Impact of Sampling on Locally Differentially Private Data Collection

Sayan Biswas[1,2], Graham Cormode[3], and Carsten Maple[4,5]

[1] *INRIA, France*
[2] *LIX, École Polytechnique, France*
[3] *Dept. of Computer Science, University of Warwick, UK*
[4] *WMG, University of Warwick, UK*
[5] *Alan Turing Institute, UK*
sayan.biswas@inria.fr, g.cormode@warwick.ac.uk, cm@warwick.ac.uk

Abstract

With the recent bloom of data, there is a huge surge in threats against individuals' private information. Various techniques for optimizing privacy-preserving data analysis are at the focus of research in the recent years. In this paper, we analyse the impact of sampling on the utility of the standard techniques of frequency estimation, which is at the core of large-scale data analysis, of the locally deferentially private data-release under a pure protocol. We study the case in a distributed environment of data sharing where the values are reported by various nodes to the central server, e.g., cross-device Federated Learning. We show that if we introduce some random sampling of the nodes in order to reduce the cost of communication, the standard existing estimators fail to remain unbiased. We propose a new unbiased estimator in the context of sampling each node with certain probability and compute various statistical summaries of the data using it. We propose a way of sampling each node with personalized sampling probabilities as a step to further generalisation, which leads to some interesting open questions in the end. We analyse the accuracy of our proposed estimators on synthetic datasets to gather some insight on the trade-off between communication cost, privacy, and utility.

1 Introduction

To address the age-old battle between privacy and utility, various optimisation techniques to analyse the data. There is a massive explosion of data in the recent few years, and with the plethora of data that is being generated everyday, their threats against their privacy is increasing manifold. Hence, the age-old battle between privacy and utility of data has become all the more important catering to the urge to dissect and analyse users' personal data for various kinds of analytics. Differential privacy (DP) [1, 2] have become the standard for privacy protection in the last few years. To efface the need of a central trusted curator, a local variant of DP called the Local Differential Privacy (LDP) [3] has been intensively studied of late. With LDP, users get an opportunity to obfuscate their data locally and this noisy data from the end of the users is reported to the central server. The privacy level can be adjusted according to the requirement of the users by , by adding some tuning the privacy parameter ϵ of the LDP mechanism.

The idea of LDP aligns well with the modern day distributed machine learning, where the idea is to reduce the dependency on a potentially adversarial central server for carrying out the model training. This gave the rise to the concept of Federated Learning (FL) [4], where a local model is trained independently at various nodes and the updates are communicated to the central server to train the main model, aggregating all the local updates. In particular, cross-device Federated Learning [5, 6, 7, 8, 9], the data from the users are used to train a local model on individual devices and the model update is communicated to the central server, making sure that one's personal data never leaves their device. However, in such a setting, often the communication cost is compromised by reporting the updates from every node to estimate the frequency of each value in the domain, and subsequently other statistical summaries of the data, in the central server.

Recently, a substantial focus of FL community focus has been on optimizing sampling techniques [10, 11, 12, 13, 14, 15]. Another branch of recent work has been in the direction of frequency estimation under LDP protocols [14, 16]. In this work, we aimed to incorporate the idea of sampling under LDP protocols and analyze its potential impact on standard frequency estimation techniques.

In this paper, we aim to look at the impact of introducing some sampling techniques on such estimates of LDP data. With millions of users holding data, a useful tool for the service providers is to gather a right number of data points which would be optimal and sufficient for performing various kinds of analysis. In summary, as a main contribution of this work, we point out that the standard estimators fail to stay unbiased when sampling techniques are introduced in a distributed learning framework. Hence, we propose a new unbiased estimator generalising the existing work by Wang et al. in [14]. We analyse the trade-off between the communication cost and the utility and performance of our estimator under a pure LDP protocol through experiments on synthetic datasets. We illustrate, empirically, the usefulness of sampling by showing that sampling a huge number of users does not drastically improve the quality of the analysis performed after a certain point, therefore, implying the necessity of setting appropriate sampling probabilities to optimize the trade-off between communication cost, privacy, and the quality of the estimators, which contribute massively in the analytics.

2 Preliminaries

Definition 2.1 (Differential privacy [1, 2])**.** For a certain query, a randomizing mechanism $\mathcal{R}$ provides *ϵ-differential privacy (DP)* if, for all neighbouring[1] datasets, D and D', and all $S \subseteq \text{Range}(\mathcal{R})$, we have

$$\mathbb{P}[\mathcal{R}(D) \in S] \leq e^{\epsilon}\, \mathbb{P}[\mathcal{R}(D') \in S]$$

Definition 2.2 (Local differential privacy [3])**.** Let $\mathcal{X}$ and $\mathcal{Y}$ denote the spaces of original and noisy data, respectively. A randomizing mechanism $\mathcal{R}$ provides *ϵ-local differential privacy (LDP)* if, for all x, $x' \in \mathcal{X}$, and all $y \in \mathcal{Y}$, we have

$$\mathbb{P}[\mathcal{R}(x) = y] \leq e^{\epsilon}\, \mathbb{P}\left[\mathcal{R}(x') = y\right]$$

Definition 2.3 (Pure LDP protocols)**.** [14] A LDP mechanism $\mathcal{R}$ is *pure* iff there exist $p^* > q^*$ such that for all v_1 and $v_2 \neq v_1$:

$$\begin{aligned} \mathcal{P}\left[\mathcal{R}(v_1) \in \{y\colon v_1 \in \text{Support}(y)\}\right] = p^*, \text{ and} \\ \mathbb{P}\left[\mathcal{R}(v_2) \in \{y\colon v_1 \in \text{Support}(y)\}\right] = q^* \end{aligned} \tag{1}$$

where for any input $x \in \mathcal{X}$, the set $\{y \in \mathcal{Y}\colon x \in \text{Support}(y)\}$ is the set of outputs in $\mathcal{Y}$ that *support* the input x with a non-zero probability of being observed via the mechanism $\mathcal{R}$, i.e., $\{y\colon x \in \text{Support}(y)\} = \{y\colon \mathbb{P}\left[\mathcal{R}(x) = y\right] \neq 0\}$.

Definition 2.4 (Direct Encoding [17])**.** Let $\mathcal{X}$ be a discrete domain of size d. Then *direct encoding* (DE) a.k.a. *k-randomized response* (k-RR) mechanism, $\mathcal{R}_{\text{DE}}$, is a locally differentially private mechanism that stochastically maps the domain $\mathcal{X}$ onto itself (i.e., $\mathcal{Y} = \mathcal{X}$), given by

$$\mathcal{R}_{\text{DE}}(y|x) = \begin{cases} p = c\, e^{\epsilon} & \text{, if } x = y \\ q = c, & \text{, otherwise} \end{cases}$$

for any x, $y \in \mathcal{X}$, where $c = \frac{1}{e^{\epsilon}+d-1}$.

In this work we focus in the setting of DE where it perturbs and fix a discrete domain $\mathcal{X}$ of size m for our analysis, supposing DE perturbs values from $\mathcal{X}$ and to some noisy values in $\mathcal{X}$. Let there are $n \in \mathbb{N}$ nodes, each holding some value from $\mathcal{X}$ obfuscated by DE. Let the Support function for DE be $\text{Support}_{\text{DE}}(i) = \{i\}$, i.e., each obfuscated output value $i \in \mathcal{X}$ supports the input $i \in \mathcal{X}$.

Remark 1. Setting $\text{Support}(i) = \{i\}$, $p^* = p$, and $q^* = q$, DE becomes is pure LDP protocol, shown by Wang et al. in [14].

Wang et al. [14] proposed an unbiased frequency estimator, $c_{\text{DE}}(i)$, of the original value i going through a pure LDP protocol as:

$$c(i) = \frac{\sum\limits_{j} \mathbb{1}_{\text{Support}(y^j)}(i) - nq^*}{p^* - q^*} \tag{2}$$

where y^j denotes the noisy value reported by the j^{th} node. Thus, using (2) in the context of DE, for any value $i \in \mathcal{X}$ we obtain:

$$c_{\text{DE}}(i) = \frac{\sum\limits_{j=1}^{n} \mathbb{1}_{\{X_j = i\}} - nq}{p - q} \tag{3}$$

where $\mathbb{1}_E$ is the indicator function for any event E such that

$$\mathbb{1}_E = \begin{cases} 1 & \text{if } E \text{ happens} \\ 0, & \text{otherwise} \end{cases}$$

We explore this idea to investigate the behaviour of c_{DE} if each node is independently sampled to report its value, perturbed with DE, with some probability π. Let S be the random variable representing the number of nodes which have been reported to the central server. Hence $\mathbb{P}(S > n) = \mathbb{P}(S < 0) = 0$. Taking the same estimator $c_{\text{DE}}(i)$ in the setup of random sampling of each node with an independent probability of π, we get:

$$\begin{aligned} \mathbb{E}\left(c_{\text{DE}}(i)\right) &= \mathbb{E}\left(\frac{\sum\limits_{j=1}^{S} \mathbb{1}_{\text{Support}(y^j)}(i) - nq^*}{p^* - q^*}\right) \\ &= \mathbb{E}\left(\frac{\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j = i\}} - nq}{p - q}\right) \\ &= \frac{\mathbb{E}(S)\mathbb{E}(\mathbb{1}_{\{X_j=i\}} - nq)}{p - q} \text{ [Wald's Equation [18]]} \\ &= \frac{n\pi(f_i p + (1 - f_i)q) - nq}{p - q} \\ &= nf_i\pi - \frac{nq(1-\pi)}{p - q} \end{aligned} \tag{4}$$

We see that putting $\pi = 1$ in (4), implying every node is sampled in each round, gives us the same result as in [14].

Theorem 2.1. If we introduce some sampling probability $\pi < 1$ for each node, c_{DE} becomes a biased frequency estimator.

Proof. We recall that $1 \geq p \geq q \geq 0$ by the definition of pure LDP protocols, and $0 \leq \pi \leq 1$. Therefore, $\frac{nq(1-\pi)}{p-q} \geq 0$, and hence, $\mathbb{E}(c_{\text{DE}}(i)) \leq nf_i\pi \leq nf_i$ and equality is attained iff $\pi = 1$. □

3 Unbiased frequency estimation

Motivated from Theorem 2.1, we proceed to device an unbiased estimator for DE, g_{DE}, incorporating the random sampling aspect, defined as follows:

$$g_{\text{DE}}(i) = \frac{c_{\text{DE}}(i)}{\pi} + \frac{nq(1-\pi)}{(p-q)\pi} \tag{5}$$

Theorem 3.1. If each node has an independent sampling probability of π, g_{DE} is an unbiased estimator of the frequencies of the values in $\mathcal{X}$ observed under DE.

Proof. Immediate from (4) in Theorem 2.1 and using the linearity of expectation. □

[1] differing in exactly one place

For the simplicity of notation, let f_i be the random variable representing the fraction of times the value $i \in \mathcal{X}$ is reported to the central server. In [14] Wang et al. define the *approximate variance* of any random variable which is a function of f_i, say $RV(f_i)$, as $\mathrm{Var}^*(RV(f_i)) = \lim\limits_{f_i \to 0} \mathrm{Var}(RV(f_i))$.

Theorem 3.2. In the event of independently sampling the nodes with some probability π, the approximate variance of c_{DE} is given by:

$$\mathrm{Var}^*(g_{\mathrm{DE}}(i)) = \frac{\mathrm{Var}^*(c_{\mathrm{DE}}(i))}{\pi^2} = \frac{n(q - q^2\pi)}{(p-q)^2\pi}$$

Proof. We start by deriving the actual variance of g_{DE}.

$$\mathrm{Var}\left(c_{\mathrm{DE}}(i)\right) = \mathrm{Var}\left(\frac{\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq}{p-q}\right)$$

[S is the r.v. representing the number of nodes sampled]

$$= \frac{\mathrm{Var}\left(\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq\right)}{(p-q)^2}$$

$$= \frac{\mathrm{Var}\left(\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j=i\}}\right)}{(p-q)^2}$$

$$= \frac{\mathbb{E}(S)\,\mathrm{Var}(\mathbb{1}_{\{X_1=i\}}) + \mathbb{E}((\mathbb{1}_{\{X_1=i\}})^2\,\mathrm{Var}(S)}{(p-q)^2}$$

[Random sums of RVs [18]]

$$= \frac{n\pi f_i p(1-p) + n\pi(1-f_i)q(1-q)}{(p-q)^2} + \frac{(f_i p + (1-f_i)q)^2(n\pi(1-\pi))}{(p-q)^2}$$

$$\therefore \mathrm{Var}^*(c_{\mathrm{DE}}(i)) = n\pi\frac{q(1-q) + q^2(1-\pi)}{(p-q)^2} = n\pi\frac{q - q^2\pi}{(p-q)^2}$$

Now we observe that $\mathrm{Var}(g_{\mathrm{DE}}(i)) = \frac{\mathrm{Var}(c_{\mathrm{DE}}(i))}{\pi^2}$, by definition of g_{DE}. Therefore,

$$\mathrm{Var}^*(g_{\mathrm{DE}}(i)) = \frac{\mathrm{Var}^*(c_{\mathrm{DE}}(i))}{\pi^2} = \frac{n(q - q^2\pi)}{(p-q)^2\pi}$$

□

Observe $\mathrm{Var}^*(g_{\mathrm{DE}}(i)) \geq \mathrm{Var}^*\, c_{\mathrm{DE}}(i))$, with equality iff $\pi = 1$, as we would expect since we are introducing more randomness and less information in $g_{\mathrm{DE}}(i)$ compared to $c_{\mathrm{DE}}(i)$ by engendering random sampling of each node.

Definition 3.1 (Normalized variance)**.** The *normalised variance* of any random variable X is defined as

$$\mathrm{Var}_{\mathrm{norm}}(X) = \frac{\mathrm{Var}(X)}{\mathbb{E}(X)}$$

Normalized variance can be useful when comparing two random variables with different means, in order to account for larger variance for larger means.

Theorem 3.3. $\mathrm{Var}^*_{\mathrm{norm}}(g_{\mathrm{DE}}(i)) = \mathcal{O}\left(\frac{1}{\pi^3 n}\right)$

Proof.

$$\mathrm{Var}_{\mathrm{norm}}\left(g_{\mathrm{DE}}(i)\right) = \mathrm{Var}\left(\frac{g_{\mathrm{DE}}(i)}{\mathbb{E}(S)}\right)$$

$$= \mathrm{Var}\left(\frac{g_{\mathrm{DE}}(i)}{n\pi}\right) = \frac{\mathrm{Var}(g_{\mathrm{DE}}(i))}{n^2\pi^2}$$

$$\implies \mathrm{Var}^*\left(\frac{g_{\mathrm{DE}}(i)}{n\pi}\right) = \frac{\mathrm{Var}^*(g_{\mathrm{DE}}(i))}{n^2\pi^2}$$

$$= \frac{n(q - q^2\pi)}{(p-q)^2 n^2 \pi^3}\ [\text{Th.3.2}] = \frac{q - q^2\pi}{(p-q)^2\pi^3 n} = \mathcal{O}\left(\frac{1}{\pi^3 n}\right)$$

□

We note that for small value of π, the normalized variance of the estimator g_{DE} would blow up as it is of the order $\frac{1}{\pi^3 n}$. But this is not unexpected, as with a low sampling probability, it is more likely that we would give rise to fewer nodes that are actually sampled to report their values, giving rise to less information for the central server, which should result in a greater variance. We acknowledge a trade-off between the bias of an estimator and its increasing variance. In particular, we see that without compensating for the bias of c_{DE} to obtain g_{DE} by scaling it with $\frac{1}{\pi}$ and adding up $\frac{nq(1-\pi)}{\pi(p-q)}$, for a small sampling probability π, we would have the bias which will grow up to be a tremendously low a quantity, always giving a massively conservative and negative estimate for the value of i as observed, especially if the number of nodes involved (n) is huge (e.g. in millions), which is often the case in federated learning. Precisely, observe from (4) that as $\lim\limits_{\pi \to 0} c_{\mathrm{DE}} = \frac{nq}{p-q}$, implying that we would be getting a constant and negative estimate for every $i \in \mathcal{X}$, which would make the analysis involving the frequencies rather absurd.

Now we look to improve upon the proposed unbiased frequency estimator g_{DE}. Let S be the random variable representing the number of nodes sampled in a round if each node is independently sampled with probablity π. We proceed to define an improved frequency estimator of the elements of $\mathcal{X}$ under DE through a very natural approach of replacing n by S in the definition of c_{DE}.

Let $\hat{c}_{\mathrm{DE}}(i) = \frac{\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - Sq}{\pi(p-q)}$. In order to use $\hat{c}_{DE}$ as the frequency estimator for any element $i \in \mathcal{X}$, it is crucial to probe if it has any bias.

Theorem 3.4. $\hat{c}_{\mathrm{DE}}$ is an unbiased estimator of the frequencies of the elements of $\mathcal{X}$ being perturbed via DE which are reported by the nodes which are sampled independently.

Proof.

$$\mathbb{E}\left(\frac{\sum\limits_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - Sq}{\pi(p-q)}\right)$$

$$= \frac{\mathbb{E}(S)\mathbb{E}(\mathbb{1}_{\{X_j=i\}}) - \mathbb{E}(Sq)}{\pi(p-q)}\ [\text{Wald's Equation [18]}]$$

$$= \frac{n\pi(f_i p + (1-f_i)q) - n\pi q}{\pi(p-q)} = nf_i$$

□

Theorem 3.5. $\mathrm{Var}(\hat{c}_{\mathrm{DE}}(i)) \geq \mathrm{Var}(g_{\mathrm{DE}}(i))$, i.e., g_{DE} gives a better (more confident) estimate for the frequencies than $\hat{c}_{\mathrm{DE}}$, which is a naive and immediate extension from c_{DE}.

Proof.

$$\begin{aligned} &\mathrm{Var}\left(\hat{c}_{\mathrm{DE}}(i)\right) \\ &= \mathrm{Var}\left(\frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - Sq}{\pi(p-q)}\right) \\ &= \frac{\mathrm{Var}\left(\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}}\right) + \mathrm{Var}(S)q^2}{\pi^2(p-q)^2} \\ &= \mathrm{Var}(g_{\mathrm{DE}}(i)) + \frac{\mathrm{Var}(S)q^2}{\pi^2(p-q)^2} \end{aligned} \tag{6}$$

It follows immediately that $\mathrm{Var}(g_{\mathrm{DE}}(i)) + \frac{\mathrm{Var}(S)q^2}{\pi^2(p-q)^2} \geq \mathrm{Var}(g_{\mathrm{DE}}(i)$ as $\frac{\mathrm{Var}(S)q^2}{\pi^2(p-q)^2} \geq 0$. □

Theorem 3.6. For every $i \in \mathcal{X}$, we have $0 \leq g_{\mathrm{DE}}(i) \leq n$ iff $0 \leq c_{\mathrm{DE}}(i) \leq n$ on an average, i.e., ensuring our proposed frequency estimate evaluating a reasonable frequency for any $i \in \mathcal{X}$ is equivalent to that of the estimate proposed by Wang et al.

Proof. We proceed to show this in two parts:

(i) $0 \leq c_{\mathrm{DE}}(i) \Leftrightarrow 0 \leq g_{\mathrm{DE}}(i)$ on an average

(ii) $n \geq c_{\mathrm{DE}}(i) \Leftrightarrow n \geq g_{\mathrm{DE}}(i)$ on an average

Proceeding with (i), we obtain:

$$\begin{aligned} c_{\mathrm{DE}}(i) \geq 0 &\Leftrightarrow \frac{\sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} - nq}{p-q} \geq 0 \\ &\Leftrightarrow \sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} - nq \geq 0 \text{ [} p \geq q \text{ for pure LDP]} \\ &\Leftrightarrow \sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} \geq nq \Leftrightarrow \mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}}\right) \geq nq \\ &\Leftrightarrow n(f_i p + (1-f_i)q) \geq nq \Leftrightarrow p \geq q \end{aligned} \tag{7}$$

That's the trivial condition assumed to make DE a pure LDP protocol.

Now focussing on g_{DE}, we get:

$$\begin{aligned} g_{\mathrm{DE}}(i) \geq 0 &\Leftrightarrow \frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq}{\pi(p-q)} + \frac{nq(1-\pi)}{(p-q)\pi} \geq 0 \\ &\Leftrightarrow \sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq\pi \geq 0 \text{ [} p \geq q \text{ for pure LDP]} \\ &\Leftrightarrow \sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} \geq nq\pi \end{aligned} \tag{8}$$

Taking the expectation of both sides:

$$\begin{aligned} &\Leftrightarrow \mathbb{E}\left(\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}}\right) \geq nq\pi \\ &\Leftrightarrow n\pi(f_i p + (1-f_i)q) \geq nq\pi \\ &\Leftrightarrow p \geq q \end{aligned} \tag{9}$$

Establishing (i), now we shift to prove (ii):

$$\begin{aligned} c_{\mathrm{DE}}(i) \leq n &\Leftrightarrow \frac{\sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} - nq}{p-q} \leq n \\ &\Leftrightarrow \sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} - nq \leq n(p-q) \\ &\Leftrightarrow \sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}} \leq np \\ &\Leftrightarrow \mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}}\right) \leq np \\ &\Leftrightarrow n(f_i p + (1-f_i)q) \leq np \\ &\Leftrightarrow q \leq p \end{aligned} \tag{10}$$

But $q \leq p$ is the trivial condition assumed to make direct encoding a pure LDP protocol.

$$\begin{aligned} g_{\mathrm{DE}}(i) \leq n &\Leftrightarrow \frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq}{\pi(p-q)} + \frac{nq(1-\pi)}{(p-q)\pi} \leq n \\ &\Leftrightarrow \sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} - nq\pi \leq n(p-q)\pi \text{ [} p > q \text{ for pure]} \\ &\Leftrightarrow \sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}} \leq np\pi \\ &\Leftrightarrow \mathbb{E}\left(\sum_{j=1}^{S} \mathbb{1}_{\{X_j=i\}}\right) \leq np\pi \\ &\Leftrightarrow n\pi(f_i p + (1-f_i)q) \leq np\pi \Leftrightarrow q \leq p \end{aligned} \tag{11}$$

□

4 Experimental results

We performed experiments on synthetic datasets to evaluate and visualize the performance of our estimator and observed that, indeed, as we increase π, the estimation by $g_{\mathrm{DE}}(i)$ approximates the original distribution better. We considered 50,000 data points sampled from a domain $\mathcal{X}$ of size 100, following the distributions Binomial$(100, 0.5)$ and Binomial$(50, 0.6)$+Binomial$(50, 0.4)$. We considered two extremes of the sampling probabilities for each node by setting $\pi = 0.1$ and $\pi = 0.9$. Figures 1 & 2 illustrate the performance of our estimator in these two settings for the two different datasets.

We computed the total variation (TV) distance between the original distribution and $\hat{c}_{\mathrm{DE}}(i)$ for the synthetically generated dataset sampled from a $Bin(100, 0.5)$ distribution and illustrated the results in Figure 3, along with communication cost for sampling probabilities ranging from $\pi = 0.1$ to $\pi = 0.9$. We can see a clear trade-off between the communication cost and the TV distance.

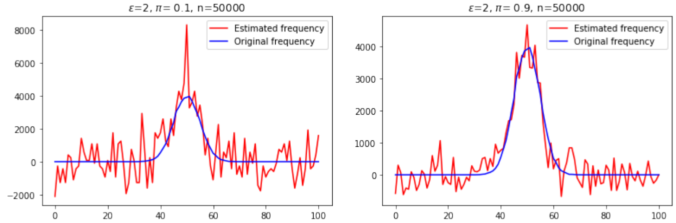

Figure 1: Data sampled from $Bin(100, 0.5)$

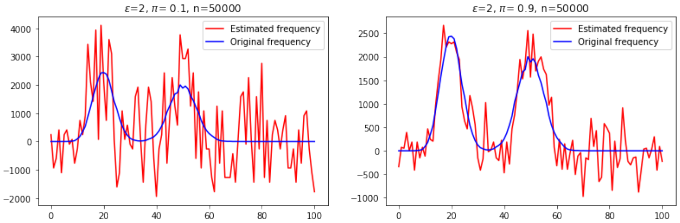

Figure 2: Data sampled from $Bin(50, 0.6) + Bin(50, 0.4)$

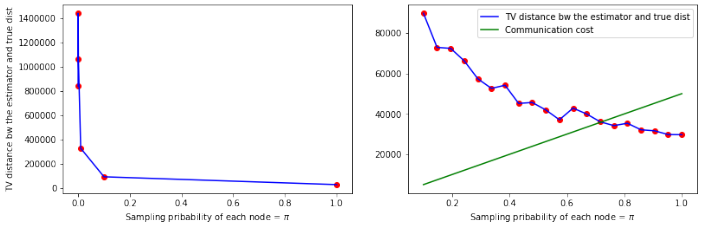

Figure 3: Total variation distance between our proposed estimator and distribution of the original data, and communication cost $= \mathcal{O}(n\pi)$ varying with different sampling probabilities

5 Generalized sampling probabilities

In all the previous results, we assumed that the values from each node is sampled independently with the same probability π. Now we enable us with the flexibility not to require the sampling probability of each node to be the same, opening doors to a lot of interesting paths of research ahead. We explore the setting where the j^{th} node is sampled independently with probability π_j for every node $j \in \{1, 2, \ldots, n\}$. Note that if we have $\pi = \pi_1 = \pi_2 = \ldots \pi_n$, we are left with the sampling environment that we addressed previously.

Let S be the random variable representing the total number of nodes sampled under this flexible setting of having personalized sampling probabilities. We proceed to derive an unbiased frequency estimator in such a generalized case.

Theorem 5.1. Let $h(i) = \frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j = i\}} - nq}{p-q}$, where X_j is the random variable denoting the value reported by the j^{th} node. Then, setting $\mathcal{T}(i)$ as $\frac{nh(i)}{\sum_{j=1}^{n} \pi_j} + \frac{nq\left(n - \sum_{j=1}^{n} \pi_j\right)}{\sum_{j=1}^{n} \pi_j (p-q)}$, it becomes an unbiased frequency estimator of every value $i \in \mathcal{X}$ with

$$\mathrm{Var}^*(\mathcal{T}(i)) = \frac{n^2 \sum_{j=1}^{n} (q\pi_j(1 - q\pi_j))}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2}$$

Remark 2. Putting $\pi = \pi_1 \ldots = \pi_n$ reduces $\mathrm{Var}^*(\mathcal{T}(i))$ to $\mathrm{Var}^*(g_{\mathrm{DE}}(i))$ and further, putting $\pi_1 = \ldots \pi_n = 1$ reduces $\mathrm{Var}^*(\mathcal{T}(i))$ to $\mathrm{Var}^*(c_{\mathrm{DE}}(i))$ as in [14], as expected.

Proof. First we aim to show that $\mathcal{T}(i)$ is an unbiased estimator for any $i \in \mathcal{X}$.

$$\mathbb{E}(h(i)) = \mathbb{E}\left(\frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j = i\}} - nq}{p-q}\right)$$

$$= \frac{\mathbb{E}\left(\sum_{j=1}^{n} \mathbb{1}_{\{X_j \text{ is sampled}\}} \mathbb{1}_{\{X_j = i\}} - nq\right)}{p-q}$$

[As sampling & privatization are independent]

$$= \frac{\sum_{j=1}^{n} \mathbb{E}\left(\mathbb{1}_{\{X_j \text{ is sampled}\}}\right) \mathbb{E}\left(\mathbb{1}_{\{X_j = i\}}\right) - nq}{p-q}$$

$$= \frac{\sum_{j=1}^{n} \mathbb{P}\left(\mathbb{1}_{\{X_j \text{ is sampled}\}}\right) \mathbb{P}\left(\mathbb{1}_{\{X_j = i\}}\right) - nq}{p-q}$$

$$= \frac{\sum_{j=1}^{n} \pi_j (f_i p + (1 - f_i) q) - nq}{p-q} = \sum_{j=1}^{n} \pi_j f_i - \frac{q(n - \sum_{j=1}^{n} \pi_j)}{p-q}$$

Note that $\mathbb{E}(S) = \mathbb{E}\left(\mathbb{1}_{\{x_j \text{ is sampled}\}}\right) = \sum_{j=1}^{n} \pi_j \leq n$. Therefore, $\frac{q(n - \sum_{j=1}^{n} \pi_j)}{p-q} \geq 0$. Hence, we define $\mathcal{T}(i) = \frac{nh(i)}{\sum_{j=1}^{n} \pi_j} + \frac{nq(n - \sum_{j=1}^{n} \pi_j)}{\sum_{j=1}^{n} \pi_j (p-q)}$ as the frequency estimate of the true value i. Because of linearity of expectation, we get $\mathbb{E}(\mathcal{T}(i)) = nf_i$, giving us an unbiased estimator for the general case where each node can have a different probability of being sampled. Putting $\pi = \pi_1 = \pi_2 = \ldots \pi_n$ reduces $\mathcal{T}(i)$ to $g_{\mathrm{DE}}(i)$ which is what we would expect.

Now we focus on computing $\mathrm{Var}^*\left(\mathcal{T})(i)\right)$ by first evaluating the actual variance of $\mathcal{T}(i)$. We obtain:

$$\mathrm{Var}(\mathcal{T}(i)) = \mathrm{Var}\left(\frac{nh(i)}{\sum_{j=1}^{n} \pi_j} + \frac{q\left(n - \sum_{j=1}^{n} \pi_j\right)}{\sum_{j=1}^{n} \pi_j (p-q)}\right) = \frac{n^2 \,\mathrm{Var}(h(i))}{(\sum_{j=1}^{n} \pi_j)^2}$$

$$
= \frac{n^2}{(\sum_{j=1}^{n} \pi_j)^2} \operatorname{Var}\left(\frac{\sum_{j=1}^{S} \mathbb{1}_{\{X_j = i\}} - nq}{p-q} \right) = \frac{n^2 \operatorname{Var}\left(\sum_{j=1}^{S} \mathbb{1}_{\{X_j = i\}} \right)}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2}
$$

$$
= \frac{n^2 \operatorname{Var}\left(\sum_{j=1}^{n} \mathbb{1}_{\{X_j \text{ is sampled}\}} \mathbb{1}_{\{X_j = i\}} \right)}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2}
$$

$$
\begin{aligned}
= \frac{n^2}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2} \Bigg(\sum_{j=1}^{n} (\operatorname{Var}\left(\mathbb{1}_{\{X_j \text{ is sampled}\}}\right) \operatorname{Var}\left(\mathbb{1}_{\{X_j = i\}}\right) \\
+ \operatorname{Var}\left(\mathbb{1}_{\{X_j \text{ is sampled}\}}\right) \mathbb{E}\left(\mathbb{1}_{\{X_j = i\}}\right)^2 \\
+ \mathbb{E}\left(\mathbb{1}_{\{X_j \text{ is sampled}\}}\right)^2 \operatorname{Var}\left(\mathbb{1}_{\{X_j = i\}}\right) \Bigg)
\end{aligned}
$$

$$
= \frac{n^2 \sum_{j=1}^{n} (\pi(1-\pi)((f_i p(1-p) + (1-f_i)q(1-q))}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2}
$$

$$
+ \frac{(f_i p + (1-f_i)q)^2)}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2} + \pi^2 (f_i p(1-p) + (1-f_i)q(1-q)))
$$

$$
\implies \operatorname{Var}^*(\mathcal{T}(i)) = \frac{n^2 \sum_{j=1}^{n} (q\pi_j(1-q\pi_j))}{(\sum_{j=1}^{n} \pi_j)^2 (p-q)^2}
$$

□

6 Conclusion and way forward

Sampling of nodes and its impact on accuracy of the trained models, statistical analysis of the data, and aspects of privacy have been at the epicentre of research in the areas of federated learning. The results in this paper enable us to have an unbiased estimate for the frequency of elements of a domain of values which are held by the users. We also get an insight on how the sampling affects the utility of the estimators and the accuracy of estimating the true distribution of the data.

In Figure 3, we observe that after a point, the TV distance doesn't decrease significantly compared to how much the communication cost increases, raising some interesting open questions: Should we go on till sampling every single node? Where should we stop? In fact, the first plot of Figure 3 shows sampling each node with probability 0.1 and sampling every single node do not engender a drastic difference in the TV distance. In particular, we would like to highlight some interesting open questions leading on from this work:

i) *Uniform sampling*: As we proposed an unbiased frequency estimator of the values which are sampled from the users with any arbitrary probability distribution, it would be an interesting area of analysis to first get an initial idea of the sampling distribution in the first round using $\mathcal{T}$, and then use that to our advantage to revise the sampling probabilities of each value inversely proportional to their frequencies so that we can ensure that a sample of the dataset we wish to derive doesn't over-represent a certain value and under-represent some others. In the context of FL, this can be a key area for ensuring a fair model which is not heavily influenced by the mode of the data, making the model more biased towards the majorities, which might not be the desirable outcome for certain tasks, e.g., facial recognition, text prediction, etc. It would be a challenging area to investigate how such a mechanism would perform in the aspect of the communication cost vs utility trade-off against the state-of-the-art differentially private FL techniques [10], especially for high dimensional data.

ii) *Shuffling*: Privacy amplification methods have been recently studied a lot involving the shuffle model. If we look to apply shuffling to the LDP data using DE as the local randomizer, that should mean we should have a high level of central differential privacy guarantee using a lower intensity of local noise using the recent advancements and studies for deriving the amplified formal central differential privacy guarantees using shuffling [19, 20, 21, 22]. As the estimators we proposed, both c_{DE} and $\mathcal{T}$, are functions function of the underlying LDP mechanism used – in particular, the obfuscating probability distribution which is dependant on ϵ – it is obvious that a higher value of ϵ will engender a better bound. The introduction of shuffling would guarantee that the privacy of the users would not be compromised, as we can tune the final level of central DP guarantee quite high for even a high value of the privacy parameter of DE, which is the local randomizer used in this process.

   Thus, it would be an interesting comparison to have between variance bounds of the estimated frequencies of the shuffle model with DE using our proposed estimates, and the variance of the observed data under the central Gaussian mechanism, which is essentially the maximum likelihood estimate of the original distribution of the data, under the same level of the privacy parameters. Depending on the behaviour, we could hypothesize on the requirement of the number of samples and the sampling probabilities that would ensure a tighter variance for our proposed estimates.

iii) *Personalised sampling*: Another very interesting direction this work leads on to is to see if techniques like the Lagrange multipliers could be used to find the optimal sampling distribution $(\pi_1, \ldots, \pi_n)$ that would minimize the variance of the estimator that we derived under the constraint that $(\pi_1, \ldots, \pi_n)$ is a probability distribution. In other words, we would like to focus on the optimization problem where we wish to $\operatorname{Var}^*(\mathcal{T})(i)$ for every value $i \in \mathcal{X}$ such that $0 \leq \pi_j \leq 1$ for every $j \in \{1, \ldots, n\}$ and $\sum_{j=1}^{n}$. The problem would be straightforward if we wished to minimize $\operatorname{Var}^*(\mathcal{T})(i)$ for some fixed i, but becomes increasingly challenging when we are dealing with minimizing all the variances at an the minimum, under some multi-dimensional metric, giving us the optimal $(\pi_1, \ldots, \pi_n)$. This approach would enable us to find the optimal sampling probability that would give the minimum variance for our proposed unbiased estimators.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[2] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006*, S. Vaudenay, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 486–503.

[3] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 2013, pp. 429–438.

[4] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *CoRR*, vol. abs/1602.05629, 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[5] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," 2019.

[6] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," 2018.

[7] M. Chen, R. Mathews, T. Ouyang, and F. Beaufays, "Federated learning of out-of-vocabulary words," 2019.

[8] S. Ramaswamy, R. Mathews, K. Rao, and F. Beaufays, "Federated learning for emoji prediction in a mobile keyboard," 2019.

[9] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," 2019.

[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318

[11] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-private streaming algorithms." in *ics*, 2010, pp. 66–80.

[12] E. Rizk, S. Vlaski, and A. H. Sayed, "Federated learning under importance sampling," 2020.

[13] ——, "Optimal importance sampling for federated learning," 2020. [Online]. Available: https://arxiv.org/abs/2010.13600

[14] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 729–745. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao

[15] L. Cai, D. Lin, J. Zhang, and S. Yu, "Dynamic sample selection for federated learning with heterogeneous data in fog computing," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[16] G. Cormode, S. Maddock, and C. Maple, "Frequency estimation under local differential privacy," *Proc. VLDB Endow.*, vol. 14, no. 11, p. 2046–2058, jul 2021. [Online]. Available: https://doi.org/10.14778/3476249.3476261

[17] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2436–2444.

[18] A. Wald, "Small summaries for big data," https://en.wikipedia.org/wiki/Wald\%27s_equation, accessed: 12-Jul-2020.

[19] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Annual International Cryptology Conference*. Springer, 2019, pp. 638–667.

[20] V. Feldman, A. McMillan, and K. Talwar, "Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling," *arXiv preprint arXiv:2012.12803*, 2020.

[21] A. Koskela, M. A. Heikkilä, and A. Honkela, "Tight accounting in the shuffle model of differential privacy," *arXiv preprint arXiv:2106.00477*, 2021.

[22] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 2468–2479.