 and 1024 events regardless of outcome, so that the choice of floor can be checked rather than trusted. Metric choices follow the nearest well-posed benchmarks rather than local preference: ARI (Hubert & Arabie, 1985) with B-cubed precision, recall and F, the only extrinsic measure satisfying the formal constraints that imbalanced partitions break (Amigó et al., 2008; Subramanian et al., 2021). AMI rather than raw NMI where cluster counts differ (Vinh et al., 2010; Romano et al., 2016); AUROC with max-logit as the mandatory open-set baseline (Vaze et al., 2022b); and the person re-identification query and gallery protocol for retrieval (Zheng et al., 2015; Ye et al., 2022). Cluster counts are never reported at oracle K alone.

I Regime differences are read against the Bland–Altman mean

Regressing $\Delta_M = \text{FULL} - \text{BEHAV}$ on BEHAV places BEHAV on both sides of the equation and produces a negative correlation under a true null, which is Oldham's fallacy. We therefore read every regime difference against the mean of the two arms. On these data the apparent $r = -0.7107$ falls to -0.4724 at $p = 0.065$ on the Bland–Altman mean, and to -0.2336 at $p = 0.384$ in the tie-rate representation we report. The reading that metadata hurts where behavior already clusters well is not adopted.

J Session-level heterogeneity in Δ_M

Δ_M is not one number applied unevenly across sessions, and the account for the dispersion is incomplete. A metadata feature with zero variance contributes exactly zero, which is a theorem rather than a measurement, and the formal test of the degeneracy account gives a slope of $+0.0065$ per decade of feature variance at a one-sided p of 0.08: the direction is right and the magnitude is not significant. Two sessions of near-identical variance sit about 0.010 apart, so a second mechanism operates at the session level and we have not identified it. Removing the degeneracy at its source, by re-expressing the metadata features as rates rather than sizes, did not remove the negative excursions: three of the four AVAX sessions do not move to four decimals, the two SOL sessions improve without reaching zero, and five of the six sessions in the affected set stay negative, as they were before. That weakens the degeneracy account, and we report the dispersion as heterogeneous.

Two registered reports could not be produced as specified. The singleton-preservation rate is undefined on the held-out session, which contains no one-order actor and whose smallest holds 49 orders. The BCa bootstrap degenerates to a percentile interval on several pairwise-agreement metrics, flagged per entry in the released results.

K Protocol timeline

The protocol is append-only and its entries are of four kinds, distinguished here because they carry different evidential weight.

entry	result visible when written	rule changed	prospective for
base protocol	none	–	every experiment
amendments 1–4	none admissible	yes	every experiment
amendment 5	T1a	yes	T1b, P02
amendments 6–7	E1, T1a (both later voided)	yes	the rebuilt E1, T1a and all successors
outcome record 1	T1b	no	–
amendment 8	the 16-session factorial	no	–
outcome record 2	the 16-session factorial	no	–
amendment 9	none of what it decides	yes	the family-sensitivity analysis
amendment 10	D1 FULL-all firings visible. BEHAV re-run and seed ensemble not yet run	yes	D1 BEHAV re-run, seed ensemble
amendment 11	D2 drift and D3 both visible. Rule fixed pre-D3 in the logged D2 adjudication	yes	D3

Amendments 1–4 are prospective corrections written while no admissible model outcome existed. One scored run had been voided by then and none of its numbers survives. Amendment 5 was written after T1a completed and before T1b, whose data decides P02, was run. Amendments 6 and 7 respond to the uninitialized-memory defect of Appendix B and to a premise that defect invalidated, and every experiment they govern was rebuilt and rerun afterwards. The first outcome record changes no rule; it documents the execution of the P02 falsifier registered in amendment 4.

Amendment 8 registers nothing either, and says so in its own status field. It discloses that no contrast in the factorial was ever registered, the factorial itself having been built to answer a review objection after the deciding data, and that the decision to cluster the cross-session inference by asset was taken after the assets were seen to separate. The second outcome record documents a rule fixed in the run brief before the 16-session factorial executed: a gain near zero or frequently reversing would withdraw the paper's title claim. The rule triggered and the claim was withdrawn.

Amendment 9 registers the family-sensitivity analysis of Appendix O: its conditions, its decision rules on each contrast and on the cross-asset chance multiple, the ban on comparing ARI levels between conditions, and the direction each outcome would move the paper. No exclusion condition and no per-family contrast had been computed when it was written. One inference it now reports, the significance of A4 − A1 on the population without the two degenerate families, falls outside the branch it registered, and is labeled exploratory where it appears.

L THE SUPERVISION AXIS

A classifier's predicted labels are a partition, so supervised and unsupervised recovery take the same metric against the same ground truth. Measured on one session (BTC 2022-05-16, BEHAV regime), with the full training window carrying 140.3 labels per actor:

labels per actor	0	1	2	4	8	16	32	full window
ARI	0.3067	0.2056	0.2428	0.2999	0.3330	0.3840	0.4388	0.5645
B-cubed F1	0.4035	0.3022	0.3535	0.3853	0.4626	0.4458	0.4644	0.6110

The axis was measured on one session and its shape elsewhere is unknown.

M THE READ-SET CONTRACT AND THE ORACLE WE DECLINE TO USE

Feature functions never receive a forbidden field, because the loader builds a restricted record view from which those fields are physically absent rather than filtering afterwards. The forbidden set is

the client identifier, the venue-assigned order identifier, the sequence number, any contamination between splits and, in the quantized regime, the sub-quantum remainder of the timestamp.

The clock embedded in a version-1 UUID deserves its own line. The difference between receipt time and embedded time is client-to-venue latency plus clock offset, a per-machine fingerprint of considerable power, and it is available to us because we can read the identifier. It is therefore a leak rather than a channel, and it is excluded everywhere. That a stream carries such an oracle and a benchmark declines to use it is the difference between measuring what an observer of the feed can do and measuring what a privileged reader can do.

The coincidence relation is a channel by the same test: in the five 2023 sessions the client identifier is absent from the feed altogether and the coincidence structure is still present. Devices have been fingerprinted from clock skew (Kohno et al., 2005) and flows linked by packet timing (Nasr et al., 2018); both estimate a statistic over a long window across two observation points, whereas this relation is a discrete coincidence between two specific events at a single observation point.

N THE SEGMENTATION FACTORIAL

Tables S2 and S3 give the values behind Figure S2 at $W = 32$, over 16 sessions covering BTC, ETH, SOL and AVAX on 2022-05-14 to 2022-05-17. Every arm shares one order set of 500,000 orders per session, one block featurizer (per-feature mean and standard deviation), one standardization, k-means at oracle K and one metric, so the only quantity that varies within a session is where a block boundary comes from. A1 applies the same clustering to individual orders. Oracle K is supplied to every arm, so cardinality supervision is held constant and is not part of the measured contrast.

Table S2: Controlled segmentation factorial at $W = 32$, 16 sessions: differences in order-level ARI. No contrast here was preregistered: the factorial was built to answer a review objection. The two that corrupt the boundaries hold their sign in all 16 sessions under two distinct corruption constructions; A4 − A1 does not.

contrast	mean	sessions positive	range
A2 contiguous − A1 no blocking	-0.1714	0/16	$[-0.2339, -0.0839]$
A3 random − A1 no blocking	-0.1749	0/16	$[-0.2383, -0.0868]$
A4 oracle − A1 no blocking	$+0.0233$	12/16	$[-0.0664, +0.0753]$
A4 oracle − A3 random	$+0.1982$	16/16	$[+0.1308, +0.2574]$

Table S3: A4 − A1 by asset at $W = 32$, four sessions each. The assets separate almost completely, and baseline recoverability does not order the reversals: mean no-blocking ARI is 0.2003 on SOL, 0.1990 on ETH and 0.1890 on AVAX, similar observed means, yet only ETH reverses.

asset	mean A4 − A1	within-asset range	reversals
BTC	$+0.0685$	$[+0.0628, +0.0753]$	0/4
AVAX	$+0.0412$	$[+0.0273, +0.0546]$	0/4
SOL	$+0.0306$	$[+0.0179, +0.0545]$	0/4
ETH	-0.0472	$[-0.0664, -0.0377]$	4/4

Between-asset variance accounts for 94.2% of the variance in A4 − A1, which is why the inference in Section 5 is clustered at the asset level rather than at the session level. That choice was made after the separation was observed and is post-hoc; it is recorded as such in the protocol amendment, together with the direction it moves the result, which is against the claim we had been making. The mean gain by window is -0.0408, -0.0077, -0.0059, $+0.0095$ and $+0.0233$ at $W = 2, 4, 8, 16, 32$, so the exploratory contrast is positive only at the two largest windows; $W = 32$ is the primary window of the base protocol, fixed there before any of these results existed, and the profile is reported whichever way it falls.

On BTC 2022-05-16 the A4 − A3 contrast carries a bootstrap interval over 1,000 resamples of $[+0.1789, +0.1801]$. Under actor subsampling in the same session it is $+0.3851$ at $K = 8$ over 38,625 orders, $+0.3251$ at $K = 16$ over 138,603 orders and $+0.2081$ at $K = 32$ over 467,072 orders, so it decays as the identity set grows and stays large across the range these data cover.

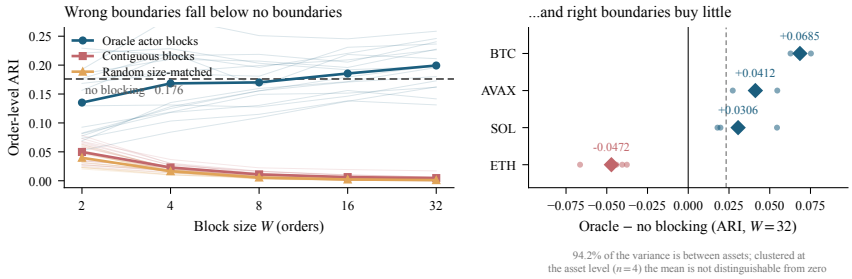


Figure S2: The segmentation factorial in full, 16 sessions. Left: order-level ARI against block size W for the three imposed-boundary arms against the no-blocking line, the panel reproduced in the main text. Right: the oracle-minus-no-blocking contrast at $W = 32$, one point per session and a diamond at each asset's mean, grouped by asset because 94.2% of its variance lies between assets; ETH is the one asset that reverses, and clustered at the asset level ($n = 4$) the mean is not distinguishable from zero.

O FAMILY SENSITIVITY OF THE FACTORIAL AND OF RETRIEVAL

This analysis was registered as a protocol amendment before any of its exclusion conditions, per-family contrasts or null replicates were computed – the baseline factorial and the composition scan it is calibrated against did already exist – and ran under 448 leakage gates, all passed. Each exclusion condition deletes the named identifier families from the session in place and rebuilds all four arms on the retained orders at the post-exclusion oracle K, so every contrast is formed inside one condition whose arms share an order set, a featurizer, a seed and a K. Dropping a family changes K and the cluster-size distribution, so ARI levels are not comparable between conditions and are never compared here; what is read across conditions is only whether a sign pattern survives.

Table S4: The segmentation factorial under four exclusion conditions, $W = 32$, 16 sessions each; order-level ARI differences with the count of sessions positive. Each row is computed inside its own condition and the rows are not comparable in level. The first row reproduces Table S2. The last column is the one-sided 95% lower bound and p for A4 − A1 clustered at the asset level, $n = 4$.

condition	A2 − A1	A3 − A1	A4 − A3	A4 − A1	LB / p
no exclusion	−0.1714, 0/16	−0.1749, 0/16	+0.1982, 16/16	+0.0233, 12/16	−0.0351 / 0.209
without `const_prefix16`	−0.1783, 0/16	−0.1830, 0/16	+0.2396, 16/16	+0.0566, 12/16	−0.0100 / 0.069
without `uuid_v1`	−0.1417, 0/16	−0.1472, 0/16	+0.1232, 16/16	−0.0240, 6/16	−0.0813 / 0.199
without `ascii_filler` and `zero_counter`	−0.1345, 0/16	−0.1385, 0/16	+0.2097, 16/16	+0.0712, 16/16	+0.0271 / 0.016

The last row is the exploratory condition of Section 5 (Table S4): the condition itself was registered, the significance branch was registered for the `const_prefix16` exclusion alone, and three exclusions were examined, so a Bonferroni correction over them leaves p near 0.048. The two families it drops are the two the validity audit of Table S1 already reported as degenerate, one at $K = 1$ with chance purity 1.0 by construction and one in which a single actor holds 99.97% of the family. Dropping them does not validate what remains: the merge battery detects 38% of synthetic known-merges, so an absent flag on a retained family is weak evidence about that family.

ETH's A4 − A1 reversal deepens without `uuid_v1` (Table S5), where all four of its sessions reverse at a mean of −0.0794. It is mixed without `const_prefix16`, at 2 of 4 and −0.0047. It is absent without the two degenerate families, at 0 of 4 and +0.0327. Two disjoint exclusions each remove it and a third deepens it, which rules out an account resting on any one family, including one resting on `const_prefix16`: the condition that removes the reversal retains that family. All four confirmed sibling groups of seven belong to `const_prefix16`, so the condition that excludes it also takes the sibling phenomenon of Appendix F out of scope.

Retrieval was rescored the same way, from scratch, with every published headline reproduced to float precision first. Cross-asset Rank-1 against the pooled gallery of 66 actors is 0.0366, 2.42

Table S5: Per-family contrasts with no family excluded, order-level B-cubed F1, mean over 16 sessions with the count of sessions positive. B-cubed rather than ARI because ARI is not point-decomposable and a per-family restricted ARI would re-import a per-family chance baseline. These rows are descriptive: identifier families are not draws from a population, so no interval or test is attached to them. `ascii_filler` and `zero_mid_node` hold too few actors per session to be eligible.

family	A2 − A1	A3 − A1	A4 − A3
`uuid_v1` (corroborated unit)	−0.2478, 0/16	−0.2575, 0/16	+0.3773, 16/16
`const_prefix16`	−0.1349, 0/16	−0.1419, 0/16	+0.2642, 16/16
`date_counter`	−0.0373, 0/16	−0.0379, 0/16	+0.1750, 16/16
`zero_counter` (secondary)	−0.1795, 0/16	−0.1861, 0/16	+0.1590, 16/16

times chance. Excluding `const_prefix16` from gallery and query alike leaves 42 gallery actors and 5,265 scored queries at Rank-1 0.0955, 4.01 times the post-exclusion chance; restricting both sides to `uuid_v1` leaves 11 actors and 1,752 queries at 0.3156, 3.47 times. Pooled per-family cross-asset Rank-1 is 0.1640 for `const_prefix16`, 0.0100 for `zero_counter` and 0.0000 for both `uuid_v1` and `date_counter`, and the 24 `const_prefix16` gallery entries account for that pattern: deleting them raises Rank-1 hits on the unchanged 5,265 queries from 8 to 503. Day retrieval under the same exclusion returns Rank-1 0.4870, 0.5566, 0.5864 and 0.5966 against post-exclusion galleries of 22, 23, 19 and 17 actors on BTC, ETH, SOL and AVAX, 10.1 to 12.8 times chance.

P Random holdouts versus naturally unseen actors

Max-logit scoring on the closed-set classifier (Vaze et al., 2022b) reaches a mean AUROC of 0.7368 on the synthetic 30% class holdout, above chance in 16 of 16 sessions and ranging from 0.5513 to 0.8616. On natural probes in the same sessions, actors genuinely absent from the training window, it reaches 0.3599, below chance in 8 of 11 scorable sessions and ranging from 0.0363 to 0.7559. Five sessions admit no genuinely unseen actor and are reported as degenerate rather than dropped.

The gap is not an orientation error. Both variants share a single code path, so a flipped known/unknown convention would invert both rather than one, and the synthetic holdout scores above chance in all 16 sessions while the natural probe scores below it in most. Closed-set balanced accuracy on the knowns runs 0.58 to 0.74, so the underlying classifier is sound and the difference is in what the two protocols ask of it. The standard protocol is measuring something easier than the problem it stands for, which is a property of the evaluation design rather than of the method, and is visible only because this corpus carries both.

The gap is not specific to the max-logit score, though the corroboration available is weaker than three detectors would suggest. Applying an energy score to the same fitted classifiers gives 0.7350 on the synthetic holdout and 0.3446 on natural probes, a gap of $+0.3904$ against max-logit's $+0.3770$; on the composition-matched pool it gives 0.6573 against max-logit's 0.6437. Max-logit is the maximum of the logit vector and energy its log-sum-exp, so these are two functions of the same quantity from the same model, not independent detectors. A class-conditional Mahalanobis score, fitted with shared covariance on the training knowns in the classifier's own feature space, reaches 0.5440 on the synthetic pool and 0.4996 on the natural one: it separates nothing on either, so it cannot speak to a contrast between them, and we report it as uninformative in this feature space rather than as disagreement. What the evidence licenses is that the disagreement is not an artifact of a single score function; it does not license calling the finding detector-robust in general.

Composition, however, is not held fixed by that comparison. The two unknown pools are nearly disjoint by identifier family: the natural pool is 82.7% `const_prefix16` and 17.3% `zero_mid_node`, with no actor of any other family in any of the 16 sessions, while the random 30% holdout is 57.5% `date_counter` and 19.2% `uuid_v1`. Those families differ in difficulty, so the gap could in principle be composition rather than novelty.

Restricting both pools to `const_prefix16` tests that directly. The natural pool then scores a mean AUROC of 0.4107 over 10 admitted sessions and the synthetic pool 0.7201 over 40 admitted session-

fold cells, a gap of 0.3094; a pooled-score diagnostic that does not weight cells equally gives 0.4130 against 0.8332, a larger gap. Composition alone therefore does not account for the disagreement. Two caveats travel with these numbers. Cell admission required at least five unknown traces and two known actors of the family, so 6 of 16 natural sessions and 8 of 48 synthetic cells are excluded, and several admitted SOL and AVAX cells rest on only one to three known actors and produce extreme per-cell values; the unweighted mean was kept for comparability with the headline above, and the pooled-score diagnostic, which lacks that weighting, moves the result in the same direction. Cohort structure remains unseparated: BTC 2022-05-14's entire natural pool is the confirmed sibling group `dd16ef01-07`, which activates together on the following day, so a real absence is a cohort coming online rather than an independent draw.

A composition-matched holdout addresses that reading directly. For each session the synthetic pool is drawn to match the size and the family composition of that session's own natural pool rather than as a uniform 30%, over five seeds, with confirmed sibling groups drawn atomically and no hyperparameter retuned. Six of the eleven natural-scorable sessions admit such a draw (Table S6); the other five fail on family-count deficiency, the admitted knowns of a required family being too few to supply a matched pool of the natural size. Pooled over the six, the matched holdout reaches 0.6437 against 0.3700 for the natural pool in the same sessions, a gap of 0.2738 (positive under every leave-one-session-out drop, $+0.1542$ to $+0.3680$; Table S7). Matching size and composition does not close the gap either.

Table S6: The composition-matched holdout, per session. Pool is the number of actors in that session's natural unknown pool, which the matched draw reproduces in size and family composition. Matched is the mean over five seeds; natural is the published AUROC for the same session. Draws is the number of distinct held-out sets the five seeds produce: where sibling atomicity leaves one feasible subset it is 1, and the seeds are then not independent replicates.

session	pool	matched	natural	gap	draws / 5
BTC 2022-05-14	7	0.9973	0.1256	$+0.8716$	1
BTC 2022-05-15	11	0.8334	0.3231	$+0.5103$	5
ETH 2022-05-14	7	0.6502	0.3405	$+0.3097$	1
ETH 2022-05-15	10	0.5021	0.2753	$+0.2268$	5
SOL 2022-05-15	2	0.4676	0.5463	-0.0786	5
AVAX 2022-05-15	2	0.4117	0.6089	-0.1972	4

The base is narrow and four features of it belong beside that number. Only 6 of the 11 natural-scorable sessions are measurable at all. The two sessions that reverse are exactly the two whose natural pool holds two actors, where an AUROC over two unknowns is not estimable and the seed spread is widest, at standard deviations of 0.128 and 0.220; restricting to pools above two actors leaves 4 of 4 positive at a mean gap of 0.4796, but that restriction was chosen after seeing which sessions reversed and is post-hoc rather than the headline. Two of the four positive sessions have exactly one feasible matched draw under sibling-group atomicity, so their zero standard deviation across seeds is structural rather than sampled. The result therefore rests on BTC 2022-05-15 and ETH 2022-05-15, the two sessions carrying both an adequate pool and independent draws, with two determined-draw sessions alongside them.

A leave-one-session-out sweep of the pooled headline (Table S7, pure recomputation on the frozen per-session scores of Table S6, no refitting) confirms the sign is not carried by one session: dropping any one of the six leaves the gap positive, $+0.1542$ to $+0.3680$. The magnitude is concentration-sensitive, though – dropping BTC 2022-05-14, the session with the highest matched AUROC (0.9973), nearly halves the pooled gap, and every other single drop leaves it at or above $+0.2265$.

One asymmetry cannot be designed away and is disclosed rather than corrected. A matched holdout is carved out of the known set, drawn and then removed from training; natural unknowns are additional actors present in the session that never entered the labeled set. No matching makes these the same kind of object. What the three stages establish is that the disagreement between the two protocols survives both a within-family restriction and a composition-matched holdout on the sessions where it can be measured, not that naturalness is the mechanism corpus-wide.

Table S7: Leave-one-session-out sensitivity of the pooled matched-vs-natural gap. Each row drops the named session and recomputes the pooled matched and natural means over the remaining five (`t3_matched_pool_loso.json`).

session dropped	matched (n=5)	natural (n=5)	gap	vs. all-6 (+0.2738)
BTC 2022-05-14	0.5730	0.4188	+0.1542	−0.1196
BTC 2022-05-15	0.6058	0.3793	+0.2265	−0.0473
ETH 2022-05-14	0.6424	0.3759	+0.2666	−0.0072
ETH 2022-05-15	0.6721	0.3889	+0.2832	+0.0094
SOL 2022-05-15	0.6790	0.3347	+0.3442	+0.0705
AVAX 2022-05-15	0.6901	0.3222	+0.3680	+0.0942

Q THE REGISTERED TIMING-CHANNEL PREDICTIONS IN FULL

P01 holds that in a segmented task the native-microsecond regime adds at most 0.02, because inside a trace the events already share a source and coincidence says nothing about the grouping. The supervised experiment gives $\Delta_M = +0.0074$; over the 16 unsupervised sessions the mean is -0.0012, with a two-sided 95% interval of $[-0.0135, +0.0110]$, so the prediction passes under the stricter of the two registered readings, its session-level dispersion having an account covering part of it and an unidentified remainder (Appendix J). P02 states the reverse for the unsegmented task, at least $+0.10$ ARI, because there every co-timestamped pair is a same-actor edge and the only near-certain supervision available in the problem; its decision rule was fixed while T1b had not run and its data was unseen. Measured $\Delta_M = -0.1174$: opposite in sign and about four times the registered falsifier threshold of 0.03. Union-find over exact-microsecond equality returns 486,106 clusters over 500,000 orders, so near-perfect pairwise precision at 4.86% coverage does not compose into a partition, and any grouping method built on a sparse high-precision linking oracle inherits that ceiling. The consequence registered in advance was executed rather than renegotiated: Channel M leaves the decomposition, which becomes R, S and B, and is retained as the label-validity certificate and the defense result.

The session-level dispersion of Δ_M has an account covering part of it and an unidentified remainder; see Appendix J.

R THE LABEL-VALIDITY CERTIFICATE BY FAMILY, IN FULL

Computed within each family against that family's own chance baseline, the coincidence relation lifts pair purity by $9.32\times$ and $7.66\times$ in the two counter-keyed families and by about $2\times$ in the two node-keyed ones (Appendix Table S1). For `zero_counter` the lift is $1.01\times$, because one actor holds 99.97% of that family and chance purity is already 0.988, and for `ascii_filler` it is $1.00\times$ at $K = 1$, where chance is 1.0 by construction. The battery flags `date_counter` at 2.5 times the `uuid_v1` false-positive baseline, disclosed as a candidate-merge signal, and at 0.38 power the absence of a flag is weak evidence of cleanliness. Families are themselves separable from behavior at 0.9717 and 0.9317 balanced accuracy at $K = 10$ against a chance of 0.100, which is how the deflationary objection arises: a method might be reading the SDK rather than the actor. Holding family constant removes that possibility by construction and identity survives, at 0.9688 and 0.9981 inside `uuid_v1` and at 0.5107 and 0.4274 inside `const_prefix16`, the largest family. Those two differ by a factor of two, so a pooled actor-recovery number averages over populations of very different difficulty.

S THE SUPERVISED CLOSED-SET RESULT IN FULL

A gradient-boosted classifier separates 29 BTC actors and 30 ETH actors at a mean of 0.7064 balanced accuracy against a chance of 0.0339, about 21 times chance (Appendix Figure S1). Every variant except the control returns $p < 0.005$ over 200 permutation repetitions, BEHAV-all on BTC carrying a BCa interval of $[0.7178, 0.7541]$ against a null mean of 0.0347 while the label-shuffle control sits at 0.0228. Rate alone reaches four times chance, so actors do differ in it, yet removing rate from the full set costs 0.0069 and removing schedule 0.0052, and removing both leaves

0.7137, above the full set: in these two sessions the supervised signal is preserved by the rate- and schedule-free representation (Appendix B).

Table S8: Identity recovery, all 16 primary sessions, both timestamp regimes (FULL, native μs; BEHAV, rate- and schedule-free). Balanced accuracy against per-session chance; every cell $p < 0.005$ over 200 permutation repetitions, bootstrap 95% intervals never touch chance (`e1_16session_replication.json`, `e1_16session_replication_behavall.json`). BTC and ETH 2022-05-16 are the anchors reproduced from the published table above. These 16 sessions extend the same corpus and capture regime; the replication is within-corpus and not external.

session	chance	FULL	BEHAV	session	chance	FULL	BEHAV
BTC 05-14	0.0345	0.7393	0.7300	SOL 05-14	0.0417	0.6564	0.6420
BTC 05-15	0.0345	0.7349	0.7228	SOL 05-15	0.0417	0.6771	0.6831
BTC 05-16	0.0345	0.7373	0.7367	SOL 05-16	0.0417	0.6712	0.6367
BTC 05-17	0.0385	0.7442	0.7467	SOL 05-17	0.0476	0.5825	0.5938
ETH 05-14	0.0333	0.7426	0.7331	AVAX 05-14	0.0476	0.6225	0.6221
ETH 05-15	0.0333	0.7267	0.7359	AVAX 05-15	0.0476	0.5983	0.5776
ETH 05-16	0.0333	0.6925	0.6761	AVAX 05-16	0.0500	0.5965	0.5773
ETH 05-17	0.0370	0.7403	0.7445	AVAX 05-17	0.0625	0.5909	0.5979
mean by asset: BTC 0.739/0.734, ETH 0.726/0.722,				SOL 0.647/0.639, AVAX 0.602/0.594 (FULL/BEHAV)			

T LIMITATIONS IN FULL

The population limit is the tightest. Labellability is predictable from behavior at 0.8790 AUROC, so the actors we score are those whose client software leaks a structured identifier, and nothing here estimates recoverability for clients with sound identifier hygiene in either direction. Two things keep that narrowing from emptying the result: the reading of the diagnostic was fixed before it ran, so the scope follows a rule rather than an outcome, and within the population the signal is more than identifier-family classification, though an actor is itself defined as a client namespace, so actor-specific software versions and configurations are not excluded. A second bound runs alongside it: behavior identifies an actor three days and a month later far more readily than across a change of asset, so cross-asset linkage is a distinct unsolved task rather than a longer-range version of the same one, and every recovery number here is scoped twice, within identifier-leaking clients and within an asset. The rest are limits of sample rather than of kind: one venue, whose idiosyncrasy a second venue with the same field structure would separate from actor-level structure. The supervision axis on one session; the learned encoder on one venue's sessions; and session-level variation in Δ_M with an unidentified remainder.

U THE SELECTION DIAGNOSTIC

Mean AUPRC across the 17 scored sessions is 0.8591 against a mean chance of 0.3952, and the per-session AUROC ranges from 0.6750 to 0.9785. An eighteenth identifier-bearing session is not scored: it does not span the fixed split boundary, so the diagnostic is structurally undefined there rather than failed. The interpretation rule was fixed in the run brief before the diagnostic executed: an AUROC near 0.5 would have licensed a claim about anonymized event streams in general, and a high one restricts every recovery claim in this paper to identifier-leaking clients.

V THE LEARNED ENCODER

The estimated-K contrast is specific to the density-based estimator. On the withheld days HDBSCAN gives 0.2029 and 0.2444 on the learned embedding against 0.0212 and 0.0355 on hand-crafted features, while spectral eigengap selection on the same sessions gives 0.1977 against 0.1840 and 0.1689 against 0.1566. Hand features therefore do not collapse whenever the cluster count is withheld; they collapse under this one estimator. Eigengap separates the two representations too weakly to carry a reading either way, and it is unstable on both sides, returning 0.0021 for hand

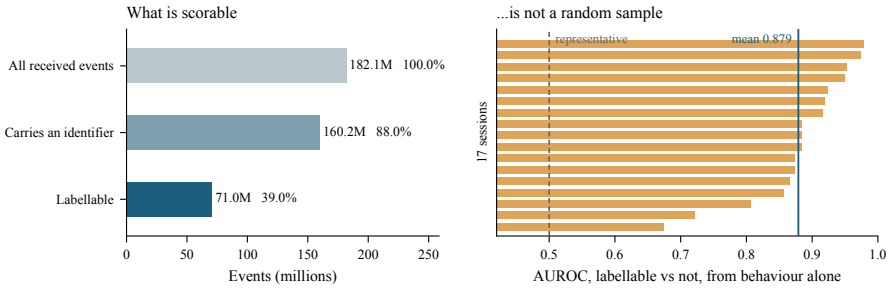


Figure S3: The scorable population is a behaviorally distinct slice of the stream. Left: of 182,119,794 received events, 160,185,756 carry a client-supplied identifier and 38.99% resolve to an actor label. Right: per-session AUROC for predicting whether an order is labellable from 20 behavioral features carrying no identifier-derived input. A representative subpopulation would sit at the dashed line.

features on SOL 2022-05-17 and 0.0091 and 0.0221 for the embedding on two single-seed cells of the same fold, so we report it as unstable rather than as a result for either representation. The density-based advantage is not an artifact of an untuned baseline: across the `min_cluster_size` sweep on the 2022-05-17 fold the hand baseline peaks at 0.0380 at 50 members against 0.0355 at the preregistered 25, and reaches 0.0003, 0.0243 and 0.0313 at 2, 5 and 10, so no setting makes it competitive under HDBSCAN; both arms use 25.

The encoder is a Set Transformer (Lee et al., 2019) over the events of a trace: two ISAB blocks with four heads at width 128 followed by a PMA head, reading per-event price, size, side, order type, position-in-trace and gap features with no identifier-derived input. Training is contrastive (Chen et al., 2020; van den Oord et al., 2018) and uses no labels: two views of a trace are formed by event dropout, feature masking and mild jitter of the continuous features, and the InfoNCE objective pulls them together against the other traces in the batch. It ran for 80 epochs over 104,742 traces on one laptop GPU, after which the weights were frozen and the embeddings scored under the T1a protocol without further fitting. All three planned seeds were trained to completion, and every encoder number here is reported across the three: oracle-K ARI 0.2495, 0.2686 and 0.2529 for seeds 1000, 1001 and 1002, mean 0.2570 with a standard deviation of 0.0102 that describes training randomness alone.

A separate set of checkpoints tests whether any of this survives holding data out of the self-supervised objective. Four folds were eventually trained at three seeds each, one per calendar day the corpus spans (2022-05-14, 05-15, 05-16, 05-17), each excluded in turn from self-supervised training and then frozen and scored on its own withheld sessions under the unchanged protocol; 05-14 and 05-17 are the extrapolation folds (the earliest and latest day), 05-15 and 05-16 the interpolation folds (interior days). Table S9 gives all four; Table S10 gives the per-session hand HDBSCAN absolutes behind the ratios in the main text. Held-out oracle-K ARI against hand-crafted features is $+0.0117$, $+0.0138$, $+0.0202$ and -0.0060 for 05-14/05-15/05-16/05-17, each measured against the hand baseline on that fold's own four withheld sessions (0.2286, 0.2310, 0.2105, 0.2158) rather than against the 16-session mean. The gap is smaller than the fold's three-seed range on 05-14, 05-15 and 05-17; on 05-16 it is larger, and all three seeds there lie above hand. That fold's seed range is also the narrowest of the four (0.0073 against 0.0236, 0.0165 and 0.0116), so the comparison against spread turns as much on the seed draw as on the gap. Across the four days (the unit a leave-one-day-out design generalizes over), the mean gap is $+0.0099$ with a fold-to-fold standard deviation of 0.0112, $t(3) = 1.77$, 95% interval -0.0079 to $+0.0278$; the same computation on the estimated-K gap returns $+0.2064$, $t(3) = 24.1$, interval $+0.179$ to $+0.234$. The transductive effect ($+0.0268$ on 05-14, $+0.0350$ on 05-17; not separately measured on the interpolation folds) is what the withdrawn oracle-K claim had rested on. Estimated-K ARI under HDBSCAN holds in all four folds (0.2029, 0.2499, 0.2419, 0.2444 against hand 0.0212, 0.0351, 0.0216, 0.0355), the two interpolation folds sitting at or above the two extrapolation folds, as preregistered before either interior fold ran. Eigengap is unstable for both representations in every fold measured (hand 0.1840/0.1605/0.1664/0.1566 against encoder 0.1977/0.2109/0.1799/0.1689 for 05-14/05-15/05-16/05-17). All 24 fold-level eval-

uations pass the leakage gates. An earlier evaluation of a one-epoch checkpoint scored 0.1905, below the hand-crafted features; reporting that as a negative result for learned representations would have been an artifact of an undertrained model, and the 80-epoch checkpoint is the comparison we report throughout. The mean estimated $\hat{K}$ over the full 16-session corpus (36.2) is within 1.4 of the true 36.75, but per-session errors span -27 to $+14$ and the count is exact in only 2 of 48 session-seed cells: that is cancellation between errors of opposite sign, not cardinality recovery, and no claim in this paper reads it otherwise.

Table S9: Held-out learned-embedding ARI by fold and estimator, three-seed means, against hand-crafted features on the same withheld sessions. 05-14/05-17 are the extrapolation folds (preregistered adjudication pair); 05-15/05-16 are the interpolation folds.

	05-14	05-15	05-16	05-17
hand, oracle-K	0.2286	0.2310	0.2105	0.2158
encoder, oracle-K (held out)	0.2403	0.2448	0.2308	0.2098
hand, HDBSCAN	0.0212	0.0351	0.0216	0.0355
encoder, HDBSCAN (held out)	0.2029	0.2499	0.2419	0.2444
hand, eigengap	0.1840	0.1605	0.1664	0.1566
encoder, eigengap (held out)	0.1977	0.2109	0.1799	0.1689

Table S10: Held-out HDBSCAN ARI, hand-crafted features, per session. Six of eight cells on the adjudication folds sit at or below 0.012, so the held-out embedding's 0.2029 and 0.2444 stand against a near-degenerate hand baseline, not against a comparably-scaled one (m4 discipline).

	BTC	ETH	SOL	AVAX
05-14	0.0002	0.0000	0.0112	0.0735
05-15	0.0010	0.0005	0.1062	0.0326
05-16	0.0004	0.0001	0.0068	0.0792
05-17	0.0004	0.0020	0.0097	0.1298

Attribution control: is the repair specific to attention? A parameter-matched, protocol-identical statistics-pooling encoder (Deep Sets family (Zaheer et al., 2017): per-event MLP, masked mean+std pooling over visible events, the same projection head, contrastive loss, augmentations, optimizer and 80-epoch budget as the Set Transformer) was trained at the same three seeds on the two preregistered adjudication folds (05-14, 05-17). Parameter matching is exact under the registered procedure, counted like for like as every trainable tensor jointly optimized during contrastive training (encoder plus shared projection head): 680,192 control parameters against 705,024 for the Set Transformer, the control 3.5% smaller, within the registered $\pm 10\%$ band; that the smaller model matches the larger one only strengthens the attribution. Decision rule R1 was fixed before the control was trained: it generalizes beyond attention if the control's three-seed mean exceeds $0.75\times$ the Set Transformer's mean on held-out HDBSCAN in both folds. It does, comfortably, in both: 0.2031 against 0.2029 on 05-14 and 0.2537 against 0.2444 on 05-17 (hand 0.0212/0.0355), so C11 is stated at the class level – learned invariant per-event representations with invariant pooling repair the HDBSCAN geometry that hand-feature moments collapse under, whether or not the pooling uses attention. The control also reproduces the oracle-K withdrawal: 0.2098 and 0.1828 against hand 0.2286/0.2158, no advantage on either fold. Table S11 gives the full comparison. On 05-17 the control's per-seed range (0.2506–0.2556) sits above the Set Transformer's (0.2421–0.2457); this is reported as a fact of the one fold measured, under no direction claimed, since the preregistered rule for a fold-consistent attention advantage (R2) did not fire and no cross-fold averaging is licensed to force a reading either way.

Order-aware comparator and order-shuffle ablation. Neither invariant encoder is order-blind. All three read the same nine per-event features, among them the position of the event in its trace, the log gap to its in-trace predecessor and a first-event flag, so the Set Transformer and the statistics-pooling control are permutation-invariant over position-tagged events rather than blind to sequence: permuting the rows of a trace with its mask permuted identically moves the frozen embeddings by at most 1.43×10^{-6} and 2.38×10^{-6} against a registered 10^{-4} threshold (anchor A3, 1024 real

traces). The comparison below therefore tests an inductive bias, consuming the events in order rather than pooling them invariantly, and not an information advantage. The comparator is a BiGRU-SP: a two-layer bidirectional GRU (Cho et al., 2014; Schuster & Paliwal, 1997) over the visible events in chronological order, masked mean and standard-deviation pooling over timesteps, then the same post-pooling MLP, projection head, NT-Xent objective at $\tau = 0.2$, augmentations, optimizer, schedule, 80-epoch budget and three seeds as the other two encoders. Hidden width was chosen before training by the registered matching rule, as the minimizer over a five-point sweep of the distance to 705,024 parameters: $g = 140$ gives 707,952, 0.42% above the Set Transformer, inside the registered $\pm 10\%$ band and on the larger side, so the order-aware arm carries the marginal capacity advantage. The readout is pooled rather than last-state, and event dropout deletes and repacks events instead of zero-filling them, since a zero row would read as a first event of zero size; the feature tensor is bitwise identical across all three architectures and only its consumption differs (anchor A6). The ablation replaces within-trace row order with a random permutation at tensor-build time, before augmentation, so training, evaluation and batching see the same permuted corpus and no train–eval order mismatch is possible. The permutation is fixed per trace, seeded by SHA-256 over the data seed, asset, date and trace identifier, and is not resampled per view: resampling would have turned the permutation into an augmentation and trained order-invariance explicitly, and it would also have left the shuffled arm approximating invariance against a permutation that changes at every step, which we take to be a harder optimization problem rather than a weaker signal. In the chronological arm the two positive views of a trace share one order; a fixed permutation preserves that sharing and changes only whether the shared order is the true one. The residual objection, that a fixed per-trace permutation is itself a shortcut cue for the contrastive task, applies to both arms alike, true order being equally a per-trace constant. The decision rules were fixed before any cell was scored, with an equivalence margin $\delta = 0.025$ (about 12% of the repair under dispute and about three times the Set Transformer's widest within-fold seed range, 0.0078), and they adjudicate the comparison against the invariant encoders separately from the comparison of the comparator against its own shuffled twin; all threshold arithmetic uses the four-decimal three-seed means printed in Table S11.

On held-out HDBSCAN the chronological arm reaches 0.1909 on 05-14 and 0.2099 on 05-17, above 0.75 of the Set Transformer's three-seed mean, the registered class-membership threshold, on every seed of both folds; its seed band sits below the Set Transformer's on both folds, below the statistics-pooling control's on 05-17 and overlapping it on 05-14. The preregistered equivalence condition against the Set Transformer holds on 05-14 (0.1909 against 0.2029) and fails on 05-17 (0.2099 against 0.2444), so the fold-inconsistent branch fires; both folds are reported in full, and no equivalence and no direction is claimed: the order-aware arm does not exceed either invariant encoder on either fold, and is inside the Set Transformer's equivalence margin on 05-14. The shuffle comparison is fold-consistent. The shuffled twin reaches 0.1868 and 0.2326, so the chronological-minus-shuffled difference is 0.0041 on 05-14 and -0.0227 on 05-17, both inside δ, with the shuffled arm above on 05-17 on all three seeds and disjoint bands and the chronological arm above on 05-14 on two seeds of three with overlapping bands. For this architecture, under this protocol, presentation order carries no recoverable signal: the order-shuffled twin moves held-out HDBSCAN ARI by less than δ on both adjudication folds, in opposite directions. Held-out oracle-K gives no advantage to either arm on either fold (0.1594 and 0.1681 chronological, 0.1615 and 0.1709 shuffled, against hand 0.2286 and 0.2158), and eigengap is as unstable for these representations as for the Set Transformer and the statistics-pooling control. Final NT-Xent loss is lower in the chronological arm than in its shuffled twin in all six matched pairs (1.9432–1.9458 against 1.9622–1.9673 on 05-14, 1.9586–1.9591 against 1.9765–1.9812 on 05-17), every value inside the registered band; the two arms are not the same network, and the loss difference does not appear as an ARI difference. One of the twelve trainings, chronological seed 1000 on 05-17, ran 3.895 h against a preregistered 3.5 h wall-clock substitution threshold while sharing the machine with an anchor re-run; the registered substitution was not applied, because its window, which closes before any evaluation metric is seen, had passed. The other ten cells ran 1.26–1.44 h per seed, and wall time enters no metric or rule. All twelve cells completed the identical 80-epoch budget. Both controls, the statistics-pooling encoder and the order-aware comparator with its shuffled twin, ran on the two adjudication folds only; 05-15 and 05-16 are disclosed as not run for either, so the class-level claim rests on the adjudication pair while the Set Transformer-only replication covers all four folds. The position and gap channel itself was not ablated, and no claim here is made about what those per-event features contribute; the reading is a property of this one recurrent architecture at this budget on this corpus, not of sequence models in general. We retain this as a diagnostic non-result rather than a resolution of the

architecture-class question; the two remaining leave-one-day-out folds were not run for the recurrent comparator or the statistics-pooling control.

Table S11: The two invariant encoders and the order-aware comparator against hand-crafted features, two adjudication folds; all learned rows are three-seed means scored on held-out sessions, with the per-seed range in parentheses under HDBSCAN, the primary estimator. The oracle-K per-seed ranges are omitted for width and are in the metrics file. Parameter counts (encoder plus shared projection head): Set Transformer 705,024, DeepSets-SP 680,192 (3.5% smaller), BiGRU-SP 707,952 (0.42% larger).

	05-14		05-17	
	oracle-K	HDBSCAN	oracle-K	HDBSCAN
hand-crafted	0.2286	0.0212	0.2158	0.0355
Set Transformer	0.2403	0.2029 (0.1995–0.2073)	0.2098	0.2444 (0.2421–0.2457)
DeepSets-SP	0.2098	0.2031 (0.1901–0.2192)	0.1828	0.2537 (0.2506–0.2556)
BiGRU-SP, chronological	0.1594	0.1909 (0.1859–0.1952)	0.1681	0.2099 (0.2059–0.2130)
BiGRU-SP, order-shuffled	0.1615	0.1868 (0.1846–0.1887)	0.1709	0.2326 (0.2296–0.2346)

W RETRIEVAL ACROSS HELD-OUT CONTEXTS, PER ASSET AND REGIME

Over all 48 ordered (gallery day, query day) pairs, 202,186 of 314,226 queries have a gallery entry and are scored. That retrievable share is split-dependent, running 0.43 to 0.86 by pair, so the single published pair is not representative of the axis. Rank-1 across the 48 pairs runs 7.97 to 31.23 times chance with a mean of 17.44. Forty-seven of the forty-eight clear the eightfold threshold registered as survival; the exception is SOL 2022-05-17 to 05-14 at 7.97, which we report rather than round up.

Time order does not survive chance normalization. Galleries earlier than their query average 0.5080 raw Rank-1 against 0.4667 the other way, which looks like an asymmetry, but their mean gallery holds 35.67 actors against 37.83 and Rank-1 falls with gallery size. Expressed as a multiple of each cell's own chance the two directions are 17.52 and 17.36, a difference of under one percent, and the effect is non-monotonic in the day gap (0.4714, 0.5384, 0.4333 at one, two and three days). We had recorded the raw difference as a directional finding; it is a gallery-size artifact of the same kind as the absolute-price-level clock documented above, and we withdraw it. Nor does retrieval decay measurably with the gallery-query gap over the range the corpus spans: chance-normalized Rank-1 is 16.8, 19.8 and 14.8 times chance at one, two and three days, a Spearman correlation with gap of -0.03 at $p = 0.82$. With only three gap values and eight cells at the longest, that is a null over an unresolvable range rather than evidence of stability; the one-month result is the only long-horizon evidence we have. On the asset axis, at a fixed day, 50,946 of 83,361 queries are retrievable; every BTC/ETH-gallery to SOL/AVAX-query cell sits at chance under raw features while the price-drift-robust representation recovers 2.5 to 4.2 times chance in all of them.

Table S13 gives the rows behind Table S12. The gallery holds one prototype per actor, the mean of that actor's gallery-side traces under a standardization fitted on the gallery alone; every query trace is ranked against it by Euclidean distance, and queries whose true actor has no gallery entry are excluded from Rank-k and mAP, which makes the protocol closed-set retrieval. The ARI column is a separate reading: it clusters the query side against itself under the gallery-frozen standardization, so it measures whether those actors are separable at all, independent of the gallery. On the asset split that column is 0.3338 while Rank-1 is 0.0366, which is the evidence that the cross-asset failure is one of correspondence rather than of representation. A tripwire registered in advance, Rank-1 above 0.97 at $K \geq 12$, did not fire on any row.

The price-robust variant drops ten price-scale columns, the mean and the 10th, 50th and 90th percentiles and standard deviation of log price and of one-second price dispersion. On the month split that removal is the difference between 0.0002 and 0.3581; on the asset split it moves Rank-1 by 0.0015.

Table S12: Retrieval across held-out contexts, BEHAV regime, one prototype per gallery actor and gallery-frozen standardization. The month split is shown twice: carrying absolute price levels across it scores two orders of magnitude below chance, because over that gap price level encodes calendar time. Per-asset rows, BCa intervals and the FULL regime are in Appendix W; every permutation p sits at the add-one floor of 0.001.

axis	gallery → query	K	Rank-1 (chance)	Rank-5	mAP
day	05-14 → 05-17, 4 assets	21–43	0.4487–0.5758 (1/43–1/21)	0.5976–0.9214	0.5390–0.7270
month	2022-04 → 2022-05, BTC	44	0.3581 (1/44)	0.6541	0.4981
month, raw	2022-04 → 2022-05, BTC	44	0.0002 (1/44)	0.0007	0.0712
asset	BTC,ETH → SOL,AVAX	66	0.0366 (1/66)	0.3090	0.1781

Table S13: Retrieval, per asset and regime. Rank-1 intervals are BCa over 1,000 resamples; every permutation p is at the add-one floor of 0.001 over 1,000 repetitions.

split	asset	regime	K	Rank-1 [95% CI]	Rank-5	mAP	ARI
day	BTC	BEHAV	43	0.4487 [0.4345, 0.4633]	0.5976	0.5390	0.1907
day	BTC	FULL	43	0.4766 [0.4622, 0.4923]	0.6183	0.5631	0.2282
day	ETH	BEHAV	43	0.5305 [0.5154, 0.5441]	0.8128	0.6547	0.2270
day	ETH	FULL	43	0.5305 [0.5161, 0.5440]	0.7836	0.6469	0.2203
day	SOL	BEHAV	24	0.5178 [0.5020, 0.5347]	0.8948	0.6720	0.1840
day	SOL	FULL	24	0.5213 [0.5056, 0.5391]	0.8933	0.6733	0.2269
day	AVAX	BEHAV	21	0.5758 [0.5551, 0.6000]	0.9214	0.7270	0.2769
day	AVAX	FULL	21	0.5395 [0.5173, 0.5611]	0.9058	0.6959	0.2433
month	BTC	BEHAV	44	0.3581 [0.3501, 0.3664]	0.6541	0.4981	0.1511
month	BTC	FULL	44	0.3833 [0.3755, 0.3919]	0.6780	0.5195	0.1605
month, prices kept	BTC	BEHAV	44	0.0002 [0.0001, 0.0007]	0.0007	0.0712	0.0104
month, prices kept	BTC	FULL	44	0.0002 [0.0001, 0.0007]	0.0010	0.0748	0.0101
asset	SOL, AVAX	BEHAV	66	0.0366 [0.0322, 0.0408]	0.3090	0.1781	0.3338
asset	SOL, AVAX	FULL	66	0.0387 [0.0341, 0.0432]	0.3209	0.1826	0.3468
asset, prices kept	SOL, AVAX	BEHAV	66	0.0351 [0.0311, 0.0395]	0.3231	0.1785	0.3572

X Coincidence availability under coarsening and after identifier removal

Worst per-session excess pair purity over per-session chance is $+0.2277$ at a 1 ms quantum, $+0.1530$ at 10 ms and $+0.1151$ at 100 ms, and it is not monotone: BTC 2022-04-14 worsens from $+0.0668$ at 10 ms to $+0.1151$ at 100 ms. In the three 2023 BTC sessions, where the client field is absent from the feed, coincidence remains available at 4.51–4.73% of all received events, against 8.01–13.95% in the six 2022 BTC sessions on the same all-events basis, a ratio of means of 0.462; coverage was already declining within 2022, so the difference is not attributable to the removal. Those sessions carry no labels, so purity is not computable there and we never report a 2023 purity number.

Y The defense program: coarsening, jitter, size discretization and the observation budget

Appendix X measures how much of the coincidence relation survives a coarser clock; it says nothing about whether the behavioral attack still works. This section measures that attack against a coarsened stream. The two are separate quantities and are never combined into a single defense number.

The transform floors `ts_us` to a quantum q on the raw stream, before featurization, and every feature is re-derived from the floored stream; the attacker sees only that stream at test time. Four panels are run, BTC and ETH 2022-05-16 under each of the two timestamp regimes, which is a counterfactual replay on the labeled 2022 sessions and applies to the same identifier-leaking population used throughout the paper. The non-adaptive attacker keeps the clean standardization and the clean classifier; the adaptive one refits the standardization and retrains the classifier from scratch on the defended traces at the magnitude it is tested against, so it sees clean data at no fitting stage. Clus-

tering is k-means at oracle K over ten seeds per cell. Fidelity gates: the $q = 1\,\mu$s point reproduces the clean feature matrix bit-identically in both sessions, the four structurally inert magnitudes in the quantized regime do the same, and Ward clustering at oracle K on the clean build reproduces the published T1a values exactly.

Table S14: Timestamp-coarsening frontier in the quantized regime, where the metadata channel is structurally absent. Clustering ARI is the registered-seed value; balanced accuracy is closed-set against a chance of 0.0339. The retrained columns are the always-retrain adaptive attacker; the best-of-both attacker is treated separately below. Magnitudes below 10 ms are structural no-ops in this regime. No ordered pair of magnitudes has disjoint seed-ensemble ranges, so no movement within this table is resolvable.

	BTC				ETH			
q	ARI		bal. acc.		ARI		bal. acc.	
	fixed	retrained	fixed	retrained	fixed	retrained	fixed	retrained
clean	0.2013	0.2013	0.7367	0.7367	0.1989	0.1989	0.6761	0.6761
10 ms	0.1650	0.1854	0.7378	0.7389	0.1958	0.1843	0.6761	0.6832
100 ms	0.1827	0.1896	0.7199	0.7402	0.1919	0.1902	0.6759	0.6904

The coarsest magnitude in the specification, $q = 1$ s, builds in none of the four panels: `burstiness_fano_mean` and `cv_of_gaps_mean` are undefined when a 32-event window collapses into too few one-second buckets. Patching them would change which traces enter the analysis, so the frontier is reported right-censored there. The boundary is a property of the featurizer and is identical under both regimes.

The ETH tie-manufacture backfire (main text) is quantified here: the coarse-q ten-seed ensemble at $q = 100$ ms sits at 0.1853–0.2005 under both standardization modes, above the clean ensemble range on ETH; the same magnitude on BTC falls inside a wider clean range, so the effect there is undetectable above restart noise rather than absent, which is why the mechanism is read as a genuine but ETH-specific backfire rather than a universal one.

16-session replication at $q = 100$ ms. The single-magnitude coarsening cell is replayed on all 16 primary sessions under both standardizer fittings (anchors reproduce the two-session panel exactly; `t7b_d1_16session_replication.json`, Table S15). The adaptive attacker is never degraded, 0 of 16 amendment_011-claimable; the non-adaptive one degrades in 5 of 16, all four on AVAX, where the defended ensemble range falls wholly below the clean range and a scaler refit alone – no model retraining – restores it, the clearest instance in the program of a non-adaptive-only evaluation certifying a defense that does not hold up. The backfire condition (defended range wholly above clean, TW6's degradation direction) fires mechanically in 2 of 16 cells, both AVAX under the adaptive arm; per the amendment_010 ten-seed-ensemble discharge rule applied to the first two firings above, this is read against the same established manufactured-ties mechanism, concentrated on AVAX's thinner book (smallest clean-SD on the metadata columns, so the largest relative amplification), and is not treated as a new tripwire condition.

Table S15: D1 coarsening at $q = 100$ ms, 16-session replication, per asset (mean of within-asset session means; never pooled across assets). Claimable/above use the amendment_011 seed-ensemble band rule; the remainder overlap the clean band.

asset	clean ARI (both arms)	defended ARI non-ad.	defended ARI adaptive	non-adaptive below/4	adaptive above/4	best-of-both bal. acc.
BTC	0.1982	0.1948	0.2148	0/4	0/4	0.7350
ETH	0.1978	0.2094	0.2082	0/4	0/4	0.7155
SOL	0.2069	0.1883	0.2261	1/4	0/4	0.6407
AVAX	0.2585	0.0764	0.2973	4/4	2/4[†]	0.6047

[†]adaptive arm wholly above clean, discharged to the manufactured-ties mechanism (thin book), not amendment_011-claimable protection.

Table S16: D1 coarsening at $q = 100$ ms, 16-session replication, per session (ten-seed ensemble ranges, native-microsecond regime). Rule is the amendment_011 outcome for non-adaptive/adaptive: BLW = wholly below clean (claimable degradation), ABV = wholly above clean (discharged backfire), OV = overlaps clean.

session	clean ARI	D1 ARI, non-adaptive	D1 ARI, adaptive	rule (na/ad)
BTC 05-14	[0.170,0.201]	[0.192,0.209]	[0.198,0.227]	ov/ov
BTC 05-15	[0.172,0.221]	[0.180,0.207]	[0.206,0.220]	ov/ov
BTC 05-16	[0.176,0.213]	[0.188,0.217]	[0.195,0.215]	ov/ov
BTC 05-17	[0.189,0.234]	[0.181,0.194]	[0.193,0.258]	ov/ov
ETH 05-14	[0.174,0.201]	[0.198,0.230]	[0.184,0.205]	ov/ov
ETH 05-15	[0.209,0.217]	[0.216,0.242]	[0.212,0.234]	ov/ov
ETH 05-16	[0.185,0.200]	[0.186,0.216]	[0.197,0.212]	ov/ov
ETH 05-17	[0.192,0.212]	[0.173,0.201]	[0.195,0.218]	ov/ov
SOL 05-14	[0.185,0.237]	[0.176,0.222]	[0.207,0.261]	ov/ov
SOL 05-15	[0.195,0.248]	[0.170,0.235]	[0.210,0.276]	ov/ov
SOL 05-16	[0.187,0.231]	[0.143,0.184]	[0.182,0.247]	BLW/ov
SOL 05-17	[0.194,0.221]	[0.154,0.202]	[0.202,0.218]	ov/ov
AVAX 05-14	[0.237,0.273]	[0.046,0.047]	[0.281,0.353]	BLW/ABV
AVAX 05-15	[0.229,0.280]	[0.032,0.033]	[0.283,0.349]	BLW/ABV
AVAX 05-16	[0.234,0.289]	[0.045,0.116]	[0.259,0.315]	BLW/ov
AVAX 05-17	[0.249,0.277]	[0.139,0.157]	[0.254,0.312]	BLW/ov

Two internal tripwires fired on the first two runs and are discharged. Both compared movements across magnitudes against a bootstrap interval computed conditional on a fixed partition, which carries no k-means restart variance and is roughly five times too narrow; a preregistration amendment, written after those firings were diagnosed and before the ensemble that adjudicates them was run, replaced that yardstick with the ten-seed ensemble range and restricted the dominance test to classification. Under the replacement no tripwire fires in any of the four panels under either standardization mode, and the earlier firings are seed-draw artifacts of the narrow yardstick. A second, mirror amendment governs the opposite direction: a defense-induced degradation of a clustering metric may be claimed only if the defended ten-seed range lies wholly below the clean range, the same session, variant and standardization mode; directional drift without that separation is reported factually and never claimed. It was fixed in the jitter (D2) adjudication below, before the size-coarsening (D3) panel ran, so its one positive application there is not a rule chosen after seeing the result it would license.

The best-of-both classification attacker keeps whichever arm scores higher on the early-stopping split, reused for arm selection. The retrained arm has an asymmetric familiarity with that split, which can change which arm is selected but cannot inflate a test number, since the test side is untouched by the selection. A post-hoc audit flags cells where the selected arm loses on test by more than the clean arm's interval half-width: one of 24 does, BTC at 100 ms in the quantized regime, where the selected arm scores 0.7199 and the unselected one 0.7402. The headline for that cell stays 0.7199 by the preregistered rule; 0.7402 is an oracle-selection upper bound, reported here once as a sensitivity. The direction, the split overrating the clean arm, is opposite to the bias the reuse would produce, so it reads as drift between the selection and test windows.

TIMESTAMP JITTER (D2)

Jitter draws $\delta_i \sim U(-\delta_{\max}, \delta_{\max})$ independently per event, re-sorts within each trace and the shared background stream, and re-derives every feature from the jittered stream; the attacker sees only that stream at test time. Unlike the floor, jitter does not collapse distinct arrival times onto one value, so every magnitude up to $\delta_{\max} = 1$ s builds in all four panels, and the frontier is uncensored where D1's was not. Fidelity: $\delta = 0$ reproduces the clean feature matrix bit-identically in all four panels; Ward oracle-K ARI matches the published values exactly on both BEHAV panels and classification within 1% on both FULL panels, the same pre-existing LightGBM determinism gap noted for D1. RMS displacement equals $\delta/\sqrt{3}$ exactly, and the fraction of scrambled adjacent pairs approaches the analytic ceiling $1 - 1/(2\delta_{\mu s})$, confirming the transform rather than the attack. No tripwire fires in any panel or standardization mode.

Table S17: Timestamp-jitter frontier, native-microsecond regime. ARI is the ten-seed clustering-ensemble mean; balanced accuracy is closed-set against a chance of 0.0339. Fixed is the non-adaptive attacker, retrained the always-retrain adaptive one.

	BTC				ETH			
δ	ARI		bal. acc.		ARI		bal. acc.	
	fixed	retrained	fixed	retrained	fixed	retrained	fixed	retrained
clean	0.1928	0.1928	0.7373	0.7373	0.1936	0.1936	0.6925	0.6925
1 ms	0.1852	0.1844	0.7243	0.7380	0.1876	0.1919	0.6881	0.6819
10 ms	0.1874	0.1843	0.7165	0.7316	0.1884	0.1895	0.6886	0.6917
100 ms	0.1789	0.1895	0.6607	0.7239	0.1810	0.1847	0.6937	0.7017
1 s	0.1747	0.1735	0.6634	0.7039	0.1846	0.1820	0.6731	0.6909

Table S18: Timestamp-jitter frontier, quantized regime (metadata channel structurally absent). Same conventions as Table S17.

	BTC				ETH			
δ	ARI		bal. acc.		ARI		bal. acc.	
	fixed	retrained	fixed	retrained	fixed	retrained	fixed	retrained
clean	0.1913	0.1913	0.7367	0.7367	0.1915	0.1915	0.6761	0.6761
1 ms	0.1907	0.1909	0.7388	0.7314	0.1916	0.1914	0.6815	0.6951
10 ms	0.1921	0.1863	0.7419	0.7323	0.1868	0.1899	0.6848	0.6839
100 ms	0.1819	0.1819	0.7052	0.7243	0.1838	0.1832	0.6837	0.6897
1 s	0.1720	0.1718	0.6632	0.7164	0.1845	0.1782	0.6851	0.6875

On BTC/FULL the non-adaptive classifier's bootstrap 95% interval at $\delta = 1$ s, $[0.6475, 0.6789]$, is disjoint from its own clean-magnitude interval, $[0.7188, 0.7550]$; the retrained arm recovers to $[0.6858, 0.7205]$, overlapping clean again. ETH shows no comparable separation at any δ on either regime, so the exhibit is reported for BTC only. Channel-M feature movement saturates at $\sim$0.27 clean standard deviations at every nonzero δ, identical at 1 ms and 1 s, and the columns' `rms_z` collapses toward $\sim$0.30 rather than the 3–9 range coarsening produced: symmetric jitter destroys exact-timestamp ties with probability $\approx 1 - 1/(2\delta_{\mu s})$ per pair, a step function saturating at the first nonzero δ, which is why the ETH backfire seen under D1 does not recur here: there is no inflated tie signal left to backfire. The best-of-both classifier scores above its own clean value in 8 of 16 nonzero cells, all four ETH/BEHAV magnitudes among them; every excess is $+0.5$ to $+1.4$ points and sits inside the undefended cell's own 95% interval, and in three of the four the clean arm itself produced the higher number on the jittered test features, so this is evaluation-side noise rather than a jittered attacker outperforming a clean one.

Observation budget and unicity (T7a)

Gallery 2022-05-14, query 2022-05-17, all four assets, seed 20260803. The retrieval budget W' truncates each query trace to its first W' events before featurization; $W' \in \{1, 2, 4, 8, 16, 24, 32\}$ is left-censored per session by a leakage gate (AVAX builds from $W' = 4$, the other three assets from $W' = 8$), and $W' = 32$ reproduces the published T4 cells exactly. Zero tripwires fire, zero unicity-monotonicity violations occur across roughly 24,000 actor-repetition sequences, and zero population-integrity failures occur.

The censored-aware summary treats BTC's two not-reached curves as right-censored above 32 rather than dropped, giving a median W^* of 24 (BEHAV) and 20 (FULL) across the four assets; pooling the three assets that do cross without that correction understates BTC's contribution and is not used.

Size discretization (D3)

Order size is rounded to the nearest point on a geometric grid, size$' = s_0(1 + g)^k$, anchored at each session's own smallest positive size, $g \in \{0, 1\%, 5\%, 10\%, 25\%, 50\%\}$. The only well-defined alternative, self-referential relative rounding with the reference equal to the order's own size, reduces

Table S19: Rank-1 against the observation budget W', symmetric truncation, both regimes. Dashes mark magnitudes left-censored by the leakage gate. W^* is the smallest built W' at which Rank-1 ≥ 0.5; none of the eight curves reaches Rank-1 ≥ 0.9 by $W' = 32$, the instrument boundary.

Asset / regime	W'=4	W'=8	W'=16	W'=24	W'=32	W^*
BTC / BEHAV	–	0.3537	0.4000	0.4256	0.4487	not reached
BTC / FULL	–	0.3777	0.4289	0.4536	0.4766	not reached
ETH / BEHAV	–	0.4781	0.5081	0.5215	0.5305	16
ETH / FULL	–	0.4697	0.5050	0.5191	0.5305	16
SOL / BEHAV	–	0.4238	0.4795	0.4991	0.5178	32
SOL / FULL	–	0.4081	0.4772	0.5027	0.5213	24
AVAX / BEHAV	0.5048	0.5421	0.5778	0.5698	0.5758	4
AVAX / FULL	0.4972	0.5199	0.5275	0.5254	0.5395	8

Table S20: Unicity, oracle matcher, P1 (dense) labeled population within one session: an upper bound on discriminability within that subpopulation, not an end-to-end re-identification rate. U-time uses only a one-second arrival bucket; U-time+attr adds side, order type and exact size. A3 check (fraction of actors with a private small-value size set) is 0.0000 in every session, so the joint uniqueness is not a private-size-grid artifact.

	U-time		U-time+attr					
Session	p=1	p=32	p=1	p=2	p=4	p=8	p=16	p=32
BTC 05-14	0.0000	0.0353	0.4947	0.5981	0.6837	0.7840	0.9284	0.9956
BTC 05-17	0.0000	0.0448	0.6562	0.7712	0.8257	0.8738	0.9524	0.9950
ETH 05-14	0.0000	0.0077	0.4884	0.6088	0.7112	0.8267	0.9581	0.9984
ETH 05-17	0.0002	0.0156	0.6177	0.7204	0.7985	0.8842	0.9685	0.9981
SOL 05-14	0.0042	0.3971	0.6913	0.8658	0.9750	0.9988	1.0000	1.0000
SOL 05-17	0.0077	0.5013	0.6506	0.8281	0.9565	0.9906	1.0000	1.0000
AVAX 05-14	0.0067	0.4919	0.7948	0.9405	0.9943	1.0000	1.0000	1.0000
AVAX 05-17	0.0112	0.8542	0.7862	0.8973	0.9250	0.9354	0.9473	0.9615

to the identity transform on every registered g because $1/g$ is an exact integer, and is not a live candidate. Fidelity, Channel-M sanity and every tripwire are clean in every cell: $g = 0$ is bit-identical, Ward oracle-K ARI and classification match the published values exactly or within 1%, and Channel-M column movement is exactly 0.0 in every cell, since size coarsening cannot touch timing by construction. The claim yardstick for a clustering degradation is the mirror rule stated above (fixed before this panel ran): a defended ten-seed range must lie wholly below the clean range. It does, at every $g \geq 1\%$, in all four panels and both attackers; jitter's directional ARI decline did not clear the same bar.

Table S21: Size-coarsening frontier, native-microsecond regime. Same conventions as Table S17. The fixed classifier is scored on features it was never fit against, since more than 99.9% of sizes change at every nonzero g; its collapse to 0.35–0.49 reflects that distribution shift, not a property of the defense, and retraining recovers it fully.

	BTC				ETH			
g	ARI		bal. acc.		ARI		bal. acc.	
	fixed	retrained	fixed	retrained	fixed	retrained	fixed	retrained
0	0.1928	0.1928	0.7373	0.7373	0.1936	0.1936	0.6925	0.6925
1%	0.1419	0.1407	0.4883	0.7191	0.1311	0.1572	0.4942	0.6976
5%	0.1393	0.1449	0.4692	0.7179	0.1264	0.1261	0.4232	0.7046
10%	0.1390	0.1432	0.3893	0.7079	0.1265	0.1250	0.4222	0.7022
25%	0.1401	0.1478	0.3986	0.7156	0.1236	0.1267	0.4591	0.6923
50%	0.1413	0.1504	0.3780	0.6976	0.1236	0.1284	0.4218	0.6899

Two anomalies are disclosed rather than investigated further. This is the third defense panel in which ETH's best-of-both classifier scores above its own clean value (8 of 20 D3 cells, Table S21

Table S22: Size-coarsening frontier, quantized regime. Same conventions as Table S21.

g	BTC				ETH			
	ARI		bal. acc.		ARI		bal. acc.	
	fixed	retrained	fixed	retrained	fixed	retrained	fixed	retrained
0	0.1913	0.1913	0.7367	0.7367	0.1915	0.1915	0.6761	0.6761
1%	0.1329	0.1349	0.4744	0.7141	0.1263	0.1554	0.4645	0.6871
5%	0.1288	0.1383	0.4469	0.7090	0.1241	0.1173	0.3942	0.6918
10%	0.1325	0.1375	0.3572	0.7079	0.1263	0.1225	0.3974	0.6992
25%	0.1322	0.1380	0.3620	0.7114	0.1259	0.1225	0.4301	0.6903
50%	0.1339	0.1417	0.3501	0.7124	0.1224	0.1259	0.3947	0.6950

Table S23: Single-order ($p = 1$) joint-attribute uniqueness under size coarsening, all eight T4 sessions, U-time+attr basis. Eight-order ($p = 8$) uniqueness on the same grid moves by at most 0.014 across every session and stays in $[0.777, 1.000]$, so uniqueness relief from this defense exists only at the smallest observation budgets.

Session	g=0	1%	5%	10%	25%	50%
BTC 05-14	0.4947	0.4798	0.4484	0.3953	0.3170	0.2547
BTC 05-17	0.6562	0.6371	0.5821	0.5281	0.4217	0.3507
ETH 05-14	0.4884	0.4714	0.4416	0.4037	0.3277	0.2581
ETH 05-17	0.6177	0.6035	0.5567	0.5029	0.4037	0.3140
SOL 05-14	0.6913	0.6875	0.6642	0.6254	0.5704	0.4821
SOL 05-17	0.6506	0.6455	0.6229	0.5968	0.5281	0.4677
AVAX 05-14	0.7948	0.7819	0.7600	0.7252	0.6624	0.5833
AVAX 05-17	0.7862	0.7746	0.7369	0.6992	0.6488	0.6104

and Table S22); as under D1 and D2, ETH's clean point estimate sits low in its own model-fit sampling distribution, all but one of the eight cells fall inside the clean bootstrap interval, and every cell shares one test set, so the excesses are strongly correlated rather than eight independent draws. It stays below the claim threshold; the reserve check, a five-draw retraining-variance range from fit-set row-order permutation ($\sim$10 CPU-minutes), is available if a reviewer asks for it, and "perturbation improves the attacker" is not a licensed reading. Separately, one restart-diagnostic cell, ETH/BEHAV at $g = 1\%$, has adaptive and non-adaptive ensemble ranges disjoint in the direction adaptive-above-non-adaptive; this is benign and mechanistically expected, since a refit scaler handles the near-constant columns the grid deletes correctly where a clean-fit scaler mis-scales them, and it is a diagnostic rather than a tripwire.

16-SESSION REPLICATION (D2, D3)

Delta = 100 ms (D2) and $g = 1\%$ (D3) were rerun on all 16 primary sessions, FULL-all, both standardizer fittings, ten-seed ensembles, best-of-both classification, under the amendment 011 rule. Internal anchors (BTC/ETH 2022-05-16) reproduce the two-session panels exactly. Table S24 gives the per-asset means; per-session detail (rule outcome for each of the 64 defense-standardizer-session cells) is in the run report (`t7b_defence_frontier_16session_replication_report.md`). D2 overlaps clean in 16/16 sessions under both standardizer fittings (0/32 cells claimable); D3 lies wholly below clean in 16/16 sessions under both standardizer fittings (32/32 cells claimable), the degradation surviving the same asset heterogeneity that broke the session-level inference behind the withdrawn title claim (Appendix O).

THE ESTIMATED-K, SEGMENTATION-GIVEN, FROZEN-ENCODER LADDER

Every defense result above scores an attacker who is given the true actor count. The ladder below withdraws that count and changes the representation with it: the frozen 2022-05-16-held-out Set Transformer checkpoints (three seeds, never retrained on defended data), embedding the defended stream and clustering it with estimated-K HDBSCAN, with trace boundaries still supplied by the

Table S24: 16-session D2/D3 replication, per-asset means of within-asset session means (never pooled across assets), non-adaptive standardizer fitting. D3 mean relative size error: 0.21–0.29% across assets.

	BTC	ETH	SOL	AVAX
clean ARI	0.1982	0.1978	0.2069	0.2585
D2 δ=100ms ARI	0.1829	0.1893	0.1981	0.2545
D2 rule (claimable/4)	0/4	0/4	0/4	0/4
D3 g=1% ARI	0.1372	0.1302	0.1351	0.1557
D3 rule (claimable/4)	4/4	4/4	4/4	4/4

segmentation oracle and with no retraining on defended observations (that arm is disclosed, deferred and not run here). Preconditions: 0/24 determinism-check failures on the defended raw stream, 0/6 identity-point anchor failures against the stored clean embeddings, and all six clean cells matching the stored held-out evaluation to full floating-point precision. Table S25 gives the full 3-seed means. D3 at $g = 1\%$, the only defense claimable against the oracle-K hand-feature attacker (Appendix Y, main text), sits inside the clean condition's own seed range on both sessions; D1 and D2 stay within ± 0.009 of clean (largest movement BTC D1 $+0.0084$). Two anomalies are flagged and not interpreted further: BTC under D2 shows HDBSCAN $\hat{K} = 29$ for all three seeds (clean $\hat{K}$ 32/37/37), a partition-granularity shift the scalar ARI and B^3 F1 do not surface and that does not recur on ETH; BTC under D1, seed 1002 reaches the highest ARI in the entire 24-cell table (0.1848) while its own B^3 F1 (0.4354) is simultaneously the lowest in the BTC block, an ARI/B^3F1 disagreement attributed to noise-to-singleton conservatism at an elevated noise count (4,173 against $\sim$3,400 baseline).

Table S25: Estimated-K ladder, instance-level ARI, three-seed means, BTC/ETH 2022-05-16; frozen held-out encoder, segmentation given, no retraining on defended data. Seed range in brackets.

condition	BTC ARI	ETH ARI
clean	0.1577 [0.1547–0.1614]	0.1711 [0.1686–0.1743]
D1 q=100ms	0.1661 [0.1474–0.1848]	0.1636 [0.1588–0.1688]
D2 δ=100ms	0.1509 [0.1500–0.1517]	0.1699 [0.1672–0.1719]
D3 g=1%	0.1585 [0.1568–0.1611]	0.1692 [0.1634–0.1738]

The no-segmentation rung. One rung remains: withdrawing the segmentation oracle as well. The three C19 checkpoints are frozen and run in inference only, with no training of any kind, over the published 500,000-order factorial slice for BTC and ETH 2022-05-16, whose identity is asserted field by field against the published run (order count, session total, slice start and end index, true actor count, actor size range). The boundary source is the only thing that varies between the three block arms; a fourth arm shuffles arrival order as a locality control. Contiguous blocks of the labeled stream are out of distribution for this encoder, which biases against the arm under test. The decision rule was registered before any cell ran. Its margin is $\varepsilon = 0.02$, the smallest realizable gain this paper reports (oracle versus no blocking, $+0.0233$), and its gap is 0.10. Branches N0–N3 require agreement on both sessions. The primary panel (k-means at oracle K, per-arm scaler) and the second panel (estimated-K HDBSCAN) were both fixed in advance. The run script prints the arithmetic and emits no verdict. On the primary panel the contiguous arm reaches 0.0012 (BTC) and 0.0013 (ETH) three-seed mean order-level ARI against registered thresholds $\max(h_2, h_3)+\varepsilon$ of 0.0228 and 0.0256, where h_2 and h_3 are the published hand-featurizer contiguous and random arms of Table S26, and the same encoder's oracle-block arm stands $+0.1303$ and $+0.1492$ above it, clearing the gap. Fitting one scaler on the contiguous arm per encoder seed leaves every contiguous-arm cell identical by construction and moves the oracle-block arm to 0.1444/0.1702 (k-means) and 0.5174/0.7267 (HDBSCAN). Branch N1 fires on both sessions and in all four panels: a frozen learned representation does not rescue an attacker who must draw its own boundaries. The segmentation-given threat model under which every learned result here is reported is therefore a measured capability assumption rather than an unmeasured one. Drawing its own boundaries means here the one rule tested, contiguous 32-order blocks of the labeled stream in which every order inherits its block's cluster; other boundary-inference methods and encoders adapted to such blocks were not tested. Gate 0, computed before any embedding was scored, assigns each contiguous block its majority true ac-

tor and scores 0.0533 (BTC) and 0.0679 (ETH): a feasible block-constant labelling, not an upper bound on block-constant ARI, so it shows only that the registered thresholds are attainable by some block-constant labelling. The contiguous arm and the two controls sit at the same floor, the largest paired per-seed difference between the contiguous arm and either of them being 0.0017, a tenth of ε. Under HDBSCAN the three-seed means reverse in sign, so nothing is read from it. The oracle-block learned arm is reported for completeness as the branch's internal reference, not as a contrast against the hand oracle arm. The two pre-registered clusterers disagree by a factor of three to five on the same embeddings (0.1315 against 0.4865 on BTC; 0.1505 against 0.7307 on ETH), and the estimated count moves in opposite directions against the true 39 and 42. No attacker-strength comparison is drawn from it here. The degenerate per-order rung ($W = 1$) was not run: it cannot pass the unmodified featurizer, which requires a row count that is a multiple of 32, and modifying the featurizer or padding the input were both excluded. It was registered as a diagnostic banned from every contrast, so no adjudication input is lost. In consequence the factorial's "worse than not blocking at all" comparison is not tested under a learned representation. The rung's scope is one encoder architecture at three seeds, frozen and non-adaptive throughout, order-level ARI on the 500,000-order slice and $W = 32$ alone, on two sessions of a single date as registered and on sixteen after the extension below; what it licenses is a statement about a frozen, non-adaptive attacker.

Table S26: The no-segmentation rung: order-level ARI, three-seed means over encoder seeds 1000/1001/1002, per-arm scaler, BTC and ETH 2022-05-16. The slice is the published factorial's 500,000 contiguous labeled orders at $W = 32$, true actor count 39 (BTC) and 42 (ETH). Per-seed range in brackets; the HDBSCAN columns also give the range of the estimated count $\hat{K}$. Coverage is 1.0000 for the contiguous and time-shuffled arms (15,625 blocks) and 0.9988 for the oracle-block and random arms (15,607 blocks, 576 leftover orders). The hand rows are read from the published factorial, which has an oracle-K panel only, hence the absent HDBSCAN entries.

	k-means at oracle K		HDBSCAN, estimated K	
arm (boundary source)	BTC	ETH	BTC	ETH
L4 oracle blocks	0.1315 [0.1228, 0.1395]	0.1505 [0.1410, 0.1586]	0.4865 ($\hat{K}$ 43–52) [0.4279, 0.5552]	0.7307 ($\hat{K}$ 28–38) [0.6023, 0.8474]
L2 contiguous	0.0012 [0.0009, 0.0016]	0.0013 [0.0009, 0.0017]	-0.0015 ($\hat{K}$ 2–4) [-0.0073, 0.0016]	-0.0185 ($\hat{K}$ 2–7) [-0.0237, -0.0145]
L3 random size-matched	0.0003 [0.0002, 0.0004]	0.0001 [0.0000, 0.0003]	0.0005 ($\hat{K}$ 2–4) [0.0001, 0.0010]	-0.0026 ($\hat{K}$ 2) [-0.0037, -0.0016]
L2 time-shuffled	0.0004 [0.0002, 0.0006]	0.0001 [0.0000, 0.0001]	0.0005 ($\hat{K}$ 2–4) [0.0001, 0.0010]	-0.0026 ($\hat{K}$ 2) [-0.0037, -0.0016]
hand A4 (published)	0.1802	0.1617	—	—
hand A2 (published)	0.0028	0.0056	—	—
hand A3 (published)	0.0007	0.0008	—	—
Gate-0 majority-label reference	0.0533	0.0679	—	—

Extension to sixteen sessions (C21-E). The rung was extended under a rule registered before any of the fourteen new sessions was scored: every session of the factorial, each scored by the only frozen checkpoint whose training held its day out, with the thresholds, panels, branches and gap above unchanged and the aggregate licence fixed in advance. Branch N1 fires on all sixteen sessions and in all four panels (Table S27). On the primary panel the contiguous arm's three-seed means run from -0.0001 to 0.0142 against thresholds of 0.0202 to 0.0370, and the oracle-block arm stands 0.1284 to 0.2452 above it. The finding that a frozen learned representation does not rescue an attacker who must draw its own boundaries, reported above on two sessions of a single date, now rests on all sixteen sessions, four assets $\times$ four days, each scored by the only frozen checkpoint that held its day out. The sixteen sessions were launched in two batches, twelve and then AVAX/SOL 2022-05-16, with the BTC/ETH 2022-05-16 cells entered unchanged from the earlier frozen run; restricted to the fourteen newly scored sessions, N1 fires in 14 of 14. The sessions are four days of four assets, not sixteen independent replications, and the extension changes neither the architecture, the seeds, the window nor the non-adaptive scope: no learned $W = 1$ arm was run, and no cell orders the learned contiguous arm against the hand one.

Table S27: The no-segmentation rung, sixteen sessions: order-level ARI, primary panel (k-means at oracle K, per-arm scaler), three-seed mean over encoder seeds 1000/1001/1002 with the per-seed [min, max] range in brackets. Each of the sixteen 500,000-order slices is centered in one asset-day; the frozen, non-adaptive Set Transformer scoring it is the checkpoint whose training left that day held out (C21-E leakage-safe fold mapping). Columns L_2/L_4/L_3/L_2 time-shuf. are the four registered boundary arms (contiguous blocks of the labeled stream, oracle actor-homogeneous, random size-matched, and L_2 under a fixed arrival-order shuffle); L_1 ($W = 1$) was not run and is not a column. 'threshold' is $\max(h_2, h_3) + \varepsilon$ with $\varepsilon = 0.02$, where h_2 and h_3 are the published hand-featurizer contiguous- and random-block order-level ARI for the identical session at $W = 32$ (`segmentation_factorial_16session.json`); neither h_2 nor h_3 is printed on its own, and no cell in this table orders the learned arm against either. Disclosure only: restricted to the fourteen newly scored sessions, N1 holds in 14 of 14; no aggregate rests on this count. 'Gate-0 reference' is the order-level ARI of assigning each contiguous block its majority true actor, computed before any embedding was scored: a feasible block-constant labelling, not an upper bound on block-constant ARI. 'batch' marks the launch group: 1 = the twelve-session batch launched together 2026-09-18; 2 = AVAX/SOL 2022-05-16, launched in a second invocation the same day; F = BTC/ETH 2022-05-16, entered by reference, unchanged, from the original two-session run (`run_id c21_lnoseg_20260910T054708Z`) and not re-run. ^{a}: this session's threshold is set by h_3 because its published h_2 is negative (Hubert–Arabie ARI is not bounded below by zero). Rows are grouped by asset and listed in chronological date order within each asset as a neutral index; no ranking or trend across sessions is implied. Coverage is 1.0000 for the L_2 and L_2 time-shuffled arms (15,625 blocks) and 0.9988–0.9994 for the L_4/L_3 arms (15,606–15,615 blocks, W=32 remainder). Source: `outputs/metrics/learned_no_segmentation_ext16.json` (`run_id c21_lnoseg_20260918T043813Z`).

					primary panel, three-seed mean [min, max] across seeds 1000/1001/1002			
asset	date	batch	threshold $\max(h_2, h_3)+\varepsilon$	Gate-0 reference	L_2 contiguous	L_4 oracle	L_3 random	L_2 time-shuf.
BTC	05-14	1	0.0232	0.0631	0.0018 [0.0015, 0.0024]	0.1547 [0.1265, 0.1739]	0.0006 [0.0001, 0.0009]	0.0006 [0.0000, 0.0009]
	05-15	1	0.0239	0.0723	0.0026 [0.0015, 0.0033]	0.1687 [0.1580, 0.1798]	0.0002 [0.0000, 0.0004]	0.0002 [0.0000, 0.0004]
	05-16	F	0.0228	0.0533	0.0012 [0.0009, 0.0016]	0.1315 [0.1228, 0.1395]	0.0003 [0.0002, 0.0004]	0.0004 [0.0002, 0.0006]
	05-17	1	0.0232	0.0743	0.0021 [0.0016, 0.0024]	0.1365 [0.1306, 0.1418]	0.0002 [0.0001, 0.0002]	0.0001 [0.0000, 0.0003]
ETH	05-14	1	0.0253	0.0664	0.0008 [0.0005, 0.0011]	0.1539 [0.1355, 0.1665]	0.0000 [-0.0001, 0.0002]	0.0000 [-0.0002, 0.0003]
	05-15	1	0.0245	0.0635	0.0015 [0.0007, 0.0023]	0.1541 [0.1499, 0.1574]	-0.0001 [-0.0003, 0.0002]	0.0000 [-0.0002, 0.0002]
	05-16	F	0.0256	0.0679	0.0013 [0.0009, 0.0017]	0.1505 [0.1410, 0.1586]	0.0001 [0.0000, 0.0003]	0.0001 [0.0000, 0.0001]
	05-17	1	0.0252	0.0696	0.0011 [0.0007, 0.0017]	0.1295 [0.1239, 0.1346]	0.0000 [0.0000, 0.0000]	-0.0001 [-0.0002, 0.0000]
SOL	05-14	1	0.0223	0.0419	0.0008 [0.0007, 0.0009]	0.2013 [0.1877, 0.2163]	0.0008 [0.0007, 0.0009]	0.0009 [0.0007, 0.0010]
	05-15	1	0.0225	0.0287	0.0013 [0.0011, 0.0018]	0.2465 [0.2366, 0.2643]	0.0014 [0.0012, 0.0016]	0.0015 [0.0012, 0.0016]
	05-16	2	0.0204^a	0.0342	-0.0001 [-0.0001, 0.0000]	0.2305 [0.2117, 0.2460]	0.0009 [0.0006, 0.0011]	0.0009 [0.0006, 0.0013]
	05-17	1	0.0202^a	0.0338	0.0010 [0.0010, 0.0011]	0.2165 [0.2116, 0.2198]	0.0019 [0.0017, 0.0021]	0.0016 [0.0013, 0.0020]
AVAX	05-14	1	0.0260	0.0328	0.0025 [0.0021, 0.0027]	0.1857 [0.1700, 0.2108]	-0.0008 [-0.0008, -0.0007]	-0.0006 [-0.0008, -0.0005]
	05-15	1	0.0262	0.0302	0.0031 [0.0029, 0.0034]	0.1959 [0.1851, 0.2157]	-0.0006 [-0.0008, -0.0003]	-0.0006 [-0.0008, -0.0002]
	05-16	2	0.0263	0.0290	0.0020 [0.0014, 0.0024]	0.1901 [0.1827, 0.2046]	-0.0006 [-0.0015, 0.0000]	-0.0006 [-0.0015, -0.0001]
	05-17	1	0.0370	0.0738	0.0142 [0.0138, 0.0147]	0.1733 [0.1645, 0.1823]	-0.0001 [-0.0006, 0.0003]	0.0000 [-0.0005, 0.0003]

Price coarsening (D4) is specified but not run: the unicity basis in Table S20 and Table S23 is carried by exact size and one-second timing, not price, so D4 would answer a question the evidence does not raise, and no claim in this paper depends on it. Also disclosed as not run: an adaptive (retrain-on-defended-data) encoder attacker. This closes the defense program: every tripwire firing across the three panels is either discharged with a named mechanism or converted into a measured, claim-scoped finding.

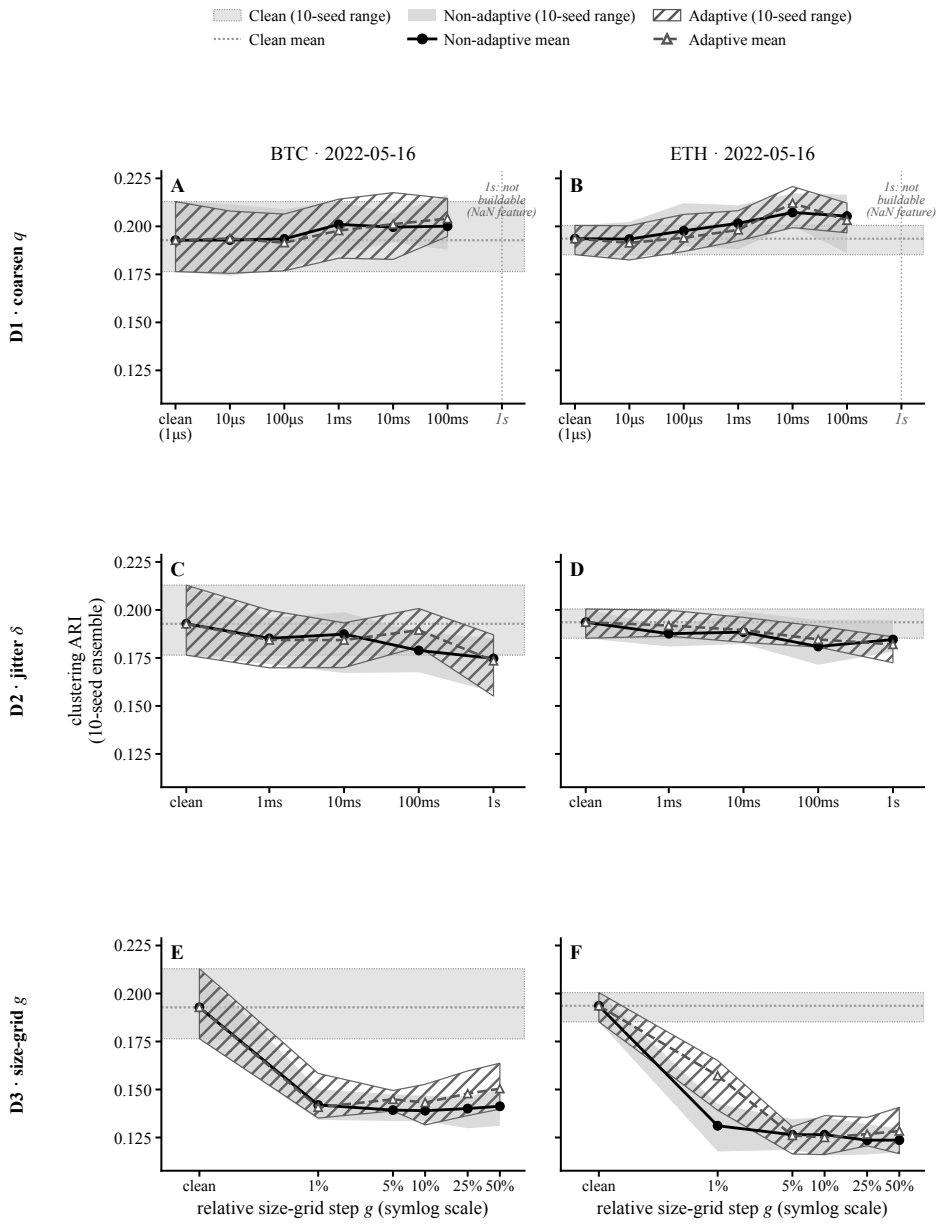


Figure S4: The y-axis in every panel is trimmed to resolve the ranges and does not start at 0; the permutation-null 95th percentile (0.0002) lies two to three orders of magnitude below the plotted range in every panel and is disclosed here rather than drawn to scale. Clustering frontier for the three defenses: each panel plots instance-level clustering ARI against defense magnitude for one (defense, session) cell, native-microsecond regime, oracle-K k-means, BTC and ETH 2022-05-16: A/B, timestamp coarsening (quantum q, log scale); C/D, timestamp jitter (half-width δ, symlog); E/F, size coarsening (relative grid step g, symlog). Solid gray ranges with filled circles, the non-adaptive attacker; hatched outlines with open triangles, the adaptive attacker; both are 10-seed k-means restart ensembles (range = min–max, marker = mean). The light gray horizontal strip is the clean ensemble range for the same cell. Coarsening's $q = 1$ s magnitude fails to build in both sessions and is marked, not plotted. Quantized-regime panels are omitted for legibility and appear in Tables S14, S18 and S22. Panel-by-panel readings, including the ETH-only coarsening backfire in B and the registered degradation rule met only by size coarsening, are in Section 5 and the text below.