ZERO-COST DENOISING FOR OPTIMIZING DIFFERENTIALLY PRIVATE SYNTHETIC DATA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Private Evolution (PE) is an API-only framework for differentially private (DP) synthetic text (Lin et al., 2024). Each PE iteration spends privacy budget on a noisy nearest-neighbor vote histogram used to select synthetic candidates. We show that this histogram is under-utilized and can be denoised without additional privacy loss under the PE framework: the denoiser is DP post-processing of the noisy histogram, the synthetic candidate graph, and the public vote sum constraints. Our estimator, GRE, solves one convex program that combines the sum-to-N simplex constraint with Laplacian smoothness on the candidate kNN graph. These are complementary where the simplex constraint preserves the known total but ignores geometry, while the graph penalty uses geometry but does not enforce the histogram constraints. Empirically, histogram error minimization, one-shot ranking accuracy maximization, and closed-loop PE generation quality are not strictly connected. We therefore tune by a degrees-of-freedom rule calibrated to closed-loop synthetic data generation behavior. Our experiments on domain-mismatched text corpora including AG News and DBpedia, graph-regularized histogram-constrained denoising improve top-N PE quality at tight privacy budgets.

1 INTRODUCTION

Differentially private (DP) synthetic text lets data holders share the statistical value of sensitive dataset without exposing individuals (Dwork & Roth, 2014; Hu et al., 2024). A recent trend is to use DP as a privacy-preserving mechanism to generate synthetic versions of sensitive data to unblock enterprise or research use-cases. Private Evolution (PE) (Lin et al., 2024) and its text instantiation AugPE (Xie et al., 2024) have emerged as the leading solution to this new privacy-preserving synthetic data generation paradigm. PE requires only black-box inference to an LLM: it draws an initial population of synthetic samples, and then iteratively (i) scores each synthetic candidate by how many private examples have it as their nearest neighbor in an embedding space, (ii) releases this vote histogram under a DP mechanism (e.g. Gaussian mechanism), and (iii) keeps or resamples the high-scoring candidates and produce variants of them into the next generation.

The single point where private data is used in the algorithm is the vote histogram, where its L_2 sensitivity is $\sqrt{2}$ (replacing a private record moves its vote between two bins; the size N being fixed and public), so T iterations compose, via zero-concentrated DP (Bun & Steinke, 2016), to a total budget of T DP releases. This makes the histogram the only statistic where privacy is consumed. However, the typical PE algorithms consumes this statistic nearly raw: it clips the noisy counts at zero, then ranks (in AugPE) or resamples (in base PE). The current paradigm of PE ignores several properties from the released DP statistic including both the known vote total and the embedding geometry of the candidates if represented in a graph, which together can determine how the clean histogram can plausibly look to improve synthetic data generation quality. The recent MAPLE framework (Chien et al., 2026) improves PE during initialization, grounding the first generation in DP-extracted tabular metadata and in-context examples; it however does not address how to consume each histogram effectively during PE. Our work is complementary to existing PE algorithms: it improves synthetic data selection by solving a constrained graph smoothing program with structures available in the released DP histograms and other public attributes without consuming additional privacy budget per post-processing property of DP.

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

Independently, the PE-means line of work (Humphries et al., 2026) observes that for PE the noisy histogram can be projected onto the simplex $\{x \geq 0,\ \mathbf{1}^\top x = N\}$, the maximum-likelihood estimate under the public constraints (namely sum-to-N simplex and candidate similarity graph smoothness) , at no privacy cost, and reports gains in downstream clustering loss. However, this projection is permutation-invariant: it treats the histogram as an unordered bin collection of counts and ignores the embedding geometry entirely. Conversely, classical graph signal denoising exploits geometry but ignores the histogram constraint.

Contributions. In our work, we show that combining the two, and choosing the smoothing strength for selection rather than for estimation, yields a better selection step for PE without additional private queries:

1. **A post-processing Graph-Regularized Estimator (GRE).** We formulate histogram denoising as a single convex program (§3) that couples the exact sum-to-N simplex projection with Laplacian smoothness over the synthetic kNN graph. Under the public-N threat model convention, this selection step is post-processing of the released histogram and synthetic population, so it costs no additional privacy loss per post-processing of DP (Proposition 1).
2. **3-way misaligned objectives among estimation, selection, and closed-loop generation.** The best smoothing strength λ depends on the objective empirically where the three objectives misalign. We notice that minimizing estimation error (MSE) calls for much less smoothing than maximizing one-shot selection accuracy: the MSE-optimal λ for the constrained estimator is smaller than the ranking-optimal one. The closed loop generation, in turn, requires less smoothing than one-shot ranking: smoothing enough to sharpen a single selection round will over-concentrate the population and restricts exploration across PE generations. We give an effective-degrees-of-freedom rule (§4.4) that targets the closed-loop optimum, with no tuned constants.
3. **Empirical gains at no extra privacy cost.** On a microbenchmark (§4.2) GRE improves MSE by up to $5\times$ at $\varepsilon = 0.41$ for example compared to AugPE, while improving top-k selection agreement at every budget level chosen. End-to-end on two domain-mismatched corpora (AG News and DBpedia), our method improves the privacy-utility tradeoff in the top-N configuration for one but fails in the other one (as predicated in Proposition 3).

2 Preliminaries

Differential privacy. We work in the *bounded* model: datasets D, D' are *neighbors* ($D \simeq D'$) if one is obtained from the other by replacing a single record, so the size $|D| = |D'| = N$ is fixed and public.

Definition 1 ((ε, δ)-DP (Dwork et al., 2006))**.** A randomized mechanism $\mathcal{A}$ is (ε, δ)-DP if for all neighbors $D \simeq D'$ and all measurable $S \subseteq \mathrm{range}(\mathcal{A})$,

$$\Pr[\mathcal{A}(D) \in S] \ \leq\ e^{\varepsilon}\ \Pr[\mathcal{A}(D') \in S] + \delta.$$

Definition 2 (Zero-concentrated DP (Bun & Steinke, 2016))**.** $\mathcal{A}$ is ρ-zCDP if for all neighbors $D \simeq D'$ and all orders $\alpha \in (1, \infty)$, the Rényi divergence satisfies $D_\alpha\big(\mathcal{A}(D) \,\|\, \mathcal{A}(D')\big) \leq \rho\,\alpha$.

We use three standard facts (Dwork & Roth, 2014; Bun & Steinke, 2016). (i) The Gaussian mechanism releasing a query of L_2 sensitivity Δ with noise $\mathcal{N}(0, \sigma^2 I)$ is ρ-zCDP with $\rho = \Delta^2/(2\sigma^2)$. (ii) *Post-processing*: any data-independent map of a ρ-zCDP output is ρ-zCDP, and *composition* is additive (k releases give $\sum_i \rho_i$). (iii) ρ-zCDP implies $\big(\rho + 2\sqrt{\rho \log(1/\delta)},\ \delta\big)$-DP for every $\delta > 0$. We report privacy in ρ and use (iii) only for cross-reference.

Private Evolution and the vote histogram. Let $X = \{x_i\}_{i=1}^N$ be the private embeddings and $S_t = \{s_j\}_{j=1}^M$ the synthetic population at iteration t. The vote histogram $h \in \mathbb{R}^M$ counts, for each synthetic point, how many private points have it as their nearest neighbor:

$$h_j \ =\ \big|\{\, i : j = \arg\min_{j'}\ d(x_i, s_{j'})\,\}\big|, \qquad \textstyle\sum_j h_j = N. \tag{1}$$

PE releases $\tilde{h} = h + \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 1/\sqrt{\rho_t}$ for a per-iteration budget ρ_t, then selects: AugPE (Xie et al., 2024) and MAPLE (Chien et al., 2026)'s keep the M_{syn} highest-scoring text

candidates (top-N selection, after the target synthetic-set size). Selected candidates are paraphrased by a variation API into the next population. Over T iterations the total budget is $\rho = \sum_t \rho_t$; AugPE splits it uniformly, $\rho_t = \rho/T$.

Top-N is a lossy decoder of the released histogram: after sorting it keeps only which bins lie above the cutoff and discards the margins, so released noisy counts $\{101, 100, 99, 98\}$ and $\{500, 10, 9, 8\}$ are treated as equally decisive even though the second is far stronger evidence. In this work, we improve this selection step by recovering the noisy voting histograms using post-processing property of DP with constraints inherent with histogram as well as the smoothness of embedding graph.

3 Method: Graph-Regularized Constrained MLE

3.1 Observations

Proposition 1 (Post-processing of DP histograms)**.** *Given any PE iteration t, suppose the mechanism releasing $\tilde{h}_t$ is ρ_t-zCDP, and that the candidate set, its embeddings , the graph $\mathbf{L}_t$, and the vote total N are public. Then, post-processing estimations that depends only on $(\tilde{h}_t, \mathbf{L}_t, N)$ incur no additional privacy loss beyond the release of $\tilde{h}$.*

The clean histogram h has two exploitable structures available without consuming additional privacy budget (see Proposition 1): 1) Count constraint. $h \in \Delta_N := \{x \in \mathbb{R}^M : x \geq 0,\ \mathbf{1}^\top x = N\}$ where, following typical PE assumptions, we treat N as public; 2) Graph smoothness. Nearby synthetic sample points attract nearest-neighbor votes from overlapping private neighborhoods, so h should vary smoothly over the kNN graph (i.e., the neighboring points receives close/"smooth" votes). Let W be the kNN adjacency of the synthetic embeddings and $\mathbf{L} = I - D^{-1/2} W D^{-1/2}$ its normalized Laplacian, smoothness is measured by $h^\top \mathbf{L} h$ (Dong et al., 2016), which will be small given the above observation. In other words, N is public by threat model convention and the graph is post-processing of the synthetic sample population.

3.2 Denoised Histogram Estimator

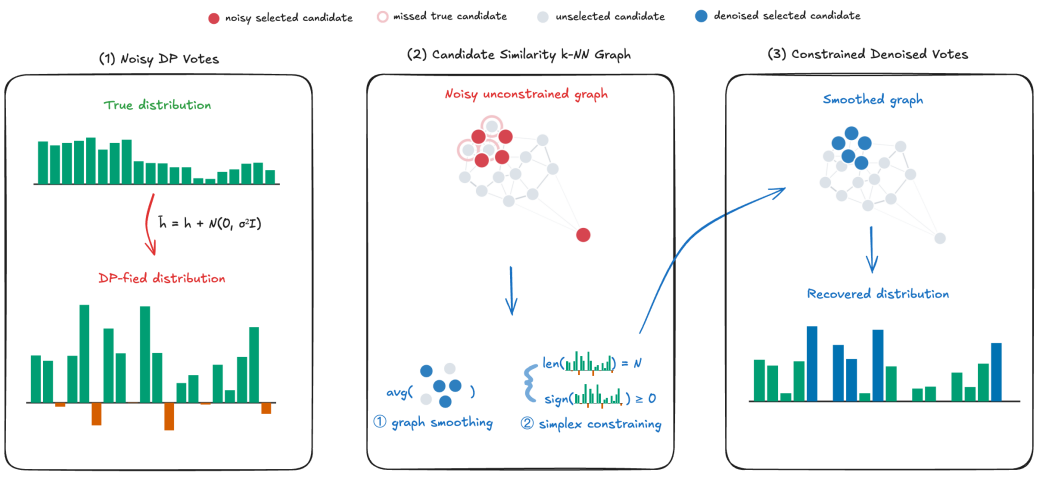

Figure 1: GRE selection pipeline. (1) The differentially private nearest-neighbor vote histogram h is released under the Gaussian mechanism as $\tilde{h} = h + \mathcal{N}(0, \sigma^2 I)$ (where some counts can be negative due to DP noises); (2) On the candidate k-NN similarity graph, GRE smooths counts over its graph neighbors and apply constraints; (3) The denoised estimate $\hat{h}$ recovers the true high-vote candidates (blue) as a non-negative histogram summing to N. All results are post-processing of the same DP release.

Algorithm 1 Private Evolution with GRE (the shaded steps are from GRE)

Input: private set X ($|X|{=}N$), embedder ϕ, generator $\mathcal{G}$, iterations T, budgets $\{\rho_t\}$, selection kept size N_{syn}, variations L
$S_0 \leftarrow$ initial samples from $\mathcal{G}$ // no private data
for $t = 1$ **to** T **do**
  $C_t \leftarrow L$ variations of each sample in S_{t-1} via $\mathcal{G}$ // candidate pool, $|C_t|{=}M$
  $h_j \leftarrow \left|\{x \in X : j = \arg\min_{c \in C_t} \|\phi(x) - \phi(c)\|\}\right|$ // NN votes, $\mathbf{1}^\top h{=}N$
  $\tilde{h}_t \leftarrow h + \mathcal{N}(0, \sigma_t^2 I)$ // Gaussian mechanism (only DP step)
  $\mathbf{L}_t \leftarrow$ normalized Laplacian of the kNN graph of $\phi(C_t)$ // public geometry
  $\hat{r} \leftarrow |\text{supp}(\Pi_{\Delta_N} \tilde{h}_t)|$; $\lambda^\star \leftarrow \arg\min_\lambda |\text{df}(\lambda) - \hat{r}|$ // choose λ, Eq. (5)
  $\hat{h} \leftarrow$ FISTA on Eq. (2) with $\lambda^\star$, warm-start $\Pi_{\Delta_N} \tilde{h}_t$ // denoise
  $S_t \leftarrow$ top N_{syn} by $\hat{h}$, or resample $N_{\text{syn}} \propto \hat{h}$ // selection
return S_T

We estimate the clean histogram by the following constrained program:

$$\hat{h} \;=\; \arg\min_{x \in \Delta_N} \; \tfrac{1}{2} \left\| x - \tilde{h} \right\|_2^2 \;+\; \tfrac{\lambda}{2}\, x^\top \mathbf{L}\, x, \tag{2}$$

where x is clean counts, $x \in \Delta_N$ means non-negative sum-to-N constraints, and λ is the regularization factor for graph smoothness constraint.

One thing to note is that under replacement (bounded-DP) adjacency where the dataset size N is fixed, a neighbor is formed by swapping one private record for another, hence the L_2 norm of that change results in a sensitivity of $\sqrt{1^2 + 1^2} = \sqrt{2}$, rather than the original sensitivity of 1.

Proposition 2 (Coupling of count-only and graph-only estimators)**.** *The* GRE *objective couples two count-denoising operations: the exact simplex projection corresponding to the count-only MLE objective proposed by (Humphries et al., 2026), and unconstrained Laplacian smoothing on the candidate graph. When* $\lambda = 0$*,* GRE *reduces to the exact simplex projection onto* Δ_N *and vice versa.*

Figure 1 and Algorithm 1 demonstrate the pipeline: it combines the histogram count simplex constraints under the DP noise and smoothness across candidate neighborhoods (as in Proposition 2), where a denoised candidate voting histogram is recovered.

Theorem 1 (Unconditional error reduction via smoothing)**.** *Let* $\hat{h}_\lambda = (I + \lambda \mathbf{L})^{-1} \tilde{h}$ *with* $\tilde{h} = h + \mathcal{N}(0, \sigma^2 I)$*, and let* $R(\lambda) = \mathbb{E} \left\| \hat{h}_\lambda - h \right\|^2$*. If* $\sigma > 0$ *and* $\mathbf{L} \neq 0$*, then there exists* $\lambda > 0$ *with lower errors than the null estimate* $\lambda = 0$*, regardless of how non-smooth the true histogram is, as* $R'(0) = -2\sigma^2 \,\text{tr}(\mathbf{L}) < 0$*.*

Theorem 1 (proof in Appendix A) explains two things we see in §4. First, the improvement scales with the noise power $\sigma^2 \propto 1/\rho_t$, so the improvements is large exactly at tight budgets, where AugPE's raw estimate is noisiest. Secondly, it explains why even a little smoothing helps: near $\lambda = 0$ the error falls at first order in λ, while the bias that smoothing introduces grows only at second order (governed by $h^\top \mathbf{L}^2 h$), so a small λ works before the bias catches up.

Theorem 1 looks at the mean squared error, but the PE loop uses only the *ranking* of $\hat{h}$. A smaller error does not guarantee to give a better ranking, so the relevant question is when the denoising step would help. The following analyzes this:

Proposition 3 (Noise-dominated modes)**.** *Write* $\mathbf{L} = \sum_j \mu_j u_j u_j^\top$ *with spectral coefficients* $c_j = u_j^\top h$*, and let* $\tilde{h} = h + \mathcal{N}(0, \sigma^2 I)$*. Then:*

1. *In mode* j*, the MSE-optimal shrinkage* $a_j^\star = c_j^2/(c_j^2 + \sigma^2)$ *reduces its error by the factor* $\sigma^2/(c_j^2 + \sigma^2)$*, which exceeds* $\frac{1}{2}$ *if and only if* $\sigma^2 > c_j^2$*. Such a mode is considered noise-dominated.*

2. *The noise-dominated signal fraction*

$$P(\sigma, \mathbf{L}, h) = \frac{\sum_j c_j^2 \, \mathbf{1}[\sigma^2 > c_j^2]}{\sum_j c_j^2} \in [0, 1] \tag{3}$$

is non-decreasing in σ (equivalently, as ρ_t decreases).

We use P as an empirical indicator: it depends on c_j and σ, so P does not by itself determine top-N recovery. What P measures is how much of the signal sits in modes the noise dominates: large at tight budgets and where the histogram spreads its mass thinly, small where a few modes contain most of it. §4.3 shows P tracks when graph denoising helps (*and more importantly, when it does not*).

We minimize the convex problem (2) by FISTA (Beck & Teboulle, 2009) with accelerated projected gradient using a fixed step $1/(1 + 2\lambda)$ and the $O(M \log M)$ simplex projection of (Duchi et al., 2008). Selection then ranks (top-N) or resamples proportionally on $\hat{h}$.

Parts of the analysis above are machine-checked in Lean [1] (see Appendix B for more details as well as Supplementary Material for `.lean` files).

3.3 Choosing λ: selection/generation is not just estimation

Now the hardest problem becomes choosing the smoothing strength λ. The smoother $S_\lambda = (I + \lambda \mathbf{L})^{-1}$ shrinks each frequency j by $1/(1 + \lambda\mu_j)$, keeping the smooth components (small μ_j) and damping the rough ones (large μ_j) where the noise sits; a larger λ smooths more.

The obvious initial choice is the λ that minimizes error. Since σ is known, Stein's unbiased risk estimate (SURE) (Stein, 1981) gives the smoother's MSE:

$$\widehat{\mathrm{MSE}}(\lambda) = \sum_j c_j^2 \left(\tfrac{\lambda\mu_j}{1+\lambda\mu_j}\right)^2 + 2\sigma^2 \sum_j \tfrac{1}{1+\lambda\mu_j}, \qquad c = U^\top \tilde{h}, \tag{4}$$

However, this is the wrong target by itself since selection depends only on the *ranking* of $\hat{h}$, not on its values, and the two optima differ. The λ that best preserves the top-k ranking is 3–6$\times$ larger than the SURE value across budgets (shown in §4.4). Heavier smoothing flattens the peaks, which the ranking ignores, while it denoises the middle of the ordering where selection would potentially makes its mistakes. The mismatch is even larger for GRE, because the simplex projection in (2) already removes noise that SURE's linear-smoother MSE still charges for.

We therefore do not minimize MSE, rather we set λ so that the smoother's effective degrees of freedom $\mathrm{df}(\lambda) = \mathrm{tr}(S_\lambda) = \sum_j 1/(1 + \lambda\mu_j)$ match the number of occupied vote regions, estimated by the positive support of the simplex-MLE:

$$\lambda^\star \;=\; \arg\min_\lambda \; \big|\, \mathrm{df}(\lambda) - \hat{r} \,\big|, \qquad \hat{r} = \big|\{\, j : (\Pi_{\Delta_N} \tilde{h})_j > 0 \,\}\big|. \tag{5}$$

The count $\hat{r}$ shrinks as the noise grows, so the rule smooths more at tight budgets on its own, with no tuned constant.

Inside the PE loop there is one more effect: the denoised histogram is not ranked once; it decides which regions survive to the next generation. Smoothing heavier than (5), such as the one-shot ranking optimum near $\hat{r}/4$, improves a single round but hurts the final result (§4.4), because too few regions survive and the loop under-explores candidates. The best λ for MSE, for one-shot ranking, and for the closed loop are thus three different values, and (5) targets the closed loop. Both (4) and (5) use only σ, the graph, and $\tilde{h}$, so choosing λ costs no additional privacy.

4 Evaluations

We evaluate GRE primarily on three questions: **Q1 (histogram recovery):** how well does each method recover the true vote histogram from its DP release, as a post-processing function of the

[1] https://github.com/leanprover/lean4

privacy budget? **Q2 (end-to-end generation):** does better selection from denoising improve the privacy-utility trade-off and convergence speed of the full PE loop? **Q3 (choosing λ)**: how do we choose λ with three objectives hence three optima?

To provide a fair comparison, all methods share identical DP setup (cosine nearest-neighbor votes, Gaussian mechanism (with same σ), uniform budget split); the only thing that varies is the selection step, so every reported difference is attributable to selection alone. While an exhaustive enumeration of every conceivable evaluation configuration across a practically unbounded space of models and datasets approximates the hardness of counting all stars of our galaxy individually, we constrain our evaluation, to the best of our efforts, to AG News (Zhang et al., 2015) and DBpedia (Auer et al., 2007) as evaluation datasets, `Qwen3-32B` as generator model and `all-MiniLM-L6-v2` as embedding model to validate our method empirically.

4.1 Detailed Experimental Setup

Privacy accounting. The only privacy-consuming step is the Gaussian mechanism releasing the vote histogram and this is natively ρ-zCDP ($\sigma = \sqrt{T/\rho}$ under $\Delta_2 = \sqrt{2}$ bounded DP, per-iteration split $\rho_t = \rho/T$), so we report the total zCDP budget ρ throughout. Each ρ implies (ε, δ)-DP via $\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$ (Bun & Steinke, 2016); the guarantee is stated in ρ and this inferring bound is only used for cross-reference. The experiments contains ρ to ε relationships as: 1)in §4.2 with δ=3.5×10^{-5}: ρ=$0.004, 0.01, 0.02, 0.04, 0.1, 0.2, 0.6$ (ε=$0.36, 0.59, 0.86, 1.25, 2.07, 3.03, 5.56$); 2) in §4.3, δ=10^{-5}: ρ=$0.0035, 0.0128, 0.0468, 0.169$ (ε=$0.36, 0.72, 1.45, 2.92$).

Embeddings. All text is embedded with `all-MiniLM-L6-v2` ($d = 384$), as in the reference AugPE/MAPLE implementation.

Estimation benchmark. Synthetic geometry: $M = 2000$ candidates and $N = 4000$ private points drawn; kNN graph: $k = 15$, Gaussian kernel with median-distance bandwidth. λ grids: log-spaced, 80 points. Reported over 8 seeds.

PE runs. $T = 5$ iterations, $M = 2000$ candidates per round ($n_{\text{syn}} = 1000$ retained $\times$ $L = 2$ variations), $N_{\text{priv}} = 5000$, $\delta = 10^{-5}$, uniform zCDP split $\rho_t = \rho/T$, top-N selection, 5 seeds. Generation uses `Qwen3-32B` (temperature 1.0) via the AugPE variation API: each retained sample is expanded by $L = 2$ LLM paraphrases to form the next candidate pool, and the initialization is generated from the generator's prior per category (the MAPLE-style domain mismatch).

4.2 Q1: Histogram Recovery

Table 1: **Histogram recovery** (normalized MSE, lower better; and top-k selection Jaccard, higher better), mean over 8 seeds.

nMSE ↓	$\rho = 0.004$	$\rho = 0.010$	$\rho = 0.020$	$\rho = 0.040$	$\rho = 0.100$	$\rho = 0.200$	$\rho = 0.600$
AugPE (w/ clip)	6.368	2.629	1.360	0.708	0.300	0.157	0.056
SimplexMLE	1.591	0.975	0.641	0.404	0.206	**0.119**	**0.048**
GraphDenoise (ours)	0.723	0.540	0.416	0.314	0.212	0.156	0.098
GRE (ours)	**0.634**	**0.484**	**0.379**	**0.291**	**0.199**	0.149	0.095
Jaccard ↑	$\rho = 0.004$	$\rho = 0.010$	$\rho = 0.020$	$\rho = 0.040$	$\rho = 0.100$	$\rho = 0.200$	$\rho = 0.600$
AugPE (w/ clip)	0.379	0.400	0.422	0.450	0.498	0.537	0.628
SimplexMLE	0.367	0.385	0.410	0.432	0.492	0.537	0.628
GraphDenoise (ours)	**0.445**	**0.475**	**0.497**	**0.526**	**0.567**	**0.606**	**0.676**
GRE (ours)	0.419	0.453	0.482	0.515	0.558	0.601	0.675

Setup. We use a controlled clustered geometry ($M = 2000$ candidates, $N = 4000$ private points, $d = 64$, 12 clusters) with a known clean histogram h that is sparse and graph-smooth. We release $\tilde{h}$ at seven budgets and score each estimator against h by normalized MSE, top-k Jaccard (the quantity

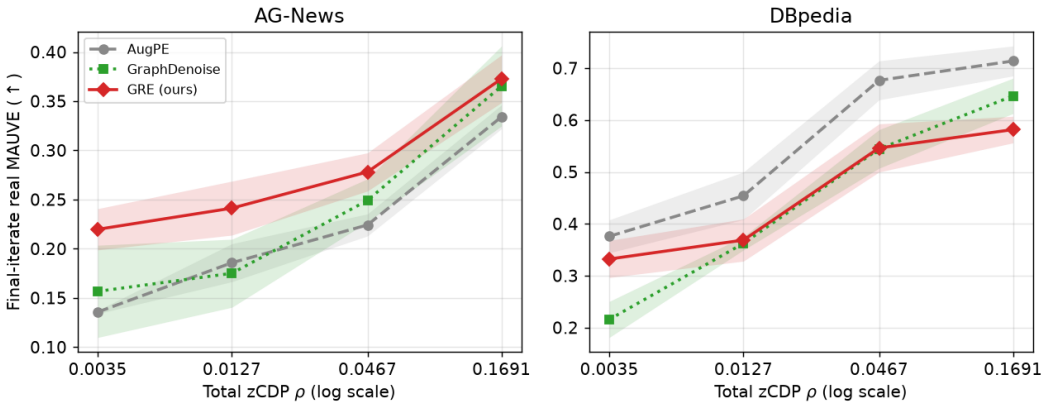

Figure 2: **End-to-end privacy–utility on text generation.** Final MAUVE vs. total zCDP ρ (log scale), top-N selection, mean over 5 seeds (shaded s.e.), `Qwen3-32B` generation. GRE improves over AugPE at every budget on AG News (left) but not on DBpedia (right), where the undenoised baseline is already strong.

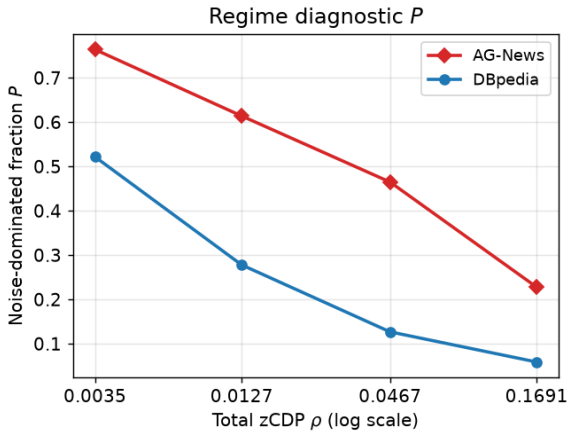

Figure 3: **Regime diagnostic.** The noise-dominated signal fraction P (Eq. 3), computed from the candidate graph, is $\sim$2–4$\times$ larger on AG News than DBpedia at every budget, matching where GRE helps (Fig. 2).

top-N selection acts on), and total-variation distance, over 8 seeds. We compare AugPE's clip, the simplex-MLE (Humphries et al., 2026), our graph-only GraphDenoise, and GRE.

As shown in Table 1, at tight budgets ($\rho \leq 0.04$) GRE has the lowest nMSE, $\sim$5.4$\times$ below clip and about $\sim$2$\times$ below the simplex-MLE at $\rho = 0.010$, and the graph methods in general have the best top-k Jaccard at every budget. The nMSE ranking reverses as noise decreases: at $\rho \geq 0.2$ clip and the simplex-MLE exceeds GRE on nMSE while GRE still leads on Jaccard metric. This benchmark shows that GRE works well for recovering noisy histogram, but lower error and better ranking are distinct objectives (§4.4), and GRE is strong on both only in the tight-budget regime, which will be discussed in the following section with end-to-end generation results.

4.3 Q2: End-to-end Private Evolution

Setup. We run the full PE loop with `Qwen3-32B` generation and score the final synthetic set against the private set with MAUVE (Pillutla et al., 2021), computed between the private set and the final synthetic set with the reference implementation over the shared `all-MiniLM-L6-v2` embedding space (20 quantization buckets, fixed seed); higher is closer. MAUVE is computed only after the loop finishes and is never fed back into selection. Following MAPLE's domain-mismatch setting,

Table 2: λ **ablation** (GRE estimator fixed; nMSE lower better, Jaccard/MAUVE higher better). nMSE and Jaccard are on the estimation benchmark (6 seeds) at three budgets; MAUVE is the end-to-end closed-loop top-N score from generation (AG News, ρ=0.0035, 3 seeds), measured only for the *df* scalings ("–" = skipped in the generation).

	$\rho = 0.010$		$\rho = 0.040$		$\rho = 0.200$		ρ=0.0035
λ rule	nMSE	Jac	nMSE	Jac	nMSE	Jac	MAUVE
SURE (MSE-opt., unconstrained)	0.467	0.495	0.304	0.535	0.131	0.594	–
df $= \hat{r}$ **(default)**	0.480	0.455	0.290	0.515	0.148	0.603	**0.24**
df $= \hat{r}/2$	0.467	0.494	0.354	0.558	0.266	0.636	0.11
df $= \hat{r}/4$	0.540	0.519	0.471	0.579	0.409	0.633	0.05
df $= 2\hat{r}$	0.628	0.417	0.339	0.465	0.121	0.541	0.14

the private target and the PE initialization are drawn from different distributions: the initialization is sampled from the generator's prior for each category, not from the private corpus. We use $T = 5$ iterations, $M = 2000$ candidates per round ($n_{\text{syn}} = 1000$ retained, $L = 2$ paraphrases each), $N_{\text{priv}} = 5000$, top-N selection, 5 seeds, at $\rho \in \{0.0035, 0.0128, 0.0468, 0.169\}$, on AG News and DBpedia.

AG News: where denoising helps at every budget. As shown in Figure 2, GRE improves MAUVE over AugPE in 17 of 20 (budget, seed) configurations at every budget, largest at the tightest ($+0.084$ at $\rho = 0.0035$) and decreases as the budget loosens (e.g., $+0.039$ at $\rho = 0.169$). Specifically, graph-only GraphDenoise is within noise of AugPE ($+0.017$, $9/20$), and GRE outperforms GraphDenoise by $+0.041$ ($13/20$), so the gain comes from coupling the graph with the simplex constraint, not from either alone.

DBpedia: where denoising does not help (aligns with Proposition 3). Same in Figure 2, on DBpedia GRE is below AugPE at every budget, as is GraphDenoise. The difference is the regime. DBpedia's undenoised MAUVE is already high (0.38–0.71, vs. 0.14–0.33 on AG News) and its histogram is more concentrated, so its signal is not noise-dominated: the noise-dominated fraction P (in Figure 3) is $\sim$2–4$\times$ smaller on DBpedia than on AG News at every budget. Where P is large (AG News) smoothing removes noise hence help with generation; where P is small (DBpedia) it mostly adds bias. GRE's benefit is thus empirically specific to the high-noise, weak-baseline regime that tight-budget DP synthetic text operates in, and P is a budget-free indicator for it.

4.4 Q3: Choosing λ: three misaligned objectives

Setup. We hold the estimator fixed at GRE and vary only the λ-selection criterion (Table 2), hence any difference is the smoothing strength, not the estimator. λ rules are chosen as SURE, the default *df* target, *df*-ratios.

As the results indicate, 1) one-shot ranking wants more smoothing than estimation where the SURE rule minimizes recovery error, but stronger smoothing keeps improving one-shot top-k Jaccard past that point. The ranking optimum sits at roughly 3–6$\times$ the SURE strength, and at 4.6–21$\times$ the constrained estimator's MSE optimum. We interpret this as that the simplex projection already does part of the denoising, so SURE over-smooths as a negative effect; 2) The closed loop wants less than one-shot ranking, shown in the closed-loop MAUVE column of Table 2 which is an inverted-U peaking at the default target c_{df}=1 ((5)) in our testing: heavier smoothing (df $= \hat{r}/2, \hat{r}/4$) and lighter smoothing (df $= 2\hat{r}$) each reduce ≥ 0.10 MAUVE, so the one-shot ranking optimum (3–6$\times$ SURE) already over-smooths the loop.

In summary, estimation, one-shot ranking, and closed-loop generation each prefer a different λ. Our default rule (5) targets the closed loop, which is the setting top-N uses.

5 Related Work

DP synthetic text splits into DP fine-tuning (Yue et al., 2023; Kurakin et al., 2023) and API-only Private Evolution (Lin et al., 2024; Xie et al., 2024). Recent PE work improves the iterative process

via multi-model fusion (Zou et al., 2025) or genetic operators (Sun et al., 2025) and improves initialization via DP metadata and in-context learning (MAPLE, Chien et al., 2026); our contribution is complimentary: it improves the selection step and composes with any of them since PE's release is fixed by the framework and we optimize the post-processing between release and decision.

Consistency post-processing of DP releases has a long history in query release: projecting noisy marginals onto a consistent polytope (Barak et al., 2007), hierarchical-histogram consistency (Hay et al., 2010), and the matrix mechanism (Li et al., 2015); for PE, Humphries et al. (2026) enforce a sum-to-N simplex constraint via truncated projection (Duchi et al., 2008). All enforce consistency among linear query answers; none uses distances among synthetic candidates or an embedding-neighbor graph, the geometric half we add and show matters for ranking-based selection under domain mismatch.

Laplacian/Tikhonov regularization on graphs is classical (Zhou & Schölkopf, 2004; Shuman et al., 2013), with SURE-style parameter selection for linear smoothers (Stein, 1981). To the best of our knowledge, we are not aware of prior work coupling graph smoothing with a simplex consistency constraint on a DP release; the three-way dissociation we observed (§4.4) is absent from that literature.

Selection under noisy directions in evolutionary computation studies how noise degrades selection and how averaging compensates (Arnold & Beyer, 2003; Jin & Branke, 2005). Our closed-loop finding, that denoising strong enough to maximize per-round accuracy can under-explore across generations, echoes that exploration–exploitation tension.

6 Conclusion

The vote histogram is the only quantity Private Evolution spends privacy on, and the existing pipeline largely under-utilizes it. We showed that denoising it with the public vote total and the synthetic candidate geometry, and tuning the denoiser for the closed loop rather than for estimation, improves both the recovered signal and the end-to-end privacy–utility trade-off without another private query under the public-N convention. Because the improvement is post-processing of the existing DP release, it is compatible with PE-family pipelines, but the utility gains we measure are specific to top-N selection from the PE initialization; the MAPLE-style composition sweep ran only the good-initialization arm and showed no final-iterate additive gain. Promising extensions include adapting the smoothing to the variation API's local step and jointly optimizing the per-iteration budget schedule with the denoiser.

The end-to-end benefit is dataset-dependent. Denoising helps when the DP noise dominates the released vote histogram and the un-denoised baseline is weak, which is the tight-budget setting where the generator starts far from the private data. It helps less, and can slightly lower the generation quality, when the baseline is already strong, because then there is little room for smoothing, which mostly adds bias. A further limit is the evaluation data. Current LLMs are heavily trained on common public benchmarks, so the generator's prior is already close to these corpora. We therefore expect the benefit to be clearest when the generator is genuinely far from the private distribution (e.g., business-specific internal sensitive data), a setting that public benchmarks understate and that we could not fully probe with the public datasets available to us.

USE OF LARGE LANGUAGE MODELS

In accordance with the ICLR 2027 policy on the use of large language models (LLMs), we disclose how LLMs were used in this work. The authors take full responsibility for all content, claims, proofs, code, and experimental results, including any produced with the aid of LLMs, and have verified them independently.

Research, proofs, and code. LLMs are used for the research ideation process on what could further improve the current state-of-the-art DP-based private evolution algorithms. The research idea of using post-processing property of DP combined with constraints of histogram simplex and candidate graph smoothing are from authors but refined with aid of LLMs.

LLM-based coding assistants were used to help implement and audit the experimental code (the Private Evolution loop, the GRE convex solver, the differential-privacy accounting, and the evaluation harness), and to help draft and check the Lean 4/Mathlib formalization of the machine-checked results in Appendix B. They were also used while developing and refining the theoretical statements (Propositions 1, 2 and Theorem 1) and their proofs. All theorem statements, proofs, and the Lean development were reviewed by the authors; the Lean proofs are additionally machine-checked (no `sorry`/`admit`/`axiom`), and every reported number was traced to the code that produced it.

Writing and figures. LLMs were used to revise for clarity and syntax as well as to help produce and refine result figures, and to assist with literature summarization and reference formatting. No text was included without author review.

LLMs as an object of study. Separately from the authoring assistance above, open-source LLMs are a studied component of the method itself: text generation uses `Qwen3-32B` and embeddings use `all-MiniLM-L6-v2`, as described in the experiments. These uses are part of the experimental design.

REFERENCES

Dirk V Arnold and Hans-Georg Beyer. On the benefits of populations for noisy optimization. *Evolutionary Computation*, 11(2):111–127, 2003.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pp. 722–735. Springer, 2007.

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 273–282, 2007.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, pp. 635–658, 2016.

Eli Chien, Yuzheng Hu, Ryan McKenna, Shanshan Wu, Zheng Xu, and Peter Kairouz. MAPLE: Metadata augmented private language evolution. In *Conference on Language Modeling (COLM)*, 2026.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pp. 272–279, 2008.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pp. 265–284, 2006.

Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1):1021–1032, 2010.

Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. SoK: Privacy-preserving data synthesis. In *IEEE Symposium on Security and Privacy (SP)*, pp. 4696–4713, 2024.

Thomas Humphries, Zinan Lin, and Sergey Yekhanin. PE-means: Improved differentially private k-means clustering through private evolution. *arXiv preprint arXiv:2606.00342*, 2026.

Yaochu Jin and Jürgen Branke. Evolutionary optimization in uncertain environments—a survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, 2005.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24 (6):757–781, 2015.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *International Conference on Learning Representations (ICLR)*, 2024.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:4816–4828, 2021.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Zhihao Sun, Zhe Tian, Yang Song, Yueyang Si, Jing Zhang, Min Huang, Kai Lu, Zhen Xiong, Xiao Liu, and Dan Li. DPGA-TextSyn: Differentially private genetic algorithm for synthetic text generation. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 16159–16179, 2025.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *International Conference on Machine Learning (ICML)*, 2024.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1321–1342, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Dengyong Zhou and Bernhard Schölkopf. A regularization framework for learning from graph data. In *ICML Workshop on Statistical Relational Learning*, 2004.

Tianyuan Zou, Yang Liu, Peng Li, Yufei Xiong, Jianqing Zhang, Jingjing Liu, Xiaozhou Ye, Ye Ouyang, and Ya-Qin Zhang. Contrastive private data synthesis via weighted multi-PLM fusion. *arXiv preprint arXiv:2502.00245*, 2025.

A PROOFS

This section contains the paper's proofs. A companion Lean 4 development (Appendix B) machine-checks the risk-derivative identity of Theorem 1 in its spectral/trace form. The remaining claims—differential privacy, the convex-equivalence of Proposition 2, and the SURE identity—are established here or by cited results, not formalized.

Proof of Proposition 1 (post-processing of DP histograms). The mechanism releases $\tilde{h} = h + \mathcal{N}(0, \sigma^2 I)$, which is ρ_t-zCDP by the Gaussian mechanism with $\Delta_2(h) = \sqrt{2}$ under bounded (replacement) adjacency: neighboring datasets swap one record keeping the size N fixed and public, so a swap moves one vote between two bins. The synthetic population S_t is produced from the previous iteration's output, which is itself a function of prior DP releases and of public randomness (the variation API and resampling use only $\tilde{h}$ and fresh randomness, never the raw private set). Hence $\mathbf{L} = \mathbf{L}(S_t)$ and N are public given the transcript of DP releases. The estimator $\hat{h} = \arg\min_{x \in \Delta_N} \frac{1}{2} \left\| x - \tilde{h} \right\|^2 + \frac{\lambda}{2} x^\top \mathbf{L} x$ is a deterministic function of $(\tilde{h}, \mathbf{L}, N, \sigma)$, and the resampling that produces the selection w is a *data-independent Markov kernel* applied to $\hat{h}$ (it uses only $\hat{h}$ and fresh randomness, never the private set). Post-processing immunity of DP covers both deterministic maps and data-independent randomization, so releasing $\hat{h}$ and w incurs no privacy cost beyond the ρ_t already charged for $\tilde{h}$. Composition over T iterations therefore yields the same total $\rho = \sum_t \rho_t$ as vanilla PE. □

Proof of Proposition 2. (i) At $\lambda = 0$ the penalty vanishes and (2) becomes $\arg\min_{x \in \Delta_N} \frac{1}{2} \left\| x - \tilde{h} \right\|^2$, the Euclidean projection of $\tilde{h}$ onto the scaled simplex $\Delta_N = \{x \geq 0, \mathbf{1}^\top x = N\}$. This is the exact $O(M \log M)$ simplex projection of Duchi et al. (2008), which Humphries et al. (2026) approximate by truncation; it is also the MLE of h under isotropic Gaussian noise subject to the two public constraints, since the negative log-likelihood is $\frac{1}{2\sigma^2} \left\| x - \tilde{h} \right\|^2$ up to constants and the feasible set is Δ_N. (ii) Relaxing the constraint set (optimizing over all of $\mathbb{R}^M$) makes (2) an unconstrained strictly convex quadratic; setting its gradient $(x - \tilde{h}) + \lambda \mathbf{L} x$ to zero gives the unique minimizer $\hat{h} = (I + \lambda \mathbf{L})^{-1} \tilde{h}$, the graph-only Laplacian smoother. Thus GRE exactly recovers the simplex estimator at $\lambda = 0$ and recovers the graph-only smoother after relaxing the simplex constraints. □

Proof of Theorem 1 (smoothing strictly helps). Write $\mathbf{L} = U \operatorname{diag}(\mu) U^\top$ and $c = U^\top h$ (clean spectral coefficients). The smoother acts spectrally with factors $a_j(\lambda) = 1/(1 + \lambda \mu_j)$, so the risk decomposes into bias and variance:

$$R(\lambda) \;=\; \sum_j (1 - a_j)^2 c_j^2 \;+\; \sigma^2 \sum_j a_j^2.$$

Differentiating with $a_j'(\lambda) = -\mu_j a_j^2$:

$$R'(\lambda) \;=\; \sum_j 2(1 - a_j)\, \mu_j a_j^2\, c_j^2 \;-\; 2\sigma^2 \sum_j \mu_j a_j^3.$$

At $\lambda = 0$ every $a_j = 1$, so the bias term vanishes and $R'(0) = -2\sigma^2 \sum_j \mu_j = -2\sigma^2 \operatorname{tr}(\mathbf{L})$. For the symmetric positive semidefinite $\mathbf{L}$ (normalized Laplacian of the symmetrized graph) the eigenvalues satisfy $\mu_j \geq 0$, so $\mathbf{L} \neq 0$ gives $\operatorname{tr}(\mathbf{L}) > 0$ and $R'(0) < 0$ whenever $\sigma > 0$. Hence the risk strictly decreases for small $\lambda > 0$. The derivative scales linearly in σ^2. The second-order Taylor expansion has variance-curvature terms and the pure bias contribution $\lambda^2 h^\top \mathbf{L}^2 h + o(\lambda^2)$, giving the stated local dependence of the achievable gain. This argument needs no smoothness assumption on h. □

SURE identity of (4). For a linear estimator $\hat{h} = S\tilde{h}$ with $\tilde{h} \sim \mathcal{N}(h, \sigma^2 I)$, Stein's lemma gives the unbiased risk estimate $\widehat{\mathrm{MSE}} = \left\| S\tilde{h} - \tilde{h} \right\|^2 + 2\sigma^2 \operatorname{tr}(S) - M\sigma^2$. Substituting the spectral form $S_\lambda = U \operatorname{diag}(1/(1 + \lambda\mu_j)) U^\top$ and $c = U^\top \tilde{h}$: $\left\| S_\lambda \tilde{h} - \tilde{h} \right\|^2 = \sum_j c_j^2 (\lambda \mu_j / (1 + \lambda \mu_j))^2$

and $\mathrm{tr}(S_\lambda) = \sum_j 1/(1+\lambda\mu_j)$, which yields (4) after dropping the constant $-M\sigma^2$. Minimizing this estimate over a grid gives the SURE-selected smoother. Our implementation uses the leading eigenpairs and a tail approximation for large graphs; the displayed identity is the full-spectrum formula. □

Proof of Proposition 3. Work in the Laplacian eigenbasis. Mode j of the release is $\tilde{c}_j = c_j + \sigma z_j$ with $z_j \sim \mathcal{N}(0,1)$ independent. Linear shrinkage $a\tilde{c}_j$ has risk $(1-a)^2c_j^2 + a^2\sigma^2$, minimized at $a_j^\star = c_j^2/(c_j^2+\sigma^2)$ with minimum $c_j^2\sigma^2/(c_j^2+\sigma^2)$. The unshrunk estimate ($a=1$) has risk σ^2, so optimal shrinkage cuts the mode's risk by $1 - c_j^2/(c_j^2+\sigma^2) = \sigma^2/(c_j^2+\sigma^2)$, which exceeds $\frac{1}{2}$ iff $\sigma^2 > c_j^2$. For (ii), as σ increases the set $\{j : \sigma^2 > c_j^2\}$ only grows, so the numerator $\sum_j c_j^2 \mathbf{1}[\sigma^2 > c_j^2]$ is non-decreasing in σ; the denominator $\sum_j c_j^2$ is independent of σ, hence P is non-decreasing in σ, and with $\sigma^2 = 1/\rho_t$, as ρ_t decreases. P is an empirical diagnostic: it uses c_j and σ but not the eigenvalues μ_j, so it does not determine the smoother's action or top-N recovery, which we assess empirically (§4.3). □

B MACHINE-CHECKED PROOFS (LEAN 4)

We accompany the paper with a *Lean 4* proof project (toolchain `leanprover/lean4:v4.33.1`) that machine-checks the core claims. It builds green and contains no `sorry`, `admit`, or `axiom`; `#print axioms` on the theorems below reports only Lean's standard `propext`, `Classical.choice`, and `Quot.sound`. We state precisely which results are machine-checked and which are proved on paper or cited.

Machine-checked in Lean.

- **Theorem 1 ($R'(0) = -2\sigma^2\,\mathrm{tr}(\mathbf{L})$), full spectral/trace form.** Writing the one-mode risk as $r(\lambda) = (1-a)^2c^2 + \sigma^2a^2$ with $a = (1+\lambda\mu)^{-1}$, we prove the calculus step `hasDerivAt_modeRisk_zero`: $r'(0) = -2\sigma^2\mu$ (via Mathlib's `HasDerivAt` API—the reciprocal rule, then the square/product/sum rules), assemble it over the spectrum (`hasDerivAt_specRisk_zero` over a list of eigenpairs and `hasDerivAt_specRiskFinset_zero` via `HasDerivAt.sum`), and finally, for a real symmetric matrix A, prove `hasDerivAt_specRisk_trace_zero`: $\frac{d}{d\lambda}\big|_0 \sum_i r_i(\lambda) = -2\sigma^2\,\mathrm{tr}(A)$, using `Matrix.IsHermitian.trace_eq_sum_eigenvalues` to identify $\sum_j \mu_j = \mathrm{tr}(A)$. This is the faithful, machine-checked statement of $R'(0) = -2\sigma^2\,\mathrm{tr}(\mathbf{L})$ for the unconstrained smoother.
- **Proposition 3, algebraic parts.** `shrinkage_halving`: the inequality that a shrinkage factor $c^2/(c^2+\sigma^2)$ exceeds $\frac{1}{2}$ iff $c^2 > \sigma^2$ (equivalently $\sigma^2/(c^2+\sigma^2) > \frac{1}{2}$ iff $\sigma^2 > c^2$); Lean checks the inequality, not that this factor is MSE-optimal (proved on paper). `P_numer_mono`: the numerator $\sum_j c_j^2\mathbf{1}[\sigma^2 > c_j^2]$ is non-decreasing in σ.
- **Structural roles (Proposition 2, logical form).** A small type encodes the two structures an estimator may enforce (the count total and graph geometry); it is checked that GRE has both while the count-only and graph-only estimators each have exactly one. This is the Boolean feature table, not the convex first-order conditions.
- **Post-processing (Proposition 1, logical form) and objective dissociation.** It is checked that composing a deterministic and a randomized map preserves an input-independence predicate (the closure shape of post-processing, a toy leakage model, *not* measure-theoretic DP), and that three tuning targets can be pairwise distinct (the logical possibility behind the MSE / one-shot-ranking / closed-loop dissociation of §4.4).

Proved on paper or cited (not in Lean).

- Proposition 1 as a *differential-privacy* statement. Mathlib v4.33.1 has no zCDP / Gaussian-mechanism / DP-composition library, so we rely on the standard post-processing theorem (Dwork & Roth, 2014; Bun & Steinke, 2016) and the proof of Appendix A; we do *not* axiomatize DP.

- The convex first-order conditions of Proposition 2 ($\lambda = 0$ simplex projection; relaxed-constraint graph smoother) and the SURE identity (4), a standard consequence of Stein's lemma (Stein, 1981).
- The two proof–practice gaps noted in the main text: Theorem 1 bounds the *unconstrained* smoother rather than the deployed simplex-constrained estimator (2), and the deployed SURE selector uses leading eigenpairs with a tail approximation whereas (4) is the full-spectrum identity.

Reproducing the build. The project lives in `lean/` with two Lake packages: `lean_grmle_mathlib` (Mathlib-backed; the `RiskDerivative` results above, including the matrix/trace form of Theorem 1; builds from the committed `lake-manifest.json` via `lake exe cache get`) and `lean_grmle_proofs` (dependency-free; the structural, post-processing, and dissociation shapes). Both build with `lake build` under Lean 4.33.1 with no `sorry`/`admit`/`axiom`. In short: Theorem 1 (matrix/trace form) is machine-checked; Propositions 1 and 2's convex step and (4) are proved on paper or cited as standard results.