Erasure-Compliant, Differentially Private Distinct Counting under Continual Observation: A Flippancy-Aware System with Edge-to-Cloud Deployment and a CCTV Case Study

Dev Sanghvi
Rice University

Draft: PAPER_DATE

Abstract

Distinct-count analytics (e.g., DAU/MAU-style metrics) are central to product analytics, IoT telemetry, ads reach, finance risk signals, and more. These settings increasingly require *both* meaningful privacy guarantees and the ability to honor erasure requests (GDPR/CCPA). We present an erasure-compliant, differentially private (DP) pipeline for *distinct* counting in *turnstile* streams (insertions and deletions) under *continual release*. The system combines: (i) **pseudonymization** with rotating salts; (ii) **Theta sketches** with union and *A-not-B* for efficient per-period and rolling-window distinct counting with deletions; and (iii) a **flippancy-aware** DP release mechanism calibrated for item-level privacy over time. We implement a working prototype with REST APIs, a privacy accountant, and synthetic workloads. Results show that at $\epsilon = EPSILON_DAU$ (daily) and $\epsilon = EPSILON_MAU$ (monthly), additive error is within ERROR_PCT% and ingestion sustains THROUGHPUT events/s on a commodity machine (placeholders). We discuss budget governance, deletion replay, and sketch choices (Theta vs. HLL++ rebuilds). To illustrate deployability, we include a brief *CCTV* case study as one application among many, without narrowing the generality of the approach.

1 Introduction

Distinct-count metrics such as daily or rolling-window "active entities" appear across domains: web/mobile user analytics, IoT fleet monitoring, ads reach, financial risk telemetry, and physical-space analytics. These analytics are increasingly subject to privacy expectations and erasure laws (e.g., GDPR Art. 17) [1]. Releasing counts continually (e.g., daily) raises composition concerns, and supporting deletions turns the problem into a *turnstile* stream.

Differential Privacy (DP) [2] bounds any single subject's influence by adding calibrated noise. Recent theoretical progress addresses *distinct counting with deletions* under continual observation via the *flippancy* parameter w—the maximum number of presence/absence flips per subject—achieving error $\tilde{O}(\sqrt{w})$ at item-level DP and proving tight lower bounds for event-level DP [3].

We contribute a practical, end-to-end pipeline for erasure-compliant, DP *distinct* analytics that is domain-agnostic:

- **Pseudonymization with rotating salts.** Transform short-lived identifiers (e.g., app/device IDs, IoT sensor tokens) into *period-scoped salted hashes*; retain no raw identities.
- **Sketch-based distinct counting.** Use *Theta* sketches for per-period distincts and rolling-window unions; support deletions via *A-not-B*. Contrast with HLL++ (merge-only; no native delete) [5, 7, 8].

- **Flippancy-aware DP releases.** Daily distincts and W-day MAW (monthly active whatever) with Laplace/Gaussian noise scaled for item-level DP; a privacy accountant tracks composition; deterministic per-period noise prevents query averaging [2, 4].

We implement a prototype with a REST service, a budget ledger, and synthetic workloads. Placeholders indicate where figures slot in. The design generalizes across domains; we include a short *CCTV case study* to demonstrate deployment in a high-sensitivity setting while keeping the paper's contribution general.

Contributions. (1) A deployment-ready *edge→cloud* recipe for erasure-aware DP distinct counting across domains; (2) a sketch layer with *A-not-B* to operationalize deletions; (3) a continual-release DP mechanism and budget service tuned to flippancy; (4) an evaluation plan and operational guidance for compliance-ready rollouts; (5) a brief case study instantiation in CCTV.

2 Background and Related Work

Differential Privacy and composition. DP bounds output distribution changes from any single subject [2]. Repeated releases compose; basic composition sums ϵ, while advanced analyses (e.g., Rényi DP) tighten bounds. Deterministic per-query/per-period noise seeds prevent adversarial re-query averaging [4]. NIST SP 800-226 offers guidance for documenting DP claims [12].

Turnstile distinct counting with deletions. Under continual observation with deletions, event-level DP faces strong lower bounds; *flippancy* w enables item-level DP with additive error $\tilde{O}(\sqrt{w})$ [3]. Many telemetry workloads naturally bound w via coalescing (e.g., one presence per subject per period).

Distinct sketches. HyperLogLog (HLL/HLL++) estimates distincts with small memory but is *append-only* (no deletions) [5]. Theta (KMV) sketches support *union*, *intersection*, and *A-not-B* (set difference) [6–8]. For large sets, order-invariant cardinality estimators can exhibit inherent DP-like behavior, though we still add explicit DP noise [9].

Adjacent privacy systems. Deployed DP analytics systems (e.g., LinkedIn's Audience Engagements) pair user-level DP with budget governance and deterministic noise [4]. Video-focused DP systems (VideoDP, Privid) address other query classes; our work targets *distinct under deletions* with continual releases [10, 11].

3 Problem Formulation

We consider many sources (apps/devices/sensors/zones) generating subject activity. For period t, define S_t as the set of (pseudonymous) subjects active that period. DAU-like metric: $\mathrm{DA}(t) = |S_t|$. For a window W (e.g., 30 days), MAW-like metric:

$$\mathrm{MA}(t) = \Big|\bigcup_{i=t-W+1}^{t} S_i\Big|.$$

Privacy. Item-level DP protects a subject's *entire* trace. We release DA/MA once per period with (ϵ_t, δ_t), tracking composition.

Turnstile and deletions. Insertions add a subject to S_t; erasures request removing a subject from all S_i where present. We maintain updated sets/sketches and ensure future releases reflect removals.

Flippancy. Let w bound per-subject flips across the horizon; period coalescing keeps w small in many telemetry workloads.

4 System Overview (Edge → Cloud)

Edge/Device

Pseudonymization. For a short-lived signal (e.g., app/device token) x, compute

$$\text{subject\_key} = H(x \,\|\, \text{salt}_{\text{period}})$$

with a rotated salt; drop x. Maintain a per-period Theta sketch T_t (or exact set for small loads). Upload sketch bytes (or DP'd scalar).

Gateway/Site

Union per-source period sketches into site-level T_t^{site}. Maintain a rolling union $U_t = \text{Union}(T_{t-W+1}, \ldots, T_t)$. For erasure of subject u across days $\mathcal{D}$, construct $U_d^{(u)}$ (tiny sketches containing u for each $d \in \mathcal{D}$) and apply $T_d \leftarrow \text{AminusB}(T_d, U_d^{(u)})$; update U_t accordingly.

Cloud/Tenant

Aggregate site sketches (unions), apply DP release (Section 5), and log to a privacy budget ledger (metric, period, ϵ, δ, seed).

5 Algorithms

Sketch layer. Backends:

- *Theta (preferred):* union/intersection/A-not-B; size k controls RSE. Efficient for rolling windows and deletions [6, 8].
- *Exact sets (baseline):* for small loads/testing.
- *HLL++ (optional):* union-only; deletions require period-level rebuild from a light index [5].

DP continual release. For count f_t (DA or MA on period t), with item-level sensitivity $\Delta = 1$, release

$$\tilde{f}_t = f_t + \eta_t, \qquad \eta_t \sim \text{Lap}\left(\frac{\Delta}{\epsilon_t}\right).$$

Use deterministic, per-(metric,period) seeding to prevent averaging via re-queries [4]. A budget service records $(t, \epsilon_t, \delta_t)$ and enforces caps. For tighter composition, swap to Gaussian/RDP.

Deletion replay. On erasure of subject u: (1) identify periods $\mathcal{D}$; (2) remove u from T_d via A-not-B (Theta) or rebuild (HLL++); (3) update window unions; (4) future releases use updated values with fresh noise. Historical releases are not retroactively DP-redactable; treat as superseded (operational policy).

6 Implementation

Python 3.11 prototype (FastAPI service, SQLite ledger). Modules:

- `sketches/`: Theta backend (via DataSketches) and exact sets.
- `pipeline.py`: period stores, rolling unions, deletions, DP releases.
- `privacy_accountant.py`: budget ledger; basic composition; warnings on cap.
- `routes.py`: `POST /event`, `GET /dau/{day}`, `GET /mau/{day}`.
- `auth.py`: optional API key.

Noise seeding. PRNG seeded by secret s and key (metric, period); rotate s periodically. **Windows.** Maintain last $W = MAU_WINDOW_DAYS$ periods (drop older). **Edge.** On constrained devices, emit per-hour sets to gateway which compacts to per-period sketches.

7 Evaluation (placeholders)

We outline the study; concrete numbers/plots are placeholders to be filled.

Setup. Synthetic telemetry over $EVAL_DAYS$ periods; $N = N_USERS$ subjects; per-period active DAILY_ACTIVE; overlap REPEAT_RATE%. Deletions: start period DELETE_-START, remove DELETE_COUNT subjects spread across prior periods.

Metrics. (1) **Accuracy:** MAE/relative error for DA/MA under $\epsilon = EPSILON_DAU, EPSILON_MAU$; (2) **Sketch impact:** Theta vs exact; (3) **Deletions:** correctness of replay and effect on MA; (4) **Latency/Throughput:** p50/p99 and ingestion rate.

Results (to be inserted).

- **Noise accuracy:** mean DA MAE = DAU_MAE (% error = DAU_REL_ERR%); MA rel. error = MAU_REL_ERR%. Figure FIG_NOISE_ACCURACY.
- **Sketch vs exact:** RSE ≈ THETA_RSE% at $k = THETA_K$; scatter near diagonal (Fig. FIG_SKETCH_VS_EXACT).
- **Deletions:** post-erasure MA drops consistent with removed subjects ± DP noise (Table TAB_DELETE).
- **Performance:** ingestion THROUGHPUT events/s; queries $< QUERY_LAT_MS$ ms p99 (Fig. FIG_PERF).

8 Case Study: CCTV Analytics (Brief)

As one concrete instantiation, we apply the pipeline to occupancy/footfall analytics in CCTV:

- **Edge pseudonymization:** daily-salted hashes of track/plate tokens on camera/NVR; no raw identities retained.
- **Sketches at gateway:** per-day Theta sketches per camera/zone; site-level unions; rolling 30-day MAU via union; deletions via *A-not-B*.
- **DP releases and ledger:** daily DAU and 30-day MAU with published (ϵ, δ); budget accounting and deterministic per-day noise.

Operational notes: Theta eases deletions; HLL++ requires day-rebuilds. Salt rotation breaks long-term linkability; retention windows align with data minimization. (Evaluation uses the same metrics as Section 7, with CCTV-like workloads; placeholders apply.)

9 Discussion

Theta vs HLL++. Theta's A-not-B simplifies deletions and rolling windows; HLL++ suits union-only or rebuild-on-delete pipelines [5, 8]. **Budgeting.** Expose customer-visible ϵ tiers; allocate budget across DA vs MA. **Deployment.** Edge pseudonymization generalizes to SDKs/mobile, IoT gateways, and NVRs.

10 Limitations and Threats to Validity

Assumptions on identifier stability and coalescing bound flippancy; synthetic traces may miss domain idiosyncrasies; sketch adversarial inputs and hash collisions; noise-seed handling and multi-instance coordination.

11 Ethics and Compliance

No raw identities beyond transient signals; period salt rotation; erasure SLAs; DP documentation and audit via [12]; DPIA/IRB if evaluating on human-subject traces.

12 Conclusion

We presented a domain-agnostic, erasure-compliant DP pipeline for distinct analytics under continual release. By combining pseudonyms, Theta sketches, and flippancy-aware DP releases with a budget service, we deliver useful DA/MA metrics while honoring deletions. Future work: tree-aggregation for smoother continual releases, RDP accounting, federated deployments, and large-scale evaluations; extended case studies (ads reach, IoT fleets).

References

[1] European Parliament and Council. *General Data Protection Regulation (EU) 2016/679 (GDPR)*, 2016.

[2] C. Dwork, F. McSherry, K. Nissim, A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[3] (Authors). Counting Distinct Elements in the Turnstile Model with Differential Privacy under Continual Observation. *NeurIPS*, 2023. (arXiv:2306.06723).

[4] R. Rogers et al. LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale. arXiv:2002.05839, 2020.

[5] P. Flajolet, É. Fusy, O. Gandouet, F. Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. *AOFA*, 2007.

[6] A. Dasgupta, K. J. Lang, L. Rhodes, J. Thaler. A Framework for Estimating Stream Expression Cardinalities. *ICDT*, 2016.

[7] Apache DataSketches Project. Theta Sketches Documentation. `https://datasketches.apache.org/`, accessed ACCESS_DATE.

[8] Apache DataSketches. A-not-B and Set Operations for Theta Sketches. `https://datasketches.apache.org/`, accessed ACCESS_DATE.

[9] C. Dickens, J. Thaler, D. Ting. Order-Invariant Cardinality Estimators Are Differentially Private. *SODA*, 2023.

[10] (Authors). VideoDP: A Platform for Differentially Private Video Analytics. *PoPETs*, 2020.

[11] (Authors). Privid: Practical, Privacy-Preserving Video Analytics Queries. *NSDI*, 2022.

[12] NIST Special Publication 800-226. *A Taxonomy and Terminology of Differential Privacy Mechanisms and Applications*. 2025.

Appendix A: Placeholders and How to Complete the Paper

Replace every token below (search for the exact token text):

- **PAPER_DATE**: Draft date (e.g., "October 2025").
- **ACCESS_DATE**: Date you accessed online documentation pages.
- **MAU_WINDOW_DAYS**: Window length (default 30).
- **EPSILON_DAU**, **EPSILON_MAU**, **DELTA**: Privacy parameters used in evaluation.
- **ERROR_PCT**: Overall relative error summary (e.g., "2").
- **THROUGHPUT**: Ingestion throughput (events/s).
- **QUERY_LAT_MS**: Query latency (ms).
- **EVAL_DAYS**, **N_USERS**, **DAILY_ACTIVE**, **REPEAT_RATE**: Synthetic trace parameters.
- **DELETE_START**, **DELETE_COUNT**: Deletion scenario parameters.
- **DAU_MAE**, **DAU_REL_ERR**, **MAU_REL_ERR**: Accuracy metrics.
- **THETA_K**, **THETA_RSE**: Theta sketch configuration and nominal RSE.
- **FIG_NOISE_ACCURACY**, **FIG_SKETCH_VS_EXACT**, **FIG_PERF**, **TAB_DELETE**: Figure/table labels; add environments or remove references.
- **ARCH_DIAGRAM_PATH**: If adding an architecture diagram figure (optional).

To finalize for submission: (1) Fill placeholders and insert figures/tables; (2) ensure all citations have full bibliographic details (replace "(Authors)" with actual author lists if needed); (3) compile and check for warnings; (4) remove this appendix section if the venue discourages placeholder notes; (5) adopt the venue's LaTeX class if required (ACM/IEEE).