Discrete Distribution Estimation under Local Privacy

Peter Kairouz $*\dagger$ KAIROUZ2@ILLINOIS.EDU
Keith Bonawitz $*$ BONAWITZ@GOOGLE.COM
Daniel Ramage $*$ DRAMAGE@GOOGLE.COM
$*$ Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043,
$\dagger$ University of Illinois, Urbana-Champaign, 1308 W Main St, Urbana, IL 61801

Abstract

The collection and analysis of user data drives improvements in the app and web ecosystems, but comes with risks to privacy. This paper examines discrete distribution estimation under local privacy, a setting wherein service providers can learn the distribution of a categorical statistic of interest without collecting the underlying data. We present new mechanisms, including hashed k-ary Randomized Response (k-RR), that empirically meet or exceed the utility of existing mechanisms at all privacy levels. New theoretical results demonstrate the order-optimality of k-RR and the existing RAPPOR mechanism at different privacy regimes.

1. Introduction

Software and service providers increasingly see the collection and analysis of user data as key to improving their services. Datasets of user interactions give insight to analysts and provide training data for machine learning models. But the collection of these datasets comes with risk—can the service provider keep the data secure from unauthorized access? Misuse of data can violate the privacy of users and substantially tarnish the provider's reputation.

One way to minimize risk is to store less data: providers can methodically consider what data to collect and how long to store it. However, even a carefully processed dataset can compromise user privacy. In a now famous study, (Narayanan & Shmatikov, 2008) showed how to de-anonymize watch histories released in the Netflix Prize, a public recommender system competition. While most providers do not intentionally release anonymized datasets, security breaches can mean that even internal, anonymized datasets have the potential to become privacy problems.

Fortunately, mathematical formulations exist that can give the benefits of population-level statistics without the collection of raw data. Local differential privacy (Duchi et al., 2013a;b) is one such formulation, requiring each device (or session for a cloud service) to share only a noised version of its raw data with the service provider's logging mechanism. No matter what computation is done to the noised output of a locally differentially private mechanism, any attempt to impute properties of a single record will have a significant probability of error. But not all differentially private mechanisms are equal when it comes to utility: some mechanisms have better accuracy than others for a given analysis, amount of data, and desired privacy level.

Private distribution estimation. This paper investigates the fundamental problem of discrete distribution estimation under local differential privacy. We focus on discrete distribution estimation because it enables a variety of useful capabilities, including usage statistics breakdowns and count-based machine learning models, e.g. naive Bayes (McCallum et al., 1998). We consider empirical, maximum likelihood, and minimax distribution estimation, and study the price of local differential privacy under a variety of loss functions and privacy regimes. In particular, we compare the performance of two recent local privacy mechanisms: (a) the Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) (Erlingsson et al., 2014), and (b) the k-ary Randomized Response (k-RR) (Kairouz et al., 2014) from a theoretical and empirical perspective.

Our contributions are:

1. For binary alphabets, we prove that Warner's randomized response model (Warner, 1965) is globally optimal for any loss function and any privacy level (Section 3).

2. For k-ary alphabets, we show that RAPPOR is order optimal in the high privacy regime and strictly sub-optimal in the low privacy regime for ℓ_1 and ℓ_2 losses using an empirical estimator. Conversely, k-RR is order optimal in the low privacy regime and strictly sub-optimal in the

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

high privacy regime (Section 4.1).

3. Large scale simulations show that the optimal decoding algorithm for both k-RR and RAPPOR depends on the shape of the true underlying distribution. For skewed distributions, the *projected estimator* (introduced here) offers the best utility across a wide variety of privacy levels and sample sizes (Section 4.4).

4. For open alphabets in which the set of input symbols is not enumerable *a priori* we construct the O-RR mechanism (an extension to k-RR using hash functions and cohorts) and provide empirical evidence that the performance of O-RR meets or exceeds that of RAPPOR over a wide range of privacy settings (Section 5).

5. We apply the O-RR mechanism to closed k-ary alphabets, replacing hash functions with permutations. We provide empirical evidence that the performance of O-RR meets or exceeds that of k-RR and RAPPOR in both low and high privacy regimes (Section 5.4).

Related work. There is a rich literature on distribution estimation under local privacy (Chan et al., 2012; Hsu et al., 2012; Bassily & Smith, 2015), of which several works are particularly relevant herein. (Warner, 1965) was the first to study the local privacy setting and propose the randomized response model that will be detailed in Section 3. (Kairouz et al., 2014) introduced k-RR and showed that it is optimal in the low privacy regime for a rich class of information theoretic utility functions. k-RR will be extended to open alphabets in Section 5.1. (Duchi et al., 2013a;b) was the first to apply differential privacy to the local setting, to study the fundamental trade-off between privacy and minimax distribution estimation in the high privacy regime, and to introduce the core of k-RAPPOR. (Erlingsson et al., 2014) proposed RAPPOR, systematically addressing a variety of practical issues for private distribution estimation, including robustness to attackers with access to multiple reports over time, and estimating distributions over open alphabets. RAPPOR has been deployed in the Chrome browser to allow Google to privately monitor the impact of malware on homepage settings. RAPPOR will be investigated in Sections 4.2 and 5.2.

Private distribution estimation also appears in the global privacy context where a trusted service provider releases randomized data (e.g., NIH releasing medical records) to protect sensitive user information (Dwork, 2006; Dwork et al., 2006; Dwork & Lei, 2009; Dwork, 2008; Diakonikolas et al., 2015; Blocki et al., 2016).

2. Preliminaries

2.1. Local differential privacy

Let X be a private source of information defined on a discrete, finite input alphabet $\mathcal{X} = \{x_1, ..., x_k\}$. A statistical privatization mechanism is a family of distributions $\boldsymbol{Q}$ that map $X = x$ to $Y = y$ with probability $\boldsymbol{Q}(y|x)$. Y, the privatized version of X, is defined on an output alphabet $\mathcal{Y} = \{y_1, ..., y_l\}$ that need not be identical to the input alphabet $\mathcal{X}$. In this paper, we will represent a privatization mechanism $\boldsymbol{Q}$ via a $k \times l$ row-stochastic matrix. A conditional distribution $\boldsymbol{Q}$ is said to be ε-locally differentially private if for all $x, x' \in \mathcal{X}$ and all $E \subset \mathcal{Y}$, we have that

$$\boldsymbol{Q}(E|x) \leq e^{\varepsilon} \boldsymbol{Q}(E|x'), \tag{1}$$

where $\boldsymbol{Q}(E|x) = \mathbb{P}(Y \in E | X = x)$ and $\varepsilon \in [0, \infty)$ (Duchi et al., 2013a) . In other words, by observing $Y \in E$, the adversary cannot reliably infer whether $X = x$ or $X = x'$ (for any pair x and x'). Indeed, the smaller the ε is, the closer the likelihood ratio of $X = x$ to $X = x'$ is to 1. Therefore, when ε is small, the adversary cannot recover the true value of X reliably.

2.2. Private distribution estimation

The private multinomial estimation problem is defined as follows. Given a vector $\boldsymbol{p} = (p_1, ..., p_k)$ on the probability simplex $\mathbb{S}^k$, samples $X_1, ..., X_n$ are drawn i.i.d. according to $\boldsymbol{p}$. An ε-locally differentially private mechanism $\boldsymbol{Q}$ is then applied independently to each sample X_i to produce $Y^n = (Y_1, \cdots, Y_n)$, the sequence of private observations. Observe that the Y_i's are distributed according to $\boldsymbol{m} = \boldsymbol{p}\boldsymbol{Q}$ and not $\boldsymbol{p}$. Our goal is to estimate the distribution vector $\boldsymbol{p}$ from Y^n.

Privacy vs. utility. There is a fundamental trade-off between utility and privacy. The more private you want to be, the less utility you can get. To formally analyze the privacy-utility trade-off, we study the following constrained minimization problem

$$r_{\ell,\varepsilon,k,n} = \inf_{\boldsymbol{Q} \in \mathcal{D}_\varepsilon} r_{\ell,\varepsilon,k,n}(\boldsymbol{Q}), \tag{2}$$

where

$$r_{\ell,\varepsilon,k,n}(\boldsymbol{Q}) = \inf_{\hat{\boldsymbol{p}}} \sup_{\boldsymbol{p}} \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{p}\boldsymbol{Q}} \ell(\boldsymbol{p}, \hat{\boldsymbol{p}})$$

is the minimax risk under $\boldsymbol{Q}$, ℓ is an application dependent loss function, and $\mathcal{D}_\varepsilon$ is the set of all ε-locally differentially private mechanisms.

This problem, though of great value, is intractable in general. Indeed, finding minimax estimators in the non-private setting is already hard for several loss functions. For instance, the minimax estimator under ℓ_1 loss is unknown even until today. However, in the high privacy regime, we are able to bound the minimax risk of any differentially private mechanism $\boldsymbol{Q}$.

Proposition 1 *For the private distribution estimation problem in* (2)*, for any ε-locally differentially private mechanism $\boldsymbol{Q}$, there exist universal constants $0 < c_l \leq c_u < 5$ such that for all $\varepsilon \in [0, 1]$,*

$$c_l \min\left\{1, \frac{1}{\sqrt{n\varepsilon^2}}, \frac{k}{n\varepsilon^2}\right\} \leq r_{\ell_2^2,\varepsilon,k,n} \leq c_u \min\left\{1, \frac{k}{n\varepsilon^2}\right\},$$

and

$$c_l \min\left\{1, \frac{k}{\sqrt{n\varepsilon^2}}\right\} \leq r_{\ell_1,\varepsilon,k,n} \leq c_u \min\left\{1, \frac{k}{\sqrt{n\varepsilon^2}}\right\}$$

Proof See (Duchi et al., 2013b). ■

This result shows that in the high privacy regime ($\varepsilon \leq 1$), the effective sample size of a dataset decreases from n to $n\varepsilon^2/k$. In other words, a factor of k/ε^2 extra samples are needed to achieve the same minimax risk. This is problematic for large alphabets. Our work shows that (a) this problem can be (partially) circumvented using a combination of cohort-style hashing and k-RR (Section 5), and (b) the dependence on the alphabet size vanishes in the moderate to low privacy regime (Section 4.3).

3. Binary Alphabets

In this section, we study the problem of private distribution estimation under binary alphabets. In particular, we show that Warner's randomized response model (W-RR) is optimal for binary distribution minimax estimation (Warner, 1965). In W-RR, interviewees flip a biased coin (that only they can see the result of), such that a fraction η of participants answer the question "Is the predicate P true (of you)?" while the remaining particants answer the negation ("Is $\neg P$ true?"), without revealing which question they answered. For $\eta = e^\varepsilon$ ($\varepsilon \geq 0$), W-RR can be described by the following 2×2 row-stochastic matrix

$$\boldsymbol{Q}_{\text{WRR}} = \frac{1}{e^\varepsilon + 1}\begin{bmatrix} e^\varepsilon & 1 \\ 1 & e^\varepsilon \end{bmatrix}. \tag{3}$$

It is easy to check that the above mechanism satisfies the constraints imposed by local differential privacy.

Theorem 2 *For all binary distributions $\boldsymbol{p}$, all loss functions ℓ, and all privacy levels ε, $\boldsymbol{Q}_{WRR}$ is the optimal solution to the private minimax distribution estimation problem in (2).*

Proof sketch. (Kairouz et al., 2014) showed that W-RR dominates all other differentially private mechanisms in a strong Markovian sense: for any binary differentially private mechanism $\boldsymbol{Q}$, there exists a 2×2 stochastic mapping $\boldsymbol{W}$ such that $\boldsymbol{Q} = \boldsymbol{W} \circ \boldsymbol{Q}_{\text{WRR}}$. Therefore, for any risk function $r(\cdot)$ that obeys the data processing inequality ($r(\boldsymbol{Q}) \leq r(\boldsymbol{Q} \circ \boldsymbol{W})$ for any stochastic mappings $\boldsymbol{Q}$ and $\boldsymbol{W}$), we have that $r(\boldsymbol{Q}_{\text{WRR}}) \leq r(\boldsymbol{Q})$ for any binary differentially private mechanism $\boldsymbol{Q}$. In Supplementary Section A, we prove that $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$ obeys the data processing inequality, thus W-RR achieves the optimal privacy-utility trade-off under minimax distribution estimation.

4. k-ary Alphabets

Above, we saw that W-RR is optimal for all privacy levels and all loss functions. However, it can only be applied to binary alphabets. In this section, we study optimal privacy mechanisms for k-ary alphabets. We show that under ℓ_1 and ℓ_2 losses, k-RAPPOR is order optimal in the high privacy regime and sub-optimal in the low privacy regime. Conversely, k-RR is order optimal in the low privacy regime and sub-optimal in the high privacy regime.

4.1. The k-ary Randomized Response

The *k-ary randomized response* (k-RR) mechanism is a locally differentially private mechanism that maps $\mathcal{X}$ stochastically onto itself (i.e., $\mathcal{Y} = \mathcal{X}$), given by

$$\boldsymbol{Q}_{\text{KRR}}(y|x) = \frac{1}{k - 1 + e^\varepsilon}\begin{cases} e^\varepsilon & \text{if } y = x, \\ 1 & \text{if } y \neq x. \end{cases} \tag{4}$$

k-RR can be viewed as a multiple choice generalization of the W-RR mechanism (note that k-RR reduces to W-RR for $k = 2$). In (Kairouz et al., 2014), the k-RR mechanism was shown to be optimal in the low privacy regime for a large class of information theoretic utility functions.

Empirical estimation under k-RR. It is easy to see that under $\boldsymbol{Q}_{\text{KRR}}$, outputs are distributed according to:

$$\boldsymbol{m} = \frac{e^\varepsilon - 1}{e^\varepsilon + k - 1}\boldsymbol{p} + \frac{1}{e^\varepsilon + k - 1} \tag{5}$$

The empirical estimate of $\boldsymbol{p}$ under $\boldsymbol{Q}_{\text{KRR}}$ is given by

$$\begin{aligned} \hat{\boldsymbol{p}} &= \hat{\boldsymbol{m}}\boldsymbol{Q}_{\text{KRR}}^{-1} \\ &= \frac{e^\varepsilon + k - 1}{e^\varepsilon - 1}\hat{\boldsymbol{m}} - \frac{1}{e^\varepsilon - 1}, \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{m}}$ is the empirical estimate of $\boldsymbol{m}$ and

$$\boldsymbol{Q}_{\text{KRR}}^{-1}(y|x) = \frac{1}{e^\varepsilon - 1}\begin{cases} e^\varepsilon + k - 2 & \text{if } y = x, \\ -1 & \text{if } y \neq x. \end{cases} \tag{7}$$

via the Sherman-Morrison formula. Observe that because $\hat{\boldsymbol{m}} \to \boldsymbol{m}$ almost surely, $\hat{\boldsymbol{p}} \to \boldsymbol{p}$ almost surely.

Proposition 3 *For the private distribution estimation problem under k-RR and its empirical estimator given in* (6)*, for all ε, n, and k, we have that*

$$\mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}) = \frac{1-\sum_{i=1}^k p_i^2}{n} + \frac{k-1}{n}\left(\frac{k+2(e^\varepsilon-1)}{(e^\varepsilon-1)^2}\right),$$

and for large n, $\mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}) \approx$

$$\sum_{i=1}^k \sqrt{\frac{2((e^\varepsilon-1)p_i+1)((e^\varepsilon-1)(1-p_i)+k-1)}{\pi n(e^\varepsilon-1)^2}},$$

where $a_n \approx b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

Proof See Supplementary Section B. ■

Observe that for $\boldsymbol{p}_{\mathrm{U}} = \left(\frac{1}{k},\cdots,\frac{1}{k}\right)$, we have that

$$\begin{aligned}\mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}) &\leq \mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}_{\mathrm{U}}) \qquad (8)\\ &= \left(1+\frac{k+2(e^\varepsilon-1)}{(e^\varepsilon-1)^2}k\right)\frac{1-\frac{1}{k}}{n},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}) &\leq \mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}_{\mathrm{U}}) \qquad (9)\\ &\approx \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)\sqrt{\frac{2(k-1)}{\pi n}}.\end{aligned}$$

Constraining empirical estimates to $\mathbb{S}^k$. It is easy to see that $||\hat{\boldsymbol{p}}_{\mathrm{KRR}}||_1 = 1$. However, some of the entries of $\hat{\boldsymbol{p}}_{\mathrm{KRR}}$ can be negative (especially for small values of n). Several remedies are available, including (a) truncating the negative entries to zero and renormalizing the entire vector to sum to 1, or (b) projecting $\hat{\boldsymbol{p}}_{\mathrm{KRR}}$ onto the probability simplex. We evaluate both approaches in Section 4.4.

4.2. k-RAPPOR

The randomized aggregatable privacy-preserving ordinal response (RAPPOR) is an open source Google technology for collecting aggregate statistics from end-users with strong local differential privacy guarantees (Erlingsson et al., 2014). The simplest version of RAPPOR, called the basic one-time RAPPOR and referred to herein as k-RAPPOR, first appeared in (Duchi et al., 2013a;b). k-RAPPOR maps the input alphabet $\mathcal{X}$ of size k to an output alphabet $\mathcal{Y}$ of size 2^k. In k-RAPPOR, we first map $\mathcal{X}$ deterministically to $\tilde{\mathcal{X}} = \mathbb{R}^k$, the k-dimensional Euclidean space. Precisely, $X = x_i$ is mapped to $\tilde{X} = e_i$, the i^{th} standard basis vector in $\mathbb{R}^k$. We then randomize the coordinates of $\tilde{X}$ independently to obtain the private vector $Y \in \{0,1\}^k$. Formally, the j^{th} coordinate of Y is given by: $Y^{(j)} = \tilde{X}^{(j)}$ with probability $e^{\varepsilon/2}/(1+e^{\varepsilon/2})$ and $1-\tilde{X}^{(j)}$ with probability $1/(1+e^{\varepsilon/2})$. The randomization in $\boldsymbol{Q}_{k\text{-RAPPOR}}$ is ε-locally differentially private (Duchi et al., 2013a; Erlingsson et al., 2014).

Under k-RAPPOR, $Y_i = [Y_i^{(1)},\cdots,Y_i^{(k)}]$ is a k-dimensional binary vector, which implies that

$$\mathbb{P}(Y_i^{(j)}=1) = \left(\frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1}\right)p_j + \frac{1}{e^{\varepsilon/2}+1}, \tag{10}$$

for all $i \in \{1,\cdots,n\}$ and $j \in \{1,\cdots,k\}$.

Empirical estimation under k-RAPPOR. Let Y^n be the $n\times k$ matrix formed by stacking the row vectors $Y_1,\cdots,Y_n$ on top of each other. The empirical estimator of $\boldsymbol{p}$ under k-RAPPOR is:

$$\hat{p}_j = \left(\frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1}\right)\frac{T_j}{n} - \frac{1}{e^{\varepsilon/2}-1}, \tag{11}$$

where $T_j = \sum_{i=1}^n Y_i^{(j)}$. Because T_j/n converges to m_j almost surely, $\hat{p}_j$ converges to p_j almost surely. As with k-RR, we can constrain $\hat{\boldsymbol{p}}$ to $\mathbb{S}^k$ through truncation and normalization or through projection (described in Section 4.1), both of which will be evaluated in Section 4.4.

Proposition 4 *For the private distribution estimation problem under k-*RAPPOR *and its empirical estimator given in* (11)*, for all ε, n, and k, we have that*

$$\mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}) = \frac{1-\sum_{i=1}^k p_i^2}{n} + \frac{ke^{\varepsilon/2}}{n(e^{\varepsilon/2}-1)^2},$$

and for large n, $\mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}) \approx$

$$\sum_{i=1}^k \sqrt{\frac{2((e^{\varepsilon/2}-1)p_i+1)((e^{\varepsilon/2}-1)(1-p_i)+1)}{\pi n(e^{\varepsilon/2}-1)^2}},$$

where $a_n \approx b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

Proof See Supplementary Section C. ■

Observe that for $\boldsymbol{p}_{\mathrm{U}} = \left(\frac{1}{k},\cdots,\frac{1}{k}\right)$, we have that

$$\begin{aligned}\mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}) &\leq \mathbb{E}\,\ell_2^2(\hat{\boldsymbol{p}},\boldsymbol{p}_{\mathrm{U}}) \qquad (12)\\ &= \left(1+\frac{k^2e^{\varepsilon/2}}{(k-1)(e^{\varepsilon/2}-1)^2}\right)\frac{1-\frac{1}{k}}{n},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}) &\leq \mathbb{E}\,\ell_1(\hat{\boldsymbol{p}},\boldsymbol{p}_{\mathrm{U}}) \qquad (13)\\ &\approx \sqrt{\frac{(e^{\varepsilon/2}+k-1)(e^{\varepsilon/2}(k-1)+1)}{(e^{\varepsilon/2}-1)^2(k-1)}}\sqrt{\frac{2(k-1)}{\pi n}}.\end{aligned}$$

4.3. Theoretical Analysis

We now analyze the performance of k-RR and k-RAPPOR relative to maximum likelihood estimation (which is equivalent to empirical estimation) on the non-privatized data

X^n. In the non-private setting, the maximum likelihood estimator has a worst case risk of $\sqrt{\frac{2(k-1)}{\pi n}}$ under the ℓ_1 loss, and a worst case risk of $\frac{1-\frac{1}{k}}{n}$ under the ℓ_2^2 loss (Lehmann & Casella, 1998; Kamath et al., 2015).

Performance under k-RR. Comparing Equation (8) to the observation above, we can see that an extra factor of $\left(1+\frac{k+2(e^\varepsilon-1)}{(e^\varepsilon-1)^2}k\right)$ samples is needed to achieve the same ℓ_2^2 loss as in the non-private setting. Similarly, from Equation (9), a factor of $\left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)^2$ samples is needed under the ℓ_1 loss. For small ε, the sample size n is effectively reduced to $n\varepsilon^2/k^2$ (under both losses). When compared to Proposition 1, this result implies that k-RR is not optimal in the high privacy regime. However, for $\varepsilon \approx \ln k$, the sample size n is reduced to $n/4$ (under both losses). This result suggests that, while k-RR is not optimal for small values of ε, it is "order" optimal for ε on the order of $\ln k$. Note that k-RR provides a natural interpretation of this low privacy regime: specifically, setting $\varepsilon = \ln k$ translates to telling the truth with probability $\frac{1}{2}$ and lying uniformly over the remainder of the alphabet with probability $\frac{1}{2}$; an intuitively reasonably notion of plausible deniability.

Performance under k-RAPPOR. Comparing Equation (12) to the observation at the beginning of this subsection, we can see that an extra factor of $\left(1+\frac{k^2e^{\varepsilon/2}}{(k-1)(e^{\varepsilon/2}-1)^2}\right)$ samples is needed to achieve the same ℓ_2^2 as in the non-private case. Similarly, from Equation (13), an extra factor of $\frac{(e^{\varepsilon/2}+k-1)(e^{\varepsilon/2}(k-1)+1)}{(e^{\varepsilon/2}-1)^2(k-1)}$ samples is needed under the ℓ_1 loss. For small ε, n is effectively reduced to $n\varepsilon^2/4k$ (under both losses). When compared to Proposition 1, this result implies that k-RAPPOR is "order" optimal in the high privacy regime. However, for $\varepsilon \approx \ln k$, n is reduced to $n/\sqrt{k}$ (under both losses). This suggests that k-RAPPOR is strictly sub-optimal in the moderate to low privacy regime.

Proposition 5 *For all $\boldsymbol{p} \in \mathbb{S}^k$ and all $\varepsilon \geq \ln(k/2)$,*

$$\mathbb{E}\,||\hat{\boldsymbol{p}}_{KRR} - \boldsymbol{p}||_2^2 \leq \mathbb{E}\,||\hat{\boldsymbol{p}}_{\text{RAPPOR}} - \boldsymbol{p}||_2^2, \qquad (14)$$

where $\hat{\mathbf{p}}_{KRR}$ is the empirical estimate of $\mathbf{p}$ under k-RR, $\hat{\mathbf{p}}_{\text{RAPPOR}}$ is the empirical estimate of $\mathbf{p}$ under k-RAPPOR, and $\hat{\mathbf{p}}$ is the empirical estimator under k-RAPPOR.

Proof See Supplementary Section D. ■

4.4. Simulation Analysis

To complement the theoretical analysis above, we ran simulations of k-RR and k-RAPPOR varying the alphabet size k, the privacy level ε, the number of users n, and the true distribution p from which the samples were drawn. In all cases, we report the mean over 10,000 evaluations of $\|\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_{\text{decoded}}\|_1$ where $\hat{\boldsymbol{p}}$ is the ground truth sample drawn from the true distribution and $\hat{\boldsymbol{p}}_{\text{decoded}}$ is the decoded k-RR or k-RAPPOR distribution. We vary ε over a range that corresponds to the moderate-to-low privacy regimes in our theoretical analysis above, observing that even large values of ε can provide plausible deniability impossible under un-noised logging.

We compare using the ℓ_1 distance of the two distributions because in most applications we want to estimate all values well, emphasizing neither very large values (as an ℓ_2 or higher metric might) nor very small values (as information theoretic metrics might). Supplementary Figures 5 and 6, analogous to the ones in this section, demonstrate that the choice of distance metric does not qualitatively affect our conclusions on the decoding strategies for k-RR or k-RAPPOR nor on the regimes in which each is superior.

The distributions we considered in simulation were binomial distributions with parameter in $\{.1, .2, .3, .4, .5\}$, Zipf distribution with parameter in $\{1, 2, 3, 4, 5\}$, multinomial distributions drawn from a symmetric Dirichlet distribution with parameter $\vec{\mathbf{1}}$, and the geometric distribution with mean $k/5$. The geometric distribution is shown in Supplementary Figure 4. We focus primarily on the geometric distribution here because qualitatively it shows the same patterns for decoding as the full set of binomial and Zipf distributions and it is sufficiently skewed to represent many real-world datasets. It is also the distribution for which k-RAPPOR does the best relative to k-RR over the largest range of k and ε in our simulations.

4.4.1. DECODING

We first consider the impact of the choice of decoding mechanism used for k-RR and k-RAPPOR. We find that the best decoder in practice for both k-RR and k-RAPPOR on skewed distributions is the *projected decoder* which projects the $\hat{\boldsymbol{p}}_{\text{KRR}}$ or $\hat{\boldsymbol{p}}_{\text{RAPPOR}}$ onto the probability simplex $\mathbb{S}^k$ using the method described in Algorithm 1 of (Wang & Carreira-Perpiñán, 2013). For k-RR, we compare the projected empirical decoder to the normalized empirical decoder (which truncates negative values and renormalizes) and to the maximum likelihood decoder (see Supplementary Section F.1). For k-RAPPOR, we compare the standard decoder, normalized decoder, and projected decoder. Figure 1 shows that the projected decoder is substantially better than the other decoders for both k-RR and k-RAPPOR for the whole range of k and ε for the geometric distribution. We find this result holds as we vary the number of users from 30 to 10^6 and for all distributions we evaluated except for the Dirichlet distribution, which is the least skewed. For the Dirichlet distribution, the normalized decoder variant is best for both k-RR and k-RAPPOR. Because the projected decoder is best on all the skewed distributions we expect to see in practice, we use it exclusively for the open-alphabet experiments in Section 5.

4.4.2. k-RR VS k-RAPPOR

To construct a fair, empirical comparison of k-RR and k-RAPPOR, we employ the same methodology used above in selecting decoders. Figure 2 shows the difference between the best k-RR decoder and the best k-RAPPOR decoder (for a particular k and ε). For most cells, the best decoder is the projected decoder described above.

Note that the best k-RAPPOR decoder is consistently better than the best k-RR decoder for relatively large k and low ε. However, k-RR is slightly better than k-RAPPOR in all conditions where $k < e^\varepsilon$ (bottom-right triangle), an empirical result for ℓ_1 that complements Proposition 5's statement about ML decoders in ℓ_2. All of the skewed distributions manifest the same pattern as the geometric distribution. As the number of users increases, k-RR's advantage over k-RAPPOR in the low privacy environment shrinks. In the next sections, we will examine the use of cohorts to improve decoding and to handle larger, open alphabets.

5. Open Alphabets, Hashing, and Cohorts

In practice, the set of values that may need to be collected may not be easily enumerable in advance, preventing a direct application of the binary and k-ary formulations of private distribution estimation. Consider a population of n users, where each user i possesses a symbol s_i drawn from a large set of symbols $\mathcal{S}$ whose membership is not known in advance. This scenario is common in practice; for example, in Chrome's estimation of the distribution of home page settings (Erlingsson et al., 2014). Building on this intuitive example, we assume for the remainder of the paper that symbols s_i are strings, but we note that the methods described are applicable to any hashable structures.

5.1. O-RR: k-RR with hashing and cohorts

k-RR is effective for privatizing over known alphabets. Inspired by (Erlingsson et al., 2014), we extend k-RR to open alphabets by combining two primary intuitions: hashing and cohorts. Let $\textsc{hash}(s)$ be a function mapping $\mathcal{S} \to \mathbb{N}$ with a low collision rate, i.e. $\textsc{hash}(s) = \textsc{hash}(s')$ with very low probability for $s' \neq s$. With hashing, we could use k-RR to guarantee local privacy over an alphabet of size k by having each client report $\boldsymbol{Q}_{\text{KRR}}(\textsc{hash}(s) \mod k)$. However, as we will see, hashing alone is not enough to provide high utility because of the increased rate of collisions introduced by the modulus.

Complementing hashing, we also apply the idea of hash *cohorts*: each user i is assigned to a cohort c_i sampled i.i.d. from the uniform distribution over $\mathcal{C} = \{1, ..., C\}$. Each cohort $c \in \mathcal{C}$ provides an independent view of the underlying distribution of strings by projecting the space of strings $\mathcal{S}$ onto a smaller space of symbols $\mathcal{X}$ using an independent hash function $\textsc{hash}_c$. The users in a cohort use their cohort's hash function to partition $\mathcal{S}$ into k disjoint subsets by computing $x_i = \textsc{hash}_{c_i}(s_i) \mod k = \textsc{hash}_{c_i}^{(k)}(s_i)$. Each subset contains approximately the same number of strings, and because each cohort uses a different hash function, the induced partitions for different cohorts are orthogonal: $\mathbb{P}(x_i = x_j | c_i \neq c_j) \approx \frac{1}{k}$ even when $s_i = s_j$.

5.1.1. ENCODING AND DECODING

For encoding, the O-RR privatization mechanism can be viewed as a sampling distribution independent of $\mathcal{C}$. Therefore, $\boldsymbol{Q}_{\text{ORR}}(y, c|s)$ is given by

$$\frac{1}{C(e^\varepsilon + k - 1)} \begin{cases} e^\varepsilon & \text{if } \textsc{hash}_c^{(k)}(s) = y, \\ 1 & \text{if } \textsc{hash}_c^{(k)}(s) \neq y. \end{cases} \tag{15}$$

For decoding, fix candidate set $\mathcal{S}$ and interpret the privatization mechanism $\boldsymbol{Q}_{\text{ORR}}$ as a $kC \times S$ row-stochastic matrix:

$$\boldsymbol{Q}_{\text{ORR}} = \frac{1}{C} \frac{1}{e^\varepsilon + k - 1} \left(\mathbf{1} + (e^\varepsilon - 1)\boldsymbol{H} \right) \tag{16}$$

where:

$$\boldsymbol{H}(y, c|s) = \mathbb{1}_{\{\textsc{hash}_c^{(k)}(s) = y\}} \tag{17}$$

Note that $\boldsymbol{H}$ is a $kC \times S$ sparse binary matrix encoding the hashed outputs for each cohort, wherein each column of $\boldsymbol{H}$ has exactly C non-zero entries.

Now $\boldsymbol{m} = \boldsymbol{p}\boldsymbol{Q}_{\text{ORR}}$ is the expected output distribution for true probability vector $\boldsymbol{p}$, allowing us to form an empirical estimator by using standard least-squares techniques to solve the linear system:

$$\hat{\boldsymbol{p}}_{\text{ORR}} \boldsymbol{H} = \frac{1}{e^\varepsilon - 1} \left(C(e^\varepsilon + k - 1)\hat{\boldsymbol{m}} - \mathbf{1} \right). \tag{18}$$

Note that when $C = 1$ and $\boldsymbol{H}$ is the identity matrix, (18) reduces to standard k-RR empirical estimator as seen in (6).

As with the k-RR empirical estimator, $\hat{\boldsymbol{p}}_{\text{ORR}}$ may have negative entries. Section 4.1 describes methods for constraining $\hat{\boldsymbol{p}}_{\text{ORR}}$ to $\mathbb{S}^k$, of which simplex projection is demonstrated to offer superior performance in Section 4.4. The remainder of the paper assumes that O-RR uses the simplex projection strategy.

5.2. O-RAPPOR

RAPPOR also extends from k-ary alphabets to open alphabets using hashing and cohorts (Erlingsson et al., 2014); we refer to this extension herein as O-RAPPOR. However, the k-RAPPOR mechanism uses a size $|\tilde{\mathcal{X}}| = 2^k$

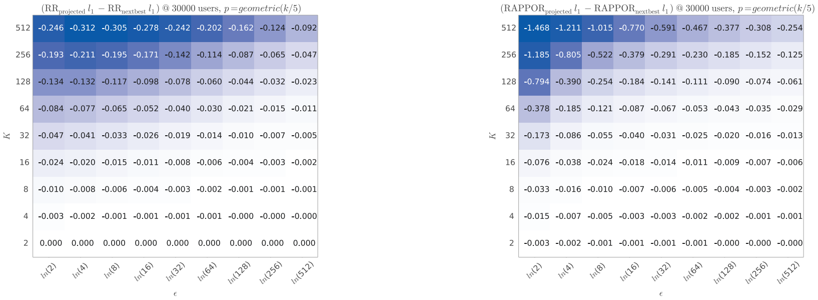


Figure 1: The improvement in ℓ_1 decoding of the projected k-RR decoder (left) and projected k-RAPPOR decoder (right). Each grid varies the size of the alphabet k (rows) and privacy parameter ε (columns). Each cell shows the difference in ℓ_1 magnitude that the projected decoder has over the ML and normalized k-RR decoders (left) or the standard and normalized k-RAPPOR decoders (right). Negative values mean improvement of the projected decoder over the next best alternative.

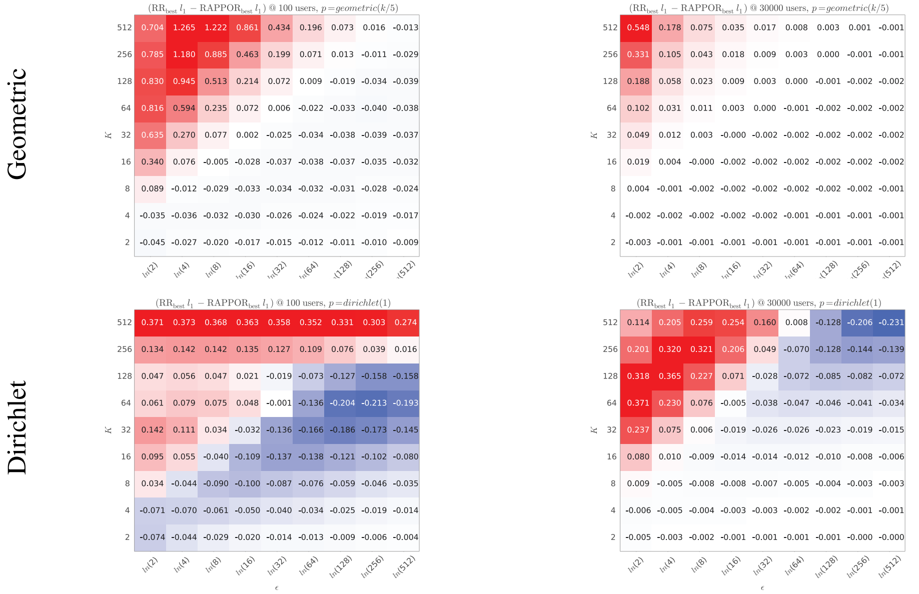
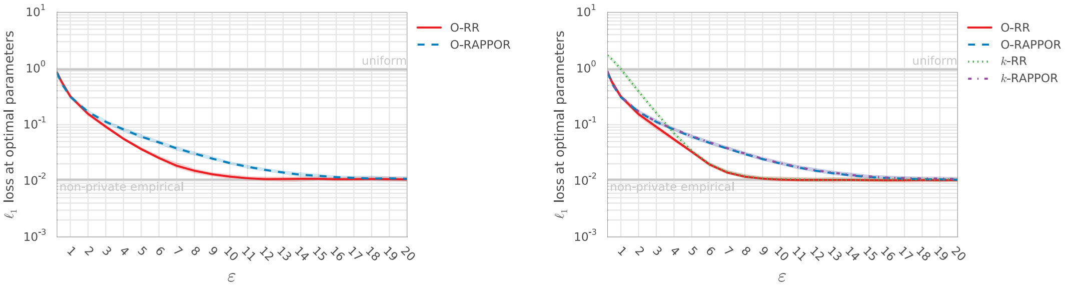


Figure 2: The improvement (negative values, blue) of the best k-RR decoder over the best k-RAPPOR decoder varying the size of the alphabet k (rows) and privacy parameter ε (columns). The left charts focus on small numbers of users (100); the right charts show a large number of users (30000, also representative of larger numbers of users). The top charts show the geometric distribution (skewed) and the bottom charts show the Dirichlet distribution (flat).



(a) Open alphabets.

(b) Closed alphabets.

Figure 3: ℓ_1 loss of O-RR and O-RAPPOR for $n = 10^6$ on the geometric distribution when applied to unknown input alphabets (via hash functions, (a)) and to known input alphabets (via perfect hashing, (b)). Lines show median ℓ_1 loss with 90% confidence intervals over 50 samples. Free parameters are set via grid search over $k \in [2, 4, 8, \ldots, 2048, 4096]$, $c \in [1, 2, 4, \ldots, 512, 1024]$, $h \in [1, 2, 4, 8, 16]$ for each ε. Note that the k-RAPPOR and O-RAPPOR lines in (b) are nearly indistinguishable. Baselines indicate expected loss from (1) using an empirical estimator directly on the input $\boldsymbol{s}$ and (2) using the uniform distribution as the $\hat{\boldsymbol{p}}$ estimate.

input representation as opposed to k-RR's size $|\mathcal{X}| = k$ representation. Taking advantage of the larger input space, O-RAPPOR uses an independent h-hash Bloom filter $\text{BLOOM}_c^{(k)}$ for each cohort before applying the k-RAPPOR mechanism—i.e. the j-th bit of x_i is 1 if $\text{HASH}_{c,h'}^{(k)}(s_i) = j$ for any $h' \in [1 \ldots h]$, where $\text{HASH}_{c,h'}^{(k)}$ are a set of hC mutually independent hash functions modulo k.

Decoding for O-RAPPOR is described in (Erlingsson et al., 2014) and follows a similar strategy as for O-RR. However, because this paper focuses on distribution estimation rather than heavy hitter detection, we eliminate both the Lasso regression stage and filtering of imputed frequencies relative to Bonferroni corrected thresholds, retaining just the regular least-squares regression.

5.3. Simulation Analysis

We ran simulations of O-RR and O-RAPPOR for $n = 10^6$ users with input drawn from an alphabet of $S = 256$ symbols under a geometric distribution with mean=$S/5$ (see Supplementary Figure 4). As described in Section 4.4, the geometric distribution is representative of actual data and relatively easy for k-RAPPOR and challenging for k-RR. Free parameters were set to minimize the median ℓ_1 loss. Similar results for $S = 4096$ and $n = 10^6$ and 10^8 are included in the Supplementary Material.

In Figure 3(a), we see that under these conditions, O-RR matches the utility of O-RAPPOR in both the very low and high privacy regimes and exceeds the utility of O-RAPPOR over midrange privacy settings.

For O-RR, we find that the optimal k depends directly on ε, that increasing C consistently improves performance in the low-to-mid privacy regime, and that $C = 1$ noticably underperforms across the range of privacy levels. For O-RAPPOR, we find that performance improves as k increases (with $k = 4096$ near the asymptotic limit), that $C = 1$ noticably underperforms across the range of privacy values, but with all $C \geq 2$ performing indistinguishably. Finally, we find that the optimal value for h is consistently 1, indicating that Bloom filters provide no utility improvement beyond simple hashing. See Supplementary Figure 11 for details.

5.4. Improved Utility for Closed Alphabets

O-RR and O-RAPPOR extend k-ary mechanisms to open alphabets through the use of hash functions and cohorts. These same mechanisms may also be applied to closed alphabets known *a priori*. While direct application is possible, the reliance on hash functions exposes both mechanism to unnecessary risk of hash collision.

Instead, we modify the O-RR and O-RAPPOR mechanisms, replacing each cohort's generic hash functions with minimal perfect hash functions mapping $\mathcal{S}$ to $[0 \ldots S-1]$ before applying the modulo k operation. In most closed-alphabet applications, $\mathcal{S} = [0 \ldots S-1]$, in which case these minimal perfect hash functions are simply permutations. Also note that in this setting, O-RR and and O-RAPPOR reduce to exactly their k-ary counterparts when C and h are both 1 except that the output symbols are permuted.

In Figure 3(b), we evaluate these modified mechanisms using the same method described in Section 5.3 (note that the utilities of k-RAPPOR and O-RAPPOR are nearly indistinguishable). O-RAPPOR benefits little from the introduction of minimal perfect hash functions. In contrast, O-RR's utility improves significantly, meeting or exceeding the utility of all other mechanisms at all considered ε.

6. Conclusion

Data improves products, services, and our understanding of the world. But its collection comes with risks to the individuals represented in the data as well as to the institutions responsible for the data's stewardship. This paper's focus on distribution estimation under local privacy takes one step toward a world where the benefits of data-driven insights are decoupled from the collection of raw data. Our new theoretical and empirical results show that combining cohort-style hashing with the k-ary extension of the classical randomized response mechanism admits practical, state of the art results for locally private logging.

In many applications, data is collected to enable the making of a specific decision. In such settings, the nature of the decision frequently determines the required level of utility, and the number of reports to be collected n is predetermined by the size of the existing user base. Thus, the differential privacy practitioner's role is often to offer users as much privacy as possible while still extracting sufficient utility at the given n. Our results suggest that O-RR may play a crucial role for such a practitioner, offering a single mechanism that provides maximal privacy at any desired utility level simply by adjusting the mechanism's parameters.

In future work, we plan to examine estimation of non-stationary distributions as they change over time, a common scenario in data logged from user interactions. We will also consider what utility improvements may be possible when some responses need more privacy than others, another common scenario in practice. Much more work remains before we can dispel the collection of un-noised data altogether.

Acknowledgements. Thanks to Úlfar Erlingsson, Ilya Mironov, and Andrey Zhmoginov for their comments on drafts of this paper.

References

Bassily, Raef and Smith, Adam. Local, private, efficient protocols for succinct histograms. *arXiv preprint arXiv:1504.04686*, 2015.

Blocki, Jeremiah, Datta, Anupam, and Bonneau, Joseph. Differentially private password frequency lists. 2016.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Chan, T-H Hubert, Li, Mingfei, Shi, Elaine, and Xu, Wenchang. Differentially private continual monitoring of heavy hitters from distributed streams. In *Privacy Enhancing Technologies*, pp. 140–159. Springer, 2012.

Diakonikolas, Ilias, Hardt, Moritz, and Schmidt, Ludwig. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems*, pp. 2557–2565, 2015.

Duchi, John, Wainwright, Martin J, and Jordan, Michael I. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems*, pp. 1529–1537, 2013a.

Duchi, John C, Jordan, Michael I, and Wainwright, Martin J. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013b.

Dwork, C. Differential privacy. In *Automata, languages and programming*, pp. 1–12. Springer, 2006.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–284. Springer, 2006.

Dwork, Cynthia. Differential privacy: A survey of results. In *Theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Erlingsson, Úlfar, Pihur, Vasyl, and Korolova, Aleksandra. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067. ACM, 2014.

Hsu, Justin, Khanna, Sanjeev, and Roth, Aaron. Distributed private heavy hitters. In *Automata, Languages, and Programming*, pp. 461–472. Springer, 2012.

Kairouz, Peter, Oh, Sewoong, and Viswanath, Pramod. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems*, pp. 2879–2887, 2014.

Kamath, Sudeep, Orlitsky, Alon, Pichapati, Venkatadheeraj, and Suresh, Ananda Theertha. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, pp. 1066–1100, 2015.

Lehmann, Erich Leo and Casella, George. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

McCallum, Andrew, Nigam, Kamal, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pp. 41–48. Citeseer, 1998.

Narayanan, Arvind and Shmatikov, Vitaly. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 111–125. IEEE, 2008.

Wang, Weiran and Carreira-Perpiñán, Miguel Á. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *CoRR*, abs/1309.1541, 2013. URL `http://arxiv.org/abs/1309.1541`.

Warner, Stanley L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.