DISSENSUS WORKING PAPER SERIES

DAI-2511

What Does Identity Access Buy? Pricing the Surveillance Increments in CBDC Transaction Monitoring

Murad Farzulla[1,2,*], Andrew Maksakov[1]

[1]Dissensus, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

February 2026

Abstract

Proposals for retail central bank digital currency increasingly accept that anti-money-laundering obligations require identity-linked transaction monitoring. The cost of that requirement has not been measured. Central-bank experiments that address privacy-preserving detection—the BIS Innovation Hub's Project Aurora, and Hertha with the Bank of England—ablate the *data-sharing* axis: what is gained when institutions pool transactions. Neither ablates the *identity* axis: what is gained from knowing who owns an account, over and above knowing which accounts move together. This paper measures that increment. We build a tiered ablation harness in which four nested evidence tiers—structure only on unlinked pseudonyms, plus pseudonymous linkage, plus identity attributes, plus watchlist access—are scored on identical data with entity-disjoint folds and entity-clustered bootstrap intervals. A pre-specified degeneracy audit fails the run if any single feature separates the classes at entity-level AUC above 0.95; the disjoint wallet-count ranges of an earlier version of this work put wallet count at AUC 1.000 by construction, which the audit rejects. Equivalence between tiers is tested by TOST rather than inferred from a null result, and a falsification world in which surveillance is built to win must return that verdict or the pipeline aborts. On synthetic data the answer depends on the detector: for a gradient-boosted model, pseudonymous linkage carries the increment and full identity access is statistically equivalent to linkage alone, robustly across a swept generating parameter; for a logistic model, identity access is decisively superior on average precision at every point of that sweep. In the regimes tested, then, the marginal value of identity access is model- and metric-dependent rather than a fixed property of the data. We report the bounds this evidence cannot cross—synthetic generation, a few hundred positive cases, two model families, and the fact that no identified public anti-money-laundering benchmark jointly exposes the behavioural, linkage, identity and watchlist axes that the full ablation requires, so the quantity cannot presently be validated on real data—and set out the architecture and governance this measurement motivates and constrains, and those it does not.

Keywords: central bank digital currency, privacy, financial surveillance, anti-money laundering, identity, zero-knowledge proofs, detection, measurement

JEL Codes: E42, E58, G28, K42, O33

Research Context

This work forms part of the Adversarial Systems Research program, which investigates stability, alignment, and friction dynamics in complex systems where competing interests generate structural conflict.

The privacy-preserving CBDC question addressed here is one such dynamic: the tension between state surveillance capacity and citizen financial autonomy. The paper extends prior work on AML regulatory failure (Farzulla, 2026), which documented how current frameworks target primitive laundering while sophisticated actors exploit derivatives and offshore structures with impunity; where that analysis diagnosed the problem, this paper asks what one of its assumed remedies, identity-linked monitoring, is actually worth.

Acknowledgments

The authors acknowledge David Chaum's foundational work on eCash and blind signatures, and the subsequent CBDC proposal with Grothoff and Moser (Chaum et al., 2021) demonstrating that privacy-preserving central bank money is institutionally viable. Andrew Maksakov contributed the PET–AML stack simulation (`pet_aml_sim.py`, Section 8). This paper benefited from collaboration with Claude (Anthropic) on analysis, literature synthesis, and drafting; the authors take full responsibility for all claims, errors, and interpretive choices. Related work in the series is available through the Adversarial Systems & Complexity Research Initiative (ASCRI). Correspondence: murad@dissensus.ai.

Contents

1 Introduction **5**

2 What Is Already Known **6**
- 2.1 CBDC design and the place of privacy . . . 6
- 2.2 Privacy-preserving compliance is a mature research program . . . 6
- 2.3 The unmeasured axis . . . 7

3 The Measurement Problem **7**

4 Instrument **8**
- 4.1 Generating process . . . 8
- 4.2 Evidence tiers . . . 8
- 4.3 Evaluation and its guards . . . 9

5 Results **10**
- 5.1 The operating point . . . 10
- 5.2 The increments, and the mechanism behind them . . . 11
- 5.3 Equivalence, and where the metrics disagree . . . 11
- 5.4 Robustness across the generating range . . . 12
- 5.5 The instrument can return the opposite answer . . . 12
- 5.6 Interpretation . . . 13

6 What the Measurement Licenses, and What It Does Not **13**
- 6.1 What it licenses . . . 13
- 6.2 What it does not license . . . 14

7 The Architecture the Measurement Licenses **14**
- 7.1 Design principles . . . 14
- 7.2 Detection, intervention, and resolution . . . 15
- 7.3 The identity firewall . . . 15
- 7.4 The abandonment mechanism and its deterrent . . . 15
- 7.5 Targeted investigation remains available . . . 16
- 7.6 Adoption and scale . . . 16
- 7.7 An unresolved tension: tipping-off . . . 16

8 Feasibility of the Stack **17**
- 8.1 Cryptographic requirements . . . 17
- 8.2 Feasibility assessment . . . 17
- 8.3 System architecture . . . 18
- 8.4 A concrete PET–AML stack . . . 18
  - 8.4.1 Threat model . . . 18
  - 8.4.2 Components . . . 18
  - 8.4.3 Performance envelope . . . 19
  - 8.4.4 Simulation: feasibility of the composition . . . 20

9 Governance **21**

10 Limitations and Future Work **21**

11 Conclusion **22**

1 Introduction

Every serious proposal for a retail central bank digital currency accepts, more or less explicitly, that anti-money-laundering obligations require identity-linked transaction monitoring. The design question is then taken to be how much privacy can be preserved *around* that requirement—how to encrypt, tier, threshold, or delay access to identity data whose collection is treated as settled. The requirement itself is rarely examined, and its price has never been measured. This paper measures it.

The requirement is worth examining because the surveillance it justifies performs poorly on its own terms. Estimates of laundered funds sit at two to five percent of global output (United Nations Office on Drugs and Crime, 2011), and of that flow the fraction interdicted is very small: reviewing the evidence, Pol (2020) concludes that seizures run at "some undetermined fraction of one percent," a rate at which, in his words, "the disruption of criminal funds hardly constitutes a rounding error." A system that imposes identity monitoring on an entire population to recover under one percent of the target flow is not obviously buying much with the identities it collects, and the burden should be on the architecture to show that it is.

The enforcement record compounds the doubt, because the friction and the failure fall on different people. Structuring, money mules, and basic layering—the primitive methods available to individuals, often coerced ones—are prosecuted vigorously, while the institutions through which the largest flows pass are resolved with financial penalties. When Danske Bank's Estonian branch was found to have processed on the order of $160–200 billion in suspicious non-resident transactions, the resolution was a guilty plea and more than $2 billion in forfeiture to the U.S. Department of Justice, with a further $413 million to the SEC (Lynch, 2022); the individuals criminally charged were branch-level staff rather than group leadership. Two billion dollars is a large number and roughly a cent on each dollar of the flow it settled. The asymmetry is not that enforcement is absent but that it lands hardest on the least sophisticated participants, which is precisely the population for whom identity monitoring is most totalizing and least necessary.

Public resistance to that monitoring is well documented and specific. Respondents to the European Central Bank's digital-euro consultation ranked privacy the most important feature of the instrument, ahead of security (European Central Bank, 2021); in a Cato Institute/YouGov survey, 74 percent of U.S. adults said they would oppose a CBDC if it let the government control how they spend, and 68 percent if it let the government monitor their spending (Cato Institute, 2023). Whatever one makes of advocacy-sponsored polling, the direction is consistent across sources and the objection is not to digital money as such but to the identity-linked visibility its proposed architectures assume.

So the question this paper asks is narrow and, we think, prior to the architecture debate: granted a detector operating on transaction data, what is the marginal value of each additional increment of identity access? Concretely, we separate three increments a monitoring regime might be granted in sequence—the ability to resolve pseudonymous wallets to a common owner (*linkage*), the attachment of verified identity attributes to that owner (*identity*), and consultation of a watchlist (*watchlist*)—and measure what each buys in detection performance over the one below it, holding the detector and the data fixed. The decomposition is the contribution: it isolates the value of *identity per se*, above and beyond the linkage of activity, which no prior evaluation we are aware of has done (Section 2).

The answer, developed in Sections 3–5, is that the marginal value of identity access is not a property of the data but of the detector applied to it. A detector capable of exploiting the interaction structure of behavior recovers almost everything from pseudonymous linkage alone, and identity attributes add nothing measurable on top; a capacity-limited detector cannot recover that structure and leans on identity

attributes to substitute for what it cannot infer. Detection capability and identity collection are, to a meaningful degree, substitutes. We establish this on a purpose-built measurement instrument rather than by architectural argument, we show it is robust across the range of a swept generating parameter rather than an artifact of one synthetic world, and we are explicit about the ceiling on the evidence: no public anti-money-laundering dataset carries an identity axis, so the quantity we decompose cannot at present be measured on real data by anyone, which is exactly why a controlled synthetic instrument is required to pose the question at all.

The remainder of the paper is organized around that instrument. Section 2 places the work against the CBDC design literature and the cryptographic-compliance lineage it builds on. Section 3 states why a naive ablation manufactures its own answer, using a failed earlier version of this work as the cautionary case. Section 4 describes the instrument and its guards, and Section 5 reports what it finds, including the robustness surface and the falsification check. Sections 6–8 then set out the architecture and cryptographic stack this measurement licenses—and the parts of it the measurement does *not* support—followed by governance (Section 9) and the limitations that bound the whole (Section 10).

2 What Is Already Known

2.1 CBDC design and the place of privacy

It is tempting to motivate a privacy-preserving architecture by asserting that the CBDC design literature treats surveillance as a default and privacy as an afterthought. That assertion does not survive contact with the literature, and we do not make it. Auer and Böhme (2020), whose taxonomy of direct, indirect, and hybrid architectures is the field's common reference, make central-bank visibility the very axis that distinguishes their categories rather than a constant across them: in the indirect model the central bank "keeps no record of individual claims," while only in the direct model does it hold "a record of all balances." They treat privacy as a design variable to be engineered, proposing that users "retain cryptographic proofs of their CBDC balances, rather than oblige the central bank to hold them," and endorsing "unlinkable pseudonyms, as envisaged in Chaum's pioneering work"—which is to say, the architecture this paper later specifies. The comprehensive review by Auer et al. (2022) likewise organizes itself partly around privacy, ties the privacy question directly to architecture, and observes that a design "in which the central bank has no information on retail transactions" is available. The design literature has not ignored privacy; it has, if anything, anticipated the architectural direction taken here.

What the design literature does tend to do, and what the framing documents make explicit, is treat the privacy–surveillance relation as a balance to be struck rather than a quantity to be measured. The ECB frames the goal as the "appropriate balance between privacy and other public policy objectives, like countering money laundering" (European Central Bank, 2024); the Federal Reserve's discussion paper poses the design problem as striking "an appropriate balance … between safeguarding the privacy rights of consumers and affording the transparency necessary to deter criminal activity" (Board of Governors of the Federal Reserve System, 2022). The metaphor of balance presupposes that the two quantities trade off along a known frontier. Whether they do, and at what exchange rate, is an empirical question that the framing documents pose but do not answer—and that this paper takes up.

2.2 Privacy-preserving compliance is a mature research program

Nor is the idea of reconciling compliance with privacy new. A cryptographic-compliance lineage stretches back three decades, from the trustee-based e-cash of Camenisch et al. (1996), through compact e-cash (Camenisch et al., 2005) and accountable privacy for decentralized ledgers (Garman et al., 2016), to recent systems that combine regulatory tracing with transaction privacy—Platypus (Wüst et al., 2022),

PEReDi (Kiayias et al., 2022), and the Privacy Pools construction (Buterin et al., 2024). Each of these establishes that selective, warranted, or threshold-gated disclosure can coexist with default transaction privacy. The contribution of the present work is not to add another such construction; the primitives it later assembles (Section 8) are drawn from this literature rather than invented here.

2.3 The unmeasured axis

The gap this paper addresses is narrower and empirical. Central banks have begun to measure privacy-preserving detection directly, at deployment scale: the BIS Innovation Hub's Project Aurora (BIS Innovation Hub, 2023) and the Hertha collaboration with the Bank of England (BIS Innovation Hub and Bank of England, 2025) both quantify what privacy-preserving analytics recover relative to full-data access. But both ablate the same axis—*data sharing*, the gain from pooling observations across institutions—and neither separates the *identity* axis from the *linkage* axis beneath it. Neither asks what knowing who owns an account buys over and above knowing which accounts move together. The nearest work in the academic literature that varies identity information, which ablates socio-demographic profile attributes for transaction classification, does so on an already entity-resolved graph with no unlinked-pseudonym tier and no watchlist axis, so it cannot make the separation this paper targets. Isolating the value of identity per se—distinct from linkage below it and watchlist access above it—is, to our knowledge, unmeasured, and the rest of this paper measures it.

3 The Measurement Problem

The quantity we want is the marginal value of each surveillance increment: how much detection performance is gained by resolving pseudonymous wallets to entities, how much is gained on top of that by attaching verified identity attributes to those entities, and how much is gained again by consulting a watchlist. Stated that way the measurement looks straightforward—train a detector on nested feature sets and difference the results. It is not, and the reason it is not is the central methodological problem of this paper.

An ablation over synthetic data measures the data-generating process, not the world. If the process that labels an entity a launderer also endows it with a distinguishing marginal feature available at some tier, that tier wins by construction, and the ablation reports the analyst's modeling choice back to them with a confidence interval attached. The failure is not subtle, but it is easy to commit and hard to see from inside, because the resulting numbers look like findings.

We can be concrete, because an earlier version of this work committed exactly this error. In that design, laundering entities were specified to control three to six wallets and legitimate entities one to two, with no overlap between the ranges. Wallet count therefore separated the two classes perfectly, and any detector with access to it achieved an area under the ROC curve of 1.000. The reported finding—that watchlist access added no marginal value, moving detection from 1.00 to 1.00—was arithmetically forced rather than observed: a saturated ceiling admits no improvement, so the ablation could not have returned anything else. The finding was not weak evidence for the paper's thesis. It was no evidence at all, and it pointed in the thesis's favor.

The error generalizes in both directions, which is what makes it dangerous rather than merely embarrassing. A process that places the discriminating signal in a structural feature manufactures the conclusion that surveillance is unnecessary; a process that places it in an identity attribute manufactures the conclusion that surveillance is essential. Neither conclusion survives contact with a differently specified process, and an ablation reported without a test for this condition carries no information about which case obtains.

An instrument adequate to the question therefore needs four properties, and the rest of this section states them as requirements before Section 4 describes how they are met. First, laundering must be a *latent behavior* realized over time—a sequence of placement, layering, and integration decisions—rather than a labeled trait that some feature reads off directly. Second, the instrument must *test itself for degeneracy* before any headline number is computed, and must fail loudly rather than warn. Third, the equivalence of two tiers must be established by an *equivalence test*, since failing to reject a difference is not evidence that none exists, and with a small positive class it is very weak evidence indeed. Fourth, the instrument must be shown to be *capable of returning the opposite answer*: a harness that reports "surveillance adds nothing" on every input it is given is not measuring anything.

4 Instrument

4.1 Generating process

Entities transact over a simulated 120-day window. Of 8,000 entities, a base rate of 0.05 designates launderers; 388 are realized in the seeded run reported here. Each entity controls a number of wallets drawn from a *single shared distribution* applied to both classes, which is the minimal fix for the failure described above: wallet count carries no class information by construction, and any predictive power it acquires must come from how wallets are used.

Laundering is realized as behavior rather than assigned as an attribute. A laundering entity runs placement, layering, and integration cycles over the window, splitting value across its wallets, routing through external counterparties, and structuring individual transfers beneath a reporting threshold with probability 0.6. An obfuscation parameter of 0.6 governs how aggressively the entity spreads activity to frustrate aggregation. Legitimate entities transact under the same mechanics without the cycle structure. The classes are consequently distinguished by *patterns of use* that a detector must reconstruct, not by any quantity it can read.

Identity attributes are drawn from overlapping class-conditional distributions rather than disjoint ones: KYC tier, account age, prior suspicious-activity-report count, and a jurisdiction risk score each differ in distribution between classes while assigning no entity a determinative value. The watchlist is imperfect in both directions, with a true-positive rate of 0.4 and a false-positive rate of 0.02, which is the property that makes the fourth tier a realistic increment rather than a label leak.

4.2 Evidence tiers

Four nested tiers correspond to four levels of access a monitoring regime might be granted:

- **T1 — structure only.** Sixteen per-wallet flow and timing features—volumes, degrees, inter-transfer lags, burstiness, proximity to the reporting threshold—modeled at *wallet* level, with an entity scored by the maximum score across its wallets. The regime observes individual pseudonymous accounts and can flag one that behaves badly, but cannot group accounts by owner.
- **T2 — plus pseudonymous linkage.** Modeling moves to entity level over nineteen features that only linkage makes computable: wallet count, cross-wallet means and maxima, internal transfer volume, chained fast-follow value, external counterparty count. The regime knows which accounts move together but not who owns them.
- **T3 — plus identity attributes.** Verified KYC tier, account age, prior SAR count, and jurisdiction risk are attached to each entity.
- **T4 — plus watchlist.** A watchlist membership bit is added.

The upper three tiers are strictly nested at entity level, T2 $\subset$ T3 $\subset$ T4, so the identity and watchlist increments are each the value of exactly one additional access right holding everything else fixed. T1 is deliberately *not* a subset of T2: it is a different unit of analysis, because the absence of linkage is not the removal of a column but the inability to form the entity-level row at all. The T1$\rightarrow$T2 step therefore prices a change of modeling regime rather than a change of feature set, and we report it as such.

One consequence of this design should be stated plainly rather than left for a reader to find. Even at T1, the *evaluation* is entity-level: wallets inherit their owner's cross-validation fold, which is required to prevent an entity's wallets from straddling the train and test partitions, and entity scores are formed by aggregating across an entity's wallets. The T1 detector is denied linkage as a feature; the T1 protocol still assumes ground-truth linkage for scoring. T1 therefore measures a regime that cannot *exploit* linkage rather than one in which linkage is unavailable in principle, and the true structure-only floor is somewhat below the figure reported here.

This is the decomposition the paper contributes: the BIS Innovation Hub's Project Aurora (BIS Innovation Hub, 2023) and the Hertha collaboration with the Bank of England (BIS Innovation Hub and Bank of England, 2025) both ablate the *data-sharing* axis, asking what is gained when institutions pool observations, and neither separates the *identity* axis from the linkage axis beneath it.

Two model classes are fit at every tier: a logistic regression and a histogram-based gradient boosting classifier. This is not a horse race, and we do not report a winner. The contrast between a model of limited capacity and one able to exploit interactions among behavioral features is itself the finding of Section 5, and a single-model design would have concealed it.

4.3 Evaluation and its guards

Scoring is at entity level with entity-disjoint cross-validation folds: no wallet belonging to an evaluated entity appears in training, which prevents the linkage tier from leaking its own answer. Confidence intervals are entity-clustered bootstraps, so the interval reflects the number of independent entities rather than the number of transactions. We run 8,000 entities at a base rate of 0.05, yielding on the order of 400 realized launderers; an earlier version of this work ran 800, and Section 5 shows why that was too few—at $\sim$29 positive cases every average-precision comparison is inconclusive, and the study's whole verdict rested on a single saturating metric.

Performance is measured on two metrics, not one, and the distinction turns out to carry the result. Area under the ROC curve saturates at these operating points—once the study is adequately powered every tier sits above 0.98 for the capable model, so "equivalent within a small margin" is close to unfalsifiable and the verdict conveys little. Average precision does not saturate at $\sim$5% prevalence, and it is the quantity a monitoring regime actually operates on when it allocates a fixed alert budget. The two can disagree, and where they do the disagreement is informative rather than a nuisance (Section 5).

A **pre-specified degeneracy audit** runs before any headline number and terminates the run if any single feature achieves entity-level AUC above 0.95, on the reasoning that a lone feature that nearly separates the classes indicates the generating process has leaked the label. In the runs reported here the worst single feature reaches 0.840 (mean fraction of fast-follow transfers) in the primary world and 0.887 (jurisdiction risk) in the falsification world, both below the threshold. The disjoint wallet-count ranges of the earlier version described in Section 3 put wallet count at an entity-level AUC of 1.000 by construction—perfect separation of two non-overlapping integer ranges—so the audit rejects that generating process outright and refuses to produce results at all.

Tier equivalence is assessed by two one-sided tests. We fix the margin at $\delta = 0.03$; this was chosen on the AUC scale before any average-precision result existed, so applying the same number to average

precision—whose baseline is prevalence rather than 0.5—would be a post hoc margin choice on a different scale, and we flag it as such wherever it appears. To avoid resting anything on that choice, every comparison also reports its *equivalence bound*: the smallest margin at which the test would declare equivalence, which is margin-free and lets a reader apply their own threshold. We report equivalence as a positive finding only where the test supports it, INCONCLUSIVE where it does not, and we never infer equivalence from a failure to reject a difference.

Two guards close the loop against the instrument flattering its author. First, the harness must demonstrate that it *can* return the opposite verdict: a second world, `surveillance_strong`, is parameterized so identity attributes genuinely dominate behavioral structure, and the pipeline exits non-zero unless that world returns SURVEILLANCE_SUPERIOR on at least one comparison, so a regression that biased the instrument toward the paper's thesis breaks the build rather than producing a publishable number. Second, because the deepest objection to any synthetic ablation is that the analyst chose the generating process, the primary result is reported not at one operating point but across a sweep of the process's most consequential free parameter (Section 5.4), so what the paper claims is an invariant of the sweep rather than a fact about a single world.

5 Results

All numbers in this section come from the seeded pipeline at the canonical operating point (8,000 entities, base rate 0.05, 388 realized launderers, seed 20260707), except Section 5.4, which sweeps the generating process, and Section 5.5, which reports the falsification world. Every figure regenerates from committed code.

5.1 The operating point

Table 1 reports detection performance by tier for both models on both metrics, with entity-clustered 95% bootstrap intervals. Reading the two metrics side by side is the point of the table, not a formality: they tell different stories about the same scores.

Table 1: Detection performance by evidence tier (8,000 entities, 388 launderers, seed 20260707). Intervals are entity-clustered bootstraps. AUC saturates for the capable model; average precision (AP) does not.

Model	Tier	AUC (95% CI)	AP (95% CI)
Logistic	T1 structure only	0.965 [0.951, 0.977]	0.851 [0.819, 0.880]
	T2 + linkage	0.981 [0.975, 0.987]	0.859 [0.829, 0.886]
	T3 + identity	0.987 [0.983, 0.992]	0.894 [0.868, 0.918]
	T4 + watchlist	0.992 [0.989, 0.995]	0.922 [0.899, 0.943]
Gradient boosting	T1 structure only	0.987 [0.981, 0.992]	0.912 [0.891, 0.933]
	T2 + linkage	0.998 [0.996, 0.999]	0.980 [0.971, 0.989]
	T3 + identity	0.998 [0.997, 0.999]	0.981 [0.971, 0.989]
	T4 + watchlist	0.999 [0.998, 0.999]	0.985 [0.977, 0.992]

The AUC columns for the gradient-boosted model are visually flat—0.987 to 0.999 across four tiers—and this is the trap the earlier version of this work fell into. When every tier scores above 0.98, differences among them are compressed into the last percent of a bounded scale, and any equivalence test with a plausible margin will return equivalence almost regardless of what the tiers actually contain. Average precision, which weights the ranking of the $\sim$5% of entities that are positive rather than the

whole ROC surface, keeps its resolution: the logistic model moves from 0.851 to 0.922 across the tiers, room in which real differences can show.

5.2 The increments, and the mechanism behind them

Table 2 differences adjacent tiers to give the marginal value of each access right. We read the average-precision column, for the reason just given.

Table 2: Marginal gain from each additional access right at the canonical operating point. Read the AP column; the AUC column is shown to make the saturation visible.

Model	Metric	Linkage (T1→T2)	Identity (T2→T3)	Watchlist (T3→T4)
Logistic	AP	+0.009	+0.035	+0.027
	AUC	+0.016	+0.006	+0.005
Gradient boosting	AP	+0.069	+0.000	+0.004
	AUC	+0.011	+0.000	+0.001

The two models allocate the gain in opposite ways, and the contrast is the finding. On average precision the gradient-boosted model takes +0.069 from linkage and then nothing from identity (+0.000) and almost nothing from the watchlist (+0.004). The logistic model takes almost nothing from linkage (+0.009) and then +0.035 from identity and +0.027 from the watchlist.

This is not two noisy estimates of one quantity; it is a mechanism. The signal that separates launderers from legitimate entities lives in the *interaction structure* of behavior across an entity's wallets—how fast value chains between accounts, how activity bursts and disperses. A gradient-boosted model can exploit that structure once linkage lets it see all of an entity's wallets together, which is why linkage carries its entire gain and identity attributes add nothing a sufficiently capable model has not already recovered from behavior. A linear model cannot represent those interactions, so linkage buys it little; it leans instead on identity attributes, which supply the class signal directly rather than through structure it cannot model. Identity access substitutes for a capability the weak model lacks and the strong model does not need.

5.3 Equivalence, and where the metrics disagree

Two one-sided tests compare the identity-free tier (T2) against full access (T4). The verdicts depend on both the model and the metric, and Table 3 shows how.

Table 3: TOST equivalence, T2 (linkage only) vs T4 (full surveillance access), $\delta = 0.03$. The equivalence bound is the smallest margin that would yield EQUIVALENT and is margin-free. δ was set on the AUC scale; its use on AP is flagged post hoc (Section 4).

Model	Metric	Δ (90% CI)	Equiv. bound	Verdict
Gradient boosting	AUC	+0.001 [−0.000, +0.002]	0.002	EQUIVALENT
	AP	+0.004 [−0.001, +0.010]	0.010	EQUIVALENT
Logistic	AUC	+0.011 [+0.007, +0.015]	0.015	EQUIVALENT
	AP	+0.062 [+0.048, +0.078]	0.078	SURVEILLANCE SUPERIOR

For the gradient-boosted model, T2 and T4 are equivalent on both metrics, with an average-precision equivalence bound of 0.010: full surveillance access buys the capable detector essentially nothing over pseudonymous linkage, and that holds at any margin above one AP point.

For the logistic model the two metrics part ways, and the disagreement is diagnostic rather than awkward. On AUC the comparison reads EQUIVALENT (+0.011); on average precision it reads SURVEILLANCE SUPERIOR (+0.062, the entire interval above the margin). The same scores, the same tiers, and the same test return opposite verdicts because AUC has saturated and average precision has not. This is the concrete form of the warning in Section 4: the retracted analysis certified its central null on AUC, the one metric that cannot see the effect it was used to rule out.

The intervals in Table 3 condition on a single generating-process realization and a single cross-validation fit; they capture entity-level resampling variance but not the variance from the generator seed or the fold assignment. The equivalence bound of 0.010 is therefore a within-one-world quantity. The sweep of the next subsection bounds the missing generator variance directly: across its fifteen conditions the gradient-boosted T2-vs-T4 average-precision equivalence bound ranges from 0.010 to 0.036, so the capable model's identity increment stays within four average-precision points of zero even under the least favorable seed and obfuscation level.

5.4 Robustness across the generating range

Because the deepest objection to a synthetic ablation is that its answer is an artifact of the chosen generating process, we do not report the result at a single operating point. We sweep the process's most consequential free parameter—laundering `obfuscation`, the degree to which laundering entities spread and disguise their behavior—across five levels from 0.5 to 0.9, at three seeds each, holding everything else fixed. Fifteen conditions per model. The degeneracy audit passes at every one.

The point estimates wander, as point estimates from 400-odd positive cases do. The *verdict* does not. Table 4 reports the T2-vs-T4 equivalence outcome on average precision across all fifteen conditions.

Table 4: T2-vs-T4 equivalence on average precision across the obfuscation sweep (5 levels × 3 seeds). The invariant, not the point estimates, is the result.

Model	SURV. SUPERIOR	EQUIVALENT / INCONCL.	Δ range
Logistic	15 / 15	0 / 15	+0.048 to +0.076
Gradient boosting	0 / 15	14 / 1	+0.004 to +0.025

Across the whole range, the logistic model finds full surveillance access decisively superior to linkage—SURVEILLANCE SUPERIOR in every one of fifteen conditions—and the gradient-boosted model never does, in any condition, on either metric (0 of 60). The capable model's average-precision verdict is EQUIVALENT in fourteen of fifteen conditions and INCONCLUSIVE in one (highest obfuscation, one seed); it is never SURVEILLANCE SUPERIOR. So the strongest honest statement about the capable model is that the identity increment is equivalent to zero or, at worst, not distinguishable from it—never that surveillance wins. The gap between the two models is not a feature of one world. It is an invariant of the generating process across the range we can move it.

5.5 The instrument can return the opposite answer

An ablation harness that reports "identity adds nothing" on every input is measuring nothing. The falsification world `surveillance_strong` is parameterized so that identity attributes genuinely dominate behavioral structure, and the pipeline aborts unless the harness detects it. It does: on average precision, T2-vs-T4 returns SURVEILLANCE SUPERIOR for both models (+0.145 gradient boosting, +0.437 logistic). The same code that reports equivalence in the primary world reports the opposite here.

One detail in this world is worth stating because it closes the argument of Section 5.3 from the other

side. For the gradient-boosted model in the falsification world, average precision returns SURVEILLANCE SUPERIOR (+0.145) while AUC returns INCONCLUSIVE (+0.033). Here surveillance genuinely does win, by construction—and AUC cannot see it. A metric that fails to detect a deliberately planted effect cannot be used to certify that effect's absence, which is precisely what the original AUC-only analysis did.

5.6 Interpretation

The marginal value of identity access is not a fixed property of financial data. It is a function of the detector applied to that data: a model that cannot reconstruct behavior from structure needs identity attributes to substitute for what it cannot infer, and a model that can reconstruct behavior finds the same attributes add nothing measurable. Detection capability and identity collection are to a meaningful degree substitutes, and a monitoring regime that treats identity access as a fixed requirement is holding its own analytical capability constant without acknowledging that it has done so. This holds across the generating range, on the metric that a monitoring regime actually operates on, and it reverses cleanly when the world is built so that it should.

Two qualifications belong here rather than in a later section. The first is that the linkage step, which carries the capable model's entire gain, is itself a surveillance capability: resolving pseudonymous wallets to a common owner is exactly the inference pseudonymity is meant to frustrate. The result of this paper is that the identity increment *above linkage* is small—a genuine finding, and a narrower privacy claim than "monitoring is unnecessary." The second is that this is measured on synthetic data, for a reason Section 10 develops: no public anti-money-laundering dataset carries an identity axis, so the quantity decomposed here cannot at present be measured on real data by us or anyone else. The synthetic instrument is not a convenience; it is currently the only instrument that can pose the question at all.

6 What the Measurement Licenses, and What It Does Not

The result of Section 5 is narrow, and it is worth being precise about what architectural conclusions it supports before the rest of the paper draws them. The measurement licenses one claim and forbids several tempting extensions of it.

6.1 What it licenses

For a detector able to exploit the interaction structure of behavior, full identity access is statistically equivalent to pseudonymous linkage alone, robustly across the generating range: the identity increment above linkage is equivalent to zero, or at worst indistinguishable from it, and never favorable to surveillance (Section 5.4). What this licenses is an architecture in which detection operates on *linked but pseudonymous* activity—an entity's wallets resolved to a common owner whose civil identity is not attached—and identity is disclosed only on a warranted, per-case basis rather than collected by default. The detection layer specified in Section 7 is built to that specification, and the measurement is what justifies it: identity-by-default is not buying the detection performance it is assumed to buy.

This also answers the objection that graph-only detection is simply worse and the privacy is paid for in missed crime. At the operating point, the capable detector's identity-free tier reaches an area under the curve of 0.998 and average precision of 0.980, statistically equivalent to full access; the performance is not sacrificed, it is recovered from structure. The privacy-preserving configuration is not a worse detector with a privacy consolation. It is, for a capable model, the same detector.

6.2 What it does not license

Three extensions do not follow, and the architecture in the following sections is careful not to assume them.

First, the result is about the increment *above linkage*, and linkage is itself a surveillance capability. Resolving an entity's pseudonymous wallets to a common owner is exactly the inference that unlinkable pseudonymity is designed to frustrate, and it carries most of the capable detector's performance. The honest reading is not that monitoring is unnecessary but that the *identity* component of monitoring, above the linkage component, adds little. An architecture that leans on this result must still take a deliberate position on linkage—who may perform it, under what authority, and whether it happens by default or on cause—rather than treating the paper as a license to link freely because "identity adds nothing."

Second, the equivalence is a property of a capable detector, and the same measurement shows a capacity-limited detector obtaining decisive value from identity access at every point of the generating range. An architecture that withholds identity data on the strength of this result is implicitly committing its operator to maintaining detection capability at the level the result assumes. That is a reasonable commitment, but it is a commitment, and it should be stated as an operational requirement rather than assumed away.

Third, the finding is measured on synthetic data, and the falsification world demonstrates that a differently constructed process returns the opposite verdict. The result is evidence that identity adds little *in a regime where laundering is a latent behavior recoverable from structure*; it is not evidence that this regime describes any particular real deployment. Section 10 takes up why the real-data question cannot presently be settled and what would settle it.

7 The Architecture the Measurement Licenses

The measurement supports an architecture in which detection operates on linked-but-pseudonymous activity and civil identity is disclosed only on cause. This section specifies that architecture. Its cryptographic components are drawn from the privacy-preserving-compliance lineage of Section 2 rather than invented here; what the measurement adds is a reason to prefer this configuration over an identity-by-default one, namely that identity-by-default is not buying the detection value it is assumed to buy. Two design elements—the abandonment primitive and opt-in deanonymization—are where the privacy headroom the measurement identifies is actually spent.

7.1 Design principles

Separation of pattern detection from identity. Transaction-graph analysis can identify structural anomalies—velocity, unusual counterparty structure, layering motifs—without knowing participant identities. The detection system outputs suspicion scores for transactions, never for users.

Transaction-level intervention. Suspicious activity affects the specific flagged transaction, not the user's broader financial access. A flagged payment is delayed; the user's ability to transact otherwise is untouched.

Opt-in deanonymization. Identity revelation is voluntary. A user may explain a flagged transaction, disclosing whatever they consider appropriate, or abandon it without explanation and without any record linking identity to the flagged pattern.

Architectural enforcement. Privacy guarantees are structural, not policy. The analytic components do not possess the capability to link transactions to civil identities without active user participation, which is a stronger guarantee than a promise to refrain from using such a capability.

7.2 Detection, intervention, and resolution

The **Detection Layer** ingests anonymized transaction data—amounts, timestamps, pseudonymous counterparty tokens—and maintains a rolling transaction graph it cannot deanonymize. A pattern engine compares structures against publicly auditable typology libraries, an anomaly system flags novel structures, and a scorer emits per-transaction suspicion scores. Its output is "transaction X has suspicion score Z," never "user Y is suspicious."

The **Intervention Layer** freezes flagged transactions in escrow and notifies both parties through the ordinary confirmation channel, without additional identifying information. This notification is where the architecture departs from the current reporting regime, and Section 7.7 states that tension rather than eliding it. Parties may explain, request time, or abandon; none of these requires identity disclosure. On timeout the transaction is canceled, funds return to sender, and no record links any identity to the flagged pattern.

The **Resolution Layer** receives voluntary explanations together with a zero-knowledge proof that the explainer is a party to the transaction, without revealing which party. Trained reviewers see the explanation, the suspicion rationale, and the amount; they do not see identities, account histories, or other transactions. Decisions—release or maintain freeze—are logged for audit, and anonymized outcomes feed back into detection so the system learns from resolved cases without holding identities.

7.3 The identity firewall

The firewall is what makes the privacy guarantee structural. Transaction-graph data and identity mapping are held by separate systems with no shared access; identity lives only in the user-facing Wallet Layer and is never transmitted to the analytic components. When a user chooses to explain, a zero-knowledge proof establishes party status without revealing identity. There is no backdoor: the components that perform analysis simply do not hold identity information, so state actors cannot compel per-transaction identity revelation without user cooperation or threshold activation. Where governance authorizes exceptional access—a court-ordered investigation—a threshold scheme requires multiple independent parties to authorize decryption for specific transactions, making bulk surveillance technically rather than merely legally unavailable.

7.4 The abandonment mechanism and its deterrent

The abandonment primitive is where the measurement's privacy headroom is spent, and its strategic structure is worth making explicit. The analysis here is informal: it identifies the choices a flagged actor faces and reasons about incentives, without deriving equilibria, and Section 10 takes up the repeated-interaction dynamics it leaves open.

A flagged actor has three options: explain and proceed, which for a genuinely illicit transaction requires producing a legitimate narrative and may create evidence; abandon silently, canceling the transaction with funds returned and no identity linked; or wait for timeout, equivalent to abandonment with delay. The deterrent follows from the first two. A laundering operation that cannot move money has failed in its core objective, regardless of whether anyone is identified, and unlike current systems—where sophisticated actors "comply through" by producing documentation—there is no comply-through option here: either a transaction can be legitimately explained, in which case it may well be legitimate, or it cannot proceed. Aggregate abandonment patterns also feed detection without identifying anyone.

For legitimate users the friction is low. Detection tuned for high-confidence anomalies flags a small, governance-controlled fraction of legitimate transactions; a flagged legitimate payment usually has a ready explanation that need only be plausible, not formally verified; abandonment imposes no penalty

beyond transaction failure for a user who prefers not to explain; and a flag never carries account-level consequence. Relative to surveillance-based designs, this trades away a class of harms: no account freezes or investigations from false positives, no repurposing of monitoring infrastructure for ends beyond its original mandate, and no honeypot of identity-linked transaction data to be breached, because that data is never assembled.

7.5 Targeted investigation remains available

The architecture prevents mass retrospective search, not investigation. Subpoenas can compel individuals to produce records; court orders can require wallet providers to disclose identity for specific users; undercover methods remain; and threshold activation can enable identity access for specific transactions when governance authorizes it. What is foreclosed is bulk access—the ability to search all transactions retrospectively for patterns associated with individuals—which is not required for targeted investigation and is the specific capability whose collection the public objects to. Rennie and Steele (2021) distinguish four privacy "losses" a CBDC can impose—anonymity, liberty, individual control, and regulatory control—and argue each needs separate treatment rather than aggregation; the state's capacities differ qualitatively from commercial data collection, since commercial entities cannot freeze property or incarcerate (International Monetary Fund, 2024). The architecture implements the principle that identity access be gated by due process structurally rather than by policy commitment.

7.6 Adoption and scale

The objection that no state would adopt a privacy-preserving CBDC has weakened. The U.S. Anti-CBDC Surveillance State Act (H.R. 1919 / S. 1124, 119th Congress), which passed the House in 2025, prohibits the Federal Reserve from issuing a CBDC usable "as a tool for surveilling Americans"; Zatti (2025) reads this as a shift from frameworks asking how CBDCs should be designed to ones rejecting surveillance architectures outright. The Act does not endorse the architecture proposed here, but it establishes that state opposition to surveillance CBDC has moved from civil-society concern onto the legislative agenda. Per Congress.gov the bill has passed the House; House passage is not enactment, and neither that step nor the introduction of S. 1124 makes either bill law. The paper's claim is in any case normative rather than predictive: demonstrating that a privacy-preserving alternative exists forecloses the argument that surveillance is a technical necessity, whatever any given jurisdiction chooses. On scale, the architecture assumes most transactions proceed without intervention and only flagged ones require attention; at a flag rate below a fraction of a percent, a system clearing millions of transactions daily produces thousands of flags, within reach of automated triage with human review reserved for contested cases. High-frequency or institutional flows can be placed under a distinct regime—accepting identity registration in exchange for reduced friction—without compromising retail privacy.

7.7 An unresolved tension: tipping-off

One objection has force the architecture does not dissolve. FATF Recommendation 21 requires that institutions and their staff be prohibited by law from disclosing that a suspicious-transaction report is being filed (Financial Action Task Force, 2012–2023), and national implementations such as section 333A of the UK Proceeds of Crime Act 2002 criminalize disclosures likely to prejudice an investigation. The anonymous notification at the center of the Intervention Layer does what these provisions forbid: it tells both parties, in real time, that their transaction has drawn suspicion. The honest statement is that the architecture is incompatible with the current reporting regime, not that it satisfies it. It presupposes a different regime in which prevention substitutes for covert reporting—the flagged transaction cannot complete, so there is no downstream investigation for notification to prejudice in the usual sense—and

whether regulators would accept that substitution is an open legal question that would require amending FATF standards before any deployment. Notification also interacts with the probing problem of Section 10: telling parties which transactions are flagged is exactly what makes the detection boundary observable. A deployable design might delay or coarsen notification, but each such change erodes the due-process property that motivates it. We record this as an unresolved compliance-design tension rather than a solved problem.

8 Feasibility of the Stack

The measurement of Section 5 licenses an architecture (Section 7) in which detection runs on linked-but-pseudonymous activity and identity is disclosed only on cause. This section addresses a separate question: whether that architecture is buildable from primitives that exist today. The answer we defend is feasibility, not optimality, and the distinction matters for how the simulation at the end of the section should be read—it establishes that the *composition* of known components stays within a plausible performance envelope, not that any component's cost has been independently validated.

8.1 Cryptographic requirements

The architecture relies on established primitives. Zero-knowledge proofs provide transaction validation and party authentication; the Zcash deployment demonstrates production viability of comparable requirements (Ben-Sasson et al., 2014). Secure multi-party computation enables pattern detection without any single party holding complete data. Threshold cryptography gates emergency identity access behind multiple independent authorizations. Blind signatures enable authorization without linkage; Chaum et al. (2021) give a complete CBDC issuance protocol in which the central bank cannot link withdrawals to spending. Ring signatures allow a party to prove membership in a transaction set without revealing which member it is, useful for anonymous explanation in the Resolution Layer.

8.2 Feasibility assessment

Each primitive has been demonstrated in production or at research scale: zero-knowledge proofs in Zcash's shielded transaction pool (Electric Coin Company, 2023); secure multi-party computation in privacy-preserving analytics such as Google's Private Join and Compute (Google, 2019); threshold cryptography in cryptocurrency custody and distributed key management (Gennaro and Goldfeder, 2018); blind signatures across privacy-preserving payment systems since their introduction (Chaum, 1982).

Performance remains a genuine consideration. Current ZKP systems carry significant proving overhead relative to transparent transactions, though proof generation can run client-side and verification is far cheaper than generation. Project Hamilton, the Federal Reserve Bank of Boston and MIT collaboration, showed a CBDC transaction processor sustaining on the order of 10^6 transactions per second with sub-second latency (Lovejoy et al., 2023), which establishes headroom for the additional cost of privacy-preserving primitives at national scale even though that throughput figure was measured without those primitives in the path.

A parallel ZKP-based compliance design merits comparison. Decker (2025) propose a zero-knowledge KYC architecture that reduces exposed user data substantially while reporting high fraud-detection accuracy, in the narrower setting of point-of-entry institutional verification. Their setting and ours are complementary rather than competing: their construction could serve as the identity-verification layer at account creation, while the architecture here governs ongoing transaction monitoring without identity exposure. We do not compare detection rates across the two, because they measure different quantities on different populations.

Progress toward practical viability continues. Michalopoulos et al. (2025) present a Secure Element-based offline CBDC system with latency comparable to commercial payment rails, and advances in secure multi-party computation continue to reduce communication overhead (Escudero et al., 2024). Zero-knowledge tooling in particular has matured into a substantial engineering base across blockchain platforms (El-Hajj and Roelink, 2024), though claims about universal industry adoption should be read with the usual caution about vendor reporting.

8.3 System architecture

A complete system separates five layers, each with a defined interface and no shared identity access beyond the first. The **Wallet Layer** is the only component holding user identity; it generates proofs and interfaces with the layers below. The **Transaction Layer** processes anonymized transactions and settles the ledger. The **Detection Layer** analyzes patterns on pseudonyms. The **Intervention Layer** manages flagged transactions and resolution, with identity access only where a user voluntarily discloses. The **Governance Layer** manages parameters, threshold key shares, and audit, exercising oversight without operational access. Compromise of any single layer does not enable identity compromise absent user cooperation or threshold activation.

8.4 A concrete PET–AML stack

The remaining subsections instantiate the high-level claim into a privacy-enhancing AML (PET–AML) stack that can be built today: private watchlist checking via Private Set Intersection (PSI) or VOPRF-based membership tests; secure risk propagation over pseudonymous transaction graphs; and transaction-time policy proofs in zero-knowledge. These map onto the layered architecture—the Wallet Layer screens and proves, the Transaction Layer verifies, the Detection Layer scores pseudonyms, and the Governance Layer controls exceptional exposure.

8.4.1 Threat model

We assume a wallet/PSP performing KYC and issuing an unlinkable credential; a ledger operator validating transaction proofs and maintaining settlement; a compliance authority maintaining watchlists and policy parameters; and a governance quorum holding threshold key shares for exceptional identity exposure. The adversaries of concern are honest-but-curious operators, malicious users seeking to launder while remaining unlinkable, and partial compromise or collusion among operational entities. There is deliberately no single super-admin trust anchor: identity access requires threshold activation and legal authorization. The security objectives are transaction privacy (the ledger validates payments and enforces limits without learning identity), watchlist privacy (neither the compliance authority nor the ledger learns non-matching queries or the list itself), enforceability (no valid transaction without satisfied policy constraints), and due process (disclosure is auditable, rate-limited, threshold-gated, with no bulk de-anonymization interface).

8.4.2 Components

Private watchlist checking (PSI/VOPRF). Sanctions and PEP screening are implemented as a privacy-preserving membership test. The wallet derives a stable identifier from KYC material and runs a PSI-style protocol against the watchlist. Unbalanced PSI is well-suited to this setting, in which the server holds a large set and each client query set is small (Wang et al., 2025); an alternative uses verifiable OPRFs (Davidson et al., 2023), with the wallet checking membership against a periodically published PRF-transformed watchlist. On a no-match the wallet can carry a compact witness proving screening occurred without revealing the identifier; on a match, policy triggers a dispute window in the Intervention

Layer rather than automatic identity leakage.

Secure risk propagation on pseudonymous graphs. The Detection Layer benefits from graph-structured signals but should not learn identities, so identity is treated as a sealed label and risk is computed on pseudonyms. Because even which pseudonyms are seeded as "known bad" can be sensitive, the seed vector is secret-shared while non-identifying graph structure remains available, and a bounded-iteration risk function is evaluated under secure computation. Cross-institution graphs can use secure graph-analytics frameworks that reduce edge-proportional communication overhead and have been demonstrated at million-node scale under semi-honest models (Koti et al., 2024; Yu et al., 2025). A privacy-preserving MPC pilot among Dutch banks reported by van Egmond et al. (2024) gives some evidence that computing risk signals across institutional boundaries, without sharing customer identities, is practically approachable rather than only theoretical.

ZK policy proofs at admission. Transaction-time proofs make policy constraints verifiable on entry: balance correctness and no double-spend; per-tier limits on amount and cumulative spend via range and sum proofs; credential validity with selective disclosure, optionally using threshold-issued credentials such as Coconut (Sonnino et al., 2019); and proof that screening was performed for the current epoch. SNARKs offer small proofs and fast verification at the cost of trusted setup, while STARKs are transparent and post-quantum secure with larger proofs (El-Hajj and Roelink, 2024); a jurisdiction can select a point on that frontier and upgrade over time.

8.4.3 Performance envelope

Table 5 gives an order-of-magnitude envelope. It is indicative, not normative, and the figures are drawn from the cited benchmarks rather than measured here.

Table 5: Indicative performance envelope for the PET–AML stack (order-of-magnitude; figures from the cited sources, not measured in this work).

Operation	**Latency**	**Bandwidth / storage**	**Notes**
Watchlist check (wallet $\rightarrow$ FIU)	sub-second interactive; amortizable via epochs	tens of KB per check; server stores $\lvert W\rvert$ items	Unbalanced PSI supports large $\lvert W\rvert$ at modest client cost (Wang et al., 2025)
Transaction ZK proof generation	sub-second to a few seconds on commodity hardware	proof size $\sim 10^2$ bytes (SNARK) to $\sim 10^4$ bytes (STARK)	Benchmarked range spans roughly 0.5 ms–3.6 s across circuits; verification typically milliseconds (El-Hajj and Roelink, 2024)
Secure risk propagation (batch)	minutes to a few hours per batch (e.g. daily)	state proportional to $\lvert V\rvert$; preprocessing amortizes multiplications	Secure graph frameworks report million-scale feasibility and large comms reductions (Koti et al., 2024; Yu et al., 2025)

To keep routine payments smooth, screening runs at wallet "ready" time, proofs are generated at transaction time from fixed audited circuits, and risk propagation runs in batches unless a high-risk typology requires near-real-time scoring.

8.4.4 Simulation: feasibility of the composition

To check that the composition of these components stays within the envelope, we implement a discrete-event simulation of the stack (`pet_aml_sim.py`).[1] The simulator models 40,000 wallets (eight PSPs of 5,000) across four KYC tiers in proportions 55/30/13/2%, draws transaction amounts from a lognormal distribution, and exercises PSI screening, ZK proof verification, and batch MPC risk propagation. Table 6 reports a representative run.

Table 6: PET–AML simulation, representative run (40,000 transactions, seed 7).

Metric	**Value**
Transactions generated	40,000
Transactions settled	25,287 (63.2%)
Transactions rejected	14,713 (36.8%)
rejected under tier-conditioned amounts (sensitivity)	3,375 (8.4%)
Sanctions hits (PSI blocked)	82
Escalations (case packets)	27 (0.068% of screened; 14 later settled)
Latency mean / p50 / p90 / p99	537 / 511 / 768 / 1,137 ms
Latency breakdown	PSI screening dominates
MPC batch runtime	0.38 hours

What the simulation shows, and what it does not. It shows that the composition is feasible: given per-component costs taken from the literature, queueing, contention, batching, and cross-PSP routing do not push end-to-end latency outside the retail envelope—mean 537 ms, 99th percentile 1.14 s, dominated by PSI screening. This is a statement about the architecture's composition, not a validation of the component costs themselves, which enter the simulator as inputs (Table 5); a simulation parameterized by an envelope cannot confirm that envelope, and we do not claim it does.

The 36.8% rejection rate is an artifact of the amount distribution, not a system failure mode: 14,631 of the 14,713 rejections are tier-0 per-transaction-limit breaches ($50 cap) and the daily cap never binds, because amounts are drawn from one global lognormal whose median exceeds that cap irrespective of tier. Drawing amounts per-wallet from a tier-scaled distribution (mean 40% of the wallet's tier limit, `-tier-amounts`) drops the rate to 8.4%, with per-transaction latency essentially unchanged; the rejection rate is therefore set by the ratio of typical spend to the cap rather than by the architecture. The 82 sanctions blocks confirm that PSI screening intercepts watchlist matches without revealing match status to the ledger.

[1] Source in the paper's repository, https://github.com/andrewmaksakov/CBDC.

The escalation count requires care in interpretation. An earlier version of the simulator reported zero escalations, an artifact of processing order: risk tiers were assigned during batch propagation *after* the transaction loop, so tier-based escalation could never fire. With propagation moved to run once per simulated day, the same two-day run yields 27 escalations. These are counted at the point of triggering, before settlement, so they are not a subset of the settled transactions—14 of the 27 subsequently settle and 13 are rejected, which is why the rate is reported against screened rather than settled transactions. The figure is a small, real exercise of the escalation pathway, not evidence about escalation rates in a deployed system, which would depend on the true prevalence and behavior of high-risk wallets.

9 Governance

Privacy-preserving does not mean ungoverned. Several parameters require ongoing oversight rather than technical default. Detection thresholds trade higher sensitivity against more false positives, a policy choice with distributional consequences. Timeout periods trade accommodation of legitimate users against friction on illicit abandonment. Pattern libraries determine what triggers a flag and should be publicly auditable to permit scrutiny and prevent discriminatory targeting. And the holders of threshold key shares—who can authorize exceptional identity access—should be distributed across independent institutions so that capture is hard and emergency capability is preserved.

Transparency is what makes this oversight real. Aggregate statistics on flag rates, resolution outcomes, and abandonment patterns can be published without compromising any individual, and they are what let the public assess whether the system works. Resolution decisions should be periodically reviewed by independent auditors for consistency and abuse; system source code should be public and, where feasible, formally verified; and every activation of exceptional identity access should be logged with justification, with unauthorized attempts triggering investigation.

These are not merely technical settings. Keister and Sanches (2023) show through general-equilibrium analysis that CBDC design choices carry non-trivial consequences for banking stability, deposit funding, and credit allocation, so the architectural properties of digital currency are determinants of macroeconomic structure rather than implementation details. When choices carry consequences of that magnitude they belong to democratic processes rather than technical defaults. The framework is consistent with stakes-weighted accounts of legitimacy, in which governance should track stakeholder preferences in proportion to the stakes involved (Farzulla, 2025): financial-privacy architecture bears more acutely on marginalized communities, dissidents, and the economically vulnerable than on those with institutional access, and governance should reflect that heterogeneity. The evolving European framework illustrates both the difficulty and the feasibility of legislating privacy requirements, as the digital euro's offline mode raises particular AML questions under the Anti-Money-Laundering Regulation, MiCAR, and PSD3 that architectural privacy guarantees could simplify relative to layered exemptions (Minto, 2024). Governance structures should accordingly include legislative authorization specifying privacy requirements, an independent oversight body with audit authority, public reporting on operation and decisions, mechanisms for citizen input on parameters, and sunset clauses requiring periodic reauthorization.

10 Limitations and Future Work

The central limitation is structural rather than incidental, and it bounds what any version of this study can claim. The quantity we decompose—the marginal value of identity access, separated from the linkage beneath it—requires a dataset in which entities carry both a linkage structure and verified identity attributes, with laundering labels. No public anti-money-laundering dataset has this shape. The two

standard benchmarks, the IBM AMLworld synthetic transaction data and the Elliptic Bitcoin graph, both lack an identity axis entirely: they support the linkage question (T1→T2) and not the identity question (T2→T3→T4). The consequence is unavoidable and we state it plainly: the paper's central quantity cannot at present be measured on real data by us or by anyone else, because the data that would measure it does not exist in the public domain. This is why a controlled synthetic instrument is not a convenience but a necessity—it is the only setting in which the identity axis can be varied at all.

That said, the synthetic setting is the study's second limitation, and the honest response to it is not realism but generator independence. The result currently rests on a data-generating process of our own construction, and the degeneracy audit, the falsification world, and the obfuscation sweep are all defenses against the objection that we designed the process to produce the answer. The strongest remaining defense is to run the parts of the decomposition that public data can support on public data. Future work will apply the linkage tier (T1→T2) of the harness to AMLworld HI-Small, whose account structure suits an account-based CBDC, marking T3 and T4 not-evaluable and reporting the linkage result on an externally generated benchmark. What this buys is not realism—AMLworld is itself synthetic—but independence from our generator on the one axis a public dataset can exercise, which is the direct answer to the objection that sank the earlier version of this work.

The third limitation is statistical. The canonical operating point realizes 388 positive cases, which is enough to identify the tier increments and to render the equivalence verdicts stable across seeds, but it is not large, and the average-precision intervals remain wide enough that the study leans on the equivalence test rather than on differencing point estimates. A larger positive class would tighten the surface further; it would not change the invariant, which is already stable across the sweep, but it would narrow the one inconclusive cell.

Finally, the architecture the measurement licenses carries a limitation of its own that the measurement does not touch. The abandonment primitive—allowing a user to walk away from a flagged transaction without deanonymization—doubles as a free adversarial probing oracle: an actor can submit transactions to learn the detection boundary and abandon those that are flagged, refining evasive behavior at no cost. This is a genuine tension between the privacy guarantee and the detection guarantee, it is not resolved here, and Section 7 notes it at the point where the primitive is specified rather than deferring it to a caveat.

11 Conclusion

Proposals for retail central bank digital currency have largely accepted that anti-money-laundering obligations require identity-linked transaction monitoring, and have taken the design problem to be how much privacy can be preserved around that requirement. This paper asked what the requirement is worth, and measured it. Decomposing surveillance access into three increments—pseudonymous linkage, identity attributes, and watchlist consultation—and pricing each against a fixed detector on fixed data, we find that the marginal value of identity access is not a property of the data but of the detector: a model that can reconstruct behavior from structure recovers almost everything from linkage and gains nothing measurable from identity, while a model that cannot leans on identity to substitute for what it cannot infer. Detection capability and identity collection are, to a meaningful degree, substitutes.

The finding is deliberately hedged where the evidence requires it. It holds robustly across the range of the generating process's most consequential parameter, on the metric a monitoring regime actually operates on, and it reverses cleanly when the process is rebuilt so that identity genuinely dominates. It is confined to the increment above linkage, and linkage is itself a surveillance capability. It is established on synthetic data, for the structural reason that no public dataset carries the identity axis the question

requires. Within those bounds, the conclusion is that identity collection by default is not buying the detection value it is assumed to buy, and that whether to collect identity by default is therefore a policy choice conditioned on detection capability rather than a technical necessity.

The architecture and cryptographic stack that this measurement licenses, set out in Sections 7–8, are assembled from a mature privacy-preserving-compliance literature rather than invented here, and their feasibility rests on existing primitives. Their contribution is that the measurement now gives a reason to prefer them: not that surveillance is impossible to build, but that, for a capable detector, the identity-linked version of it adds little the pseudonymous version does not already achieve. The instrument that establishes this, and the harness's insistence on auditing itself for degeneracy, testing equivalence rather than inferring it, and proving it can return the opposite verdict, are offered as much as the contribution as the number they produce—because in this corner of the literature, the way a result is manufactured has too often been the thing that made it wrong.

Declarations

Conflict of Interest. The authors declare no competing interests.

Funding. This research received no external funding.

Data Availability. All code and results are available at `https://github.com/andrewmaksakov/CBDC`. The detection-ablation harness of Sections 4–5 (`detection/`) is seeded and self-contained: `run_all.py` regenerates the committed results, `sweep_obfuscation.py` regenerates the robustness surface of Section 5.4, and `requirements.txt` pins the environment under which the committed figures reproduce. The PET–AML stack simulation of Section 8 (`pet_aml_sim.py`) is in the same repository.

AI Assistance. Claude (Anthropic) was used as a research collaborator for analysis, literature synthesis, LaTeX preparation, and iterative refinement. All intellectual claims and errors remain the authors' responsibility.

References

Raphael Auer and Rainer Böhme. The technology of retail central bank digital currency. *BIS Quarterly Review*, pages 85–100, March 2020. URL https://www.bis.org/publ/qtrpdf/r_qt2003j.htm. Foundational CBDC taxonomy: direct, indirect, and hybrid architectures.

Raphael Auer, Jon Frost, Leonardo Gambacorta, Cyril Monnet, Tara Rice, and Hyun Song Shin. Central bank digital currencies: Motives, economic implications, and the research frontier. *Annual Review of Economics*, 14:697–721, 2022. doi: 10.1146/annurev-economics-051420-020324. Comprehensive BIS review establishing the CBDC research frontier.

Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE Symposium on Security and Privacy*, pages 459–474. IEEE, 2014.

BIS Innovation Hub. Project aurora: The power of data, technology and collaboration to combat money laundering across institutions and borders. Technical report, Bank for International Settlements, Basel, 2023. URL https://www.bis.org/publ/othp66.pdf. Proof of concept: PETs, machine learning, and network analysis on synthetic payments data; collaborative analysis outperforms siloed monitoring.

BIS Innovation Hub and Bank of England. Project hertha: Identifying financial crime patterns in real-time retail payment systems. Technical report, Bank for International Settlements, Basel, 2025. URL https://www.bis.org/publ/othp96.pdf. Network analytics on a synthetic dataset of 1.8M accounts and 308M transactions; 12% more illicit accounts identified, 26% on novel patterns.

Board of Governors of the Federal Reserve System. Money and payments: The U.S. dollar in the age of digital transformation. Technical report, Federal Reserve, Washington, D.C., 2022.

Vitalik Buterin, Jacob Illum, Matthias Nadler, Fabian Schär, and Ameen Soleimani. Blockchain privacy and regulatory compliance: Towards a practical equilibrium. *Blockchain: Research and Applications*, 5(1):100176, 2024. doi: 10.1016/j.bcra.2023.100176. Privacy Pools: voluntary zero-knowledge proofs of dissociation from illicit funds.

Jan Camenisch, Ueli Maurer, and Markus Stadler. Digital payment systems with passive anonymity-revoking trustees. In *Computer Security — ESORICS 96*, Lecture Notes in Computer Science, pages 33–43. Springer, 1996. doi: 10.1007/3-540-61770-1_26. Trustee-based e-cash: anonymity revoked only on justified suspicion.

Jan Camenisch, Susan Hohenberger, and Anna Lysyanskaya. Compact e-cash. In *Advances in Cryptology — EUROCRYPT 2005*, Lecture Notes in Computer Science, pages 302–321. Springer, 2005. doi: 10.1007/11426639_18. Anonymous e-cash with violation-triggered traceability and no trusted third party.

Cato Institute. Cato institute 2023 central bank digital currency national survey. Survey toplines; fielded by YouGov, 27 February–8 March 2023, $n = 2{,}000$ US adults, 2023. https://www.cato.org/sites/cato.org/files/2023-05/cato-cbdc-survey-toplines.pdf.

David Chaum. Blind signatures for untraceable payments. In *Advances in Cryptology: Proceedings of Crypto 82*, pages 199–203, Boston, MA, 1982. Springer.

David Chaum, Christian Grothoff, and Thomas Moser. How to issue a central bank digital currency. SNB Working Papers 2021-03, Swiss National Bank, 2021. URL https://www.snb.ch/en/publications/research/working-papers/2021/working_paper_2021_03. Privacy-preserving CBDC design using blind signatures, co-authored with SNB board member.

Alex Davidson, Armando Faz-Hernández, Nick Sullivan, and Christopher A. Wood. Oblivious Pseudorandom Functions (OPRFs) Using Prime-Order Groups. RFC 9497, September 2023. URL https://www.rfc-editor.org/info/rfc9497.

Nicolin Decker. Proof without exposure: Zero-knowledge proofs as a cryptographic framework for institutional financial compliance. SSRN Working Paper. ZKP-based KYC reducing exposed user data by 97%, AI-enhanced fraud detection achieving 96.7% accuracy, 2025.

Mohammed El-Hajj and Bjorn Oude Roelink. Evaluating the efficiency of zk-SNARK, zk-STARK, and bulletproof in real-world scenarios: A benchmark study. *Information*, 15(8):463, 2024. doi: 10.3390/info15080463. URL https://www.mdpi.com/2078-2489/15/8/463.

Electric Coin Company. Zcash metrics. https://electriccoin.co/blog/, 2023. Shielded transaction volume statistics.

Daniel Escudero et al. Perfectly-secure multiparty computation with linear communication complexity over any modulus, 2024. URL https://eprint.iacr.org/2024/370.

European Central Bank. Eurosystem report on the public consultation on a digital euro. European Central Bank, 2021. URL https://www.ecb.europa.eu/euro/digital_euro/report/html/index.en.html. 8,221 responses; privacy ranked the most important feature of a digital euro (43% of respondents), ahead of security (18%).

European Central Bank. Making the digital euro truly private. ECB Blog, 2024. URL https://www.ecb.europa.eu/press/blog/date/2024/html/ecb.blog240613~47c255bdd4.en.html.

Murad Farzulla. Consent-theoretic framework for quantifying legitimacy: Stakes, voice, and friction in adversarial governance. *Zenodo Preprint*, 2025. doi: 10.5281/zenodo.17626762. Operationalization of consent-based legitimacy framework. Also SSRN:5918222.

Murad Farzulla. Legitimate extraction: Sophisticated laundering hides in plain sight. *SSRN Electronic Journal*, 2026. doi: 10.2139/ssrn.6145046. Fourth-stage AML framework. Targeting Oxford J. Financial Regulation. Zenodo: 10.5281/zenodo.17626621.

Financial Action Task Force. International standards on combating money laundering and the financing of terrorism & proliferation: The FATF recommendations. Technical report, FATF/OECD, Paris, 2012–2023. Updated periodically.

Christina Garman, Matthew Green, and Ian Miers. Accountable privacy for decentralized anonymous payments. In *Financial Cryptography and Data Security — FC 2016*, Lecture Notes in Computer Science, pages 81–98. Springer, 2016. doi: 10.1007/978-3-662-54970-4_5. Zero-knowledge policy enforcement (spending limits, selective tracing, tainted-coin tracing) for decentralized anonymous payments.

Rosario Gennaro and Steven Goldfeder. Fast multiparty threshold ECDSA with fast trustless setup. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1179–1194, 2018.

Google. Private join and compute. Google AI Blog, 2019. https://ai.googleblog.com/2019/06/private-join-and-compute.html.

International Monetary Fund. Central bank digital currency data use and privacy protection. *FinTech Notes*, 2024(004), 2024. URL https://www.elibrary.imf.org/view/journals/063/2024/004/article-A001-en.xml.

Todd Keister and Daniel Sanches. Should central banks issue digital currency? *Review of Economic Studies*, 90(1):404–431, 2023. doi: 10.1093/restud/rdac017. General equilibrium analysis of CBDC impact on banking.

Aggelos Kiayias, Markulf Kohlweiss, and Amirreza Sarencheh. PEReDi: Privacy-enhanced, regulated and distributed central bank digital currencies. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*, pages 1739–1752. ACM, 2022. doi: 10.1145/3548606.3560707. Regulated privacy-preserving CBDC with threshold-controlled tracing and auditing under UC security proofs.

Nishat Koti, Varsha Bhat Kukkala, Arpita Patra, and Bhavish Raj Gopal. Graphiti: Secure graph computation made more scalable. IACR Cryptology ePrint Archive, Report 2024/1756, 2024. URL https://eprint.iacr.org/2024/1756. Also published at ACM CCS 2024.

James Lovejoy, Madars Virza, Cory Fields, Kevin Karwaski, Anders Brownworth, and Neha Narula. Hamilton: A high-performance transaction processor for central bank digital currencies. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, 2023. URL https://www.usenix.org/conference/nsdi23/presentation/lovejoy. Project Hamilton (Fed Boston + MIT). 1.7M TPS. Also IACR ePrint 2022/586.

Sarah N. Lynch. Danske bank pleads guilty in $2 billion money laundering scheme. Reuters, 2022. https://www.reuters.com/business/finance/danske-bank-plead-guilty-fraud-charges-doj-says-2022-12-13/.

Panagiotis Michalopoulos, Anthony Mack, Cameron Clark, Linus Chen, Johannes Sedlmeir, and Andreas Veneris. Balancing compliance and privacy in offline CBDC transactions using a secure element-based system, 2025. URL https://arxiv.org/abs/2509.25469.

Andrea Minto. Certainties and uncertainties surrounding central bank digital currencies (CBDC) vis-a-vis the EU anti-money laundering regulatory framework. *European Company and Financial Law Review*, 21(5-6):671–704, 2024. doi: 10.1515/ecfr-2024-0020. EU AML regulatory treatment of offline vs. online digital euro.

Ronald F. Pol. Anti-money laundering: The world's least effective policy experiment? together, we can fix it. *Policy Design and Practice*, 3(1):73–94, 2020. doi: 10.1080/25741292.2020.1725366.

Ellie Rennie and Stacey Steele. Privacy and emergency payments in a pandemic: How to think about privacy and a central bank digital currency. *Law, Technology and Humans*, 3(1), 2021. Identifies four privacy losses from CBDC: anonymity, liberty, individual control, regulatory control.

Alberto Sonnino, Mustafa Al-Bassam, Shehar Bano, Sarah Meiklejohn, and George Danezis. Coconut: Threshold issuance selective disclosure credentials with applications to distributed ledgers. In *Network and Distributed System Security Symposium (NDSS)*, 2019. URL https://www.ndss-symposium.org/ndss-paper/coconut-threshold-issuance-selective-disclosure-credentials-with-applications-to-distributed-ledgers/.

United Nations Office on Drugs and Crime. Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes. Technical report, UNODC, Vienna, 2011.

Marie Beth van Egmond, Vincent Dunning, Stefan van den Berg, Thomas Rooijakkers, Alex Sangers, Ton Poppe, and Jan Veldsink. Privacy-preserving anti-money laundering using secure multi-party computation. IACR Cryptology ePrint Archive, Report 2024/065, 2024. URL https://eprint.iacr.org/2024/065. Published at Financial Cryptography 2024. Deployed MPC-based AML with ABN AMRO, Rabobank, De Volksbank.

Xiaodong Wang, Zijie Lu, Bei Liang, and Shengzhe Meng. Unbalanced PSI from client-independent relaxed oblivious PRF. *Proceedings on Privacy Enhancing Technologies*, 2025(3):475–493, 2025. doi: 10.56553/popets-2025-0109. URL https://crysp.petsymposium.org/popets/2025/popets-2025-0109.pdf.

Karl Wüst, Kari Kostiainen, Noah Delius, and Srdjan Capkun. Platypus: A central bank digital currency with unlinkable transactions and privacy-preserving regulation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*, pages 2947–2960. ACM, 2022. doi: 10.1145/3548606.3560617. Privacy-default CBDC with regulation (holding/receiving limits) enforced via zero-knowledge proofs.

Jiping Yu, Kun Chen, Yunyi Chen, Xiaoyu Fan, Xiaowei Zhu, Cheng Hong, and Wenguang Chen. GraphAce: Secure two-party graph analysis achieving communication efficiency. In *34th USENIX Security Symposium (USENIX Security 25)*, August 2025. URL https://www.usenix.org/system/files/usenixsecurity25-yu-jiping.pdf.

Filippo Zatti. From legal tender to prohibition: Competing paradigms in central bank digital currency law. *Strani pravni život*, 2025. Analyses the U.S. Anti-CBDC Surveillance State Act as paradigm shift from enabling to prohibiting CBDC.