SPEED: secure, PrivatE, and efficient deep learning

Arnaud Grivet Sébert, Rafaël Pinot, Martin Zuber, Cedric Gouy-Pailler, Renaud Sirdey

Springer Machine Learning manuscript No.
(will be inserted by the editor)

SPEED: Secure, PrivatE, and Efficient Deep learning

Arnaud Grivet Sébert · Rafaël Pinot · Martin Zuber · Cédric Gouy-Pailler · Renaud Sirdey

Received: date / Accepted: date

Abstract We introduce a deep learning framework able to deal with strong privacy constraints. Based on collaborative learning, differential privacy and homomorphic encryption, the proposed approach advances state-of-the-art of private deep learning against a wider range of threats, in particular the honest-but-curious server assumption. We address threats from both the aggregation server, the global model and potentially colluding data holders. Building upon distributed differential privacy and a homomorphic argmax operator, our method is specifically designed to maintain low communication loads and efficiency. The proposed method is supported by carefully crafted theoretical results. We provide differential privacy guarantees from the point of view of any entity having access to the final model, including colluding data holders, as a function of the ratio of data holders who kept their noise secret. This makes our method practical to real-life scenarios where data holders do not trust any third party to process their datasets nor the other data holders. Crucially the computational burden of the approach is maintained reasonable, and, to the best of our knowledge, our framework is the first one to be efficient enough to investigate deep learning applications while addressing such a large scope of threats. To assess the practical usability of our framework, experiments have been carried out on image datasets in a classification context. We present numerical results that show that the learning procedure is both accurate and private.

Keywords Data protection · Collaborative learning · Distributed differential privacy · Homomorphic encryption

Contact author: arnaud.grivetsebert@cea.fr

Arnaud Grivet Sébert, Rafaël Pinot, Martin Zuber, Cédric Gouy-Pailler, Renaud Sirdey
Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France

Rafaël Pinot
Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France

1 Introduction

Application scenarios. We consider n hospitals, each of which owns a (personal) labelled database composed of medical records from its patients and a model (e.g. neural network) trained on this database to predict if a new patient is victim of a given disease, say cancer. The hospitals' goal is to collaborate in order to improve the early detection of cancer. Building a model from a larger dataset than the personal databases would lead to improved detection capabilities. Nevertheless, these medical databases are highly-sensitive and the information they contain about the patients cannot be disclosed [37]. In such a setting, the hospitals wish to collaboratively train a global model while preserving confidentiality of their records. To do so, the idea is to rely on an aggregating institution (e.g. the World Health Organisation). This would amount to creating a three-party architecture: hospitals, aggregating institution, global model. Note that in our example, and in many real-world settings, all the training data providers may be recipients of the global model, or the global model may even be totally public. Hence, the global model may be exposed to attacks like membership inference attacks [45] that could indicate with high accuracy the probability that one patient was present in a database. Also, given a set of instances, the risk of a model inversion attack [49] which tries to infer sensitive attributes on the instances from a supposedly non-sensitive (often white-box) access to the model, is to be seriously taken into account as it would allow to infer for example that some of the hospital databases contain more ill patients than others. Besides, the aggregating institution might be the target of cyberattacks aimed at stealing data from it. For all these reasons, the three-party architecture we consider has to be resistant to threats coming from *both the aggregation server and the global model recipients*.

Another motivating example, from the field of cybersecurity, is when several actors each hold a database of cybersecurity incident signatures that have occurred on their customer networks. The actors would rely on a third-party server to train the global model. In this scenario, it is a great security issue if the global model suffers from an attack (e.g. if the model features can be inferred [46, 48, 50] with limited access to the model). In this case, this would clearly leak some information on the detection capabilities of the actors, giving a clear advantage to cyberattackers on the networks they supervise.

Deployment scenario and threat model. To perform the aggregation in a private way, we work in the tripartite setting summarised in Figure 1 and formally detailed in Section 4. The *student* (who holds the global model, a.k.a. the *student model*) is the owner of the homomorphic encryption scheme under which encrypted-domain computations will be performed by the *aggregation server*. This means that the student generates and knows both the encryption and decryption keys `pk` and `sk`. Then, when being submitted an unlabelled input, the data holders (a.k.a. the *teachers*) noise the predictions from their personal models, encrypt them under `pk` and send these encryptions to the server. The server has the responsibility to homomorphically perform the aggregation in order to produce an encryption of the output (e.g. a label) which will be sent back to the student and used by the latter for learning, after due decryption. *Homomorphic encryption* thus provides a countermeasure to confidentiality threats on the teachers' predictions from the aggregation server, while the noise introduced by the actor addresses, via *differential*

privacy, the issue of attacks against the student model. In this setting, we assume that the student model is public or at least available to all the actors of the protocol, namely the teachers, the aggregation server and, of course, the student. Our mechanism is differentially private in this context, and our guarantees still hold against a malicious teacher, who has the information of the noise she generated, or even against colluding teachers (see Section 5). On the contrary, we do not address threats whereby the student and the aggregation server collude in the sense that the student does not share sk with the server (in which case they would both get access to the teachers' predictions). We do not consider either threats where the aggregation server behaves maliciously, e.g. to prevent the student model from effectively learning from the teachers, leading to more or less stealthy forms of denial-of-service, or to perform a chosen ciphertext attack via selected queries to the student model. This is the typical scenario in which homomorphic encryption intervenes and our setting thus covers the threat model whereby the aggregation server is assumed to operate properly but may perform computations on observed data to retrieve information. This threat model is commonly known as the *honest-but-curious* model [7, 24, 26].

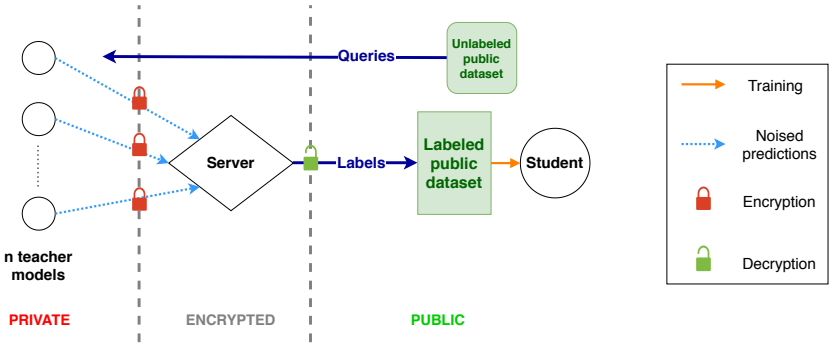

Fig. 1 SPEED - Teacher models send to the aggregation server their encrypted noisy answers to the student's queries. The server homomorphically performs the aggregation in the encrypted domain and sends the result to the student model which decrypts it and uses it for training

Our contribution. In this paper, we present a complete collaborative learning protocol which is secure along the whole workflow regarding a large scope of threats. We ensure protection of the data against any malicious actor of the protocol during the learning phase and prevent indirect information leakage from the final model using *both* homomorphic encryption and differential privacy. While our framework is agnostic to the kind of models used by both the teachers and the student, to the best of our knowledge this is the first work with this level of protection to be efficient enough to apply to deep learning, therefore allowing very good accuracy on difficult tasks such as image classification, as shown by the experiments we ran. Our framework is also bandwidth-efficient and does not require more interactions than required by the baseline protocol.

Outline of the paper. Section 2 relates our work to the literature. In Section 3, we give some technical background on differential privacy and homomorphic encryp-

tion. We describe our SPEED framework in Section 4 and analyse its differential privacy guarantees in Section 5. Section 6 presents our experimental results - SPEED achieves state-or-the-art accuracy and privacy with a mild computational overhead w.r.t previous works. Section 7 concludes the paper and states some open questions for further works.

2 Related work

Differential privacy (DP). Recent works considered to use differential privacy in collaborative settings close to the one we consider [5, 6, 12, 20, 35, 36]. Among them, the most efficient technique in terms of accuracy and privacy guarantees is Private Aggregation of Teacher Ensembles (PATE) first presented in [35] and refined in [36]. PATE uses semi-supervised learning to transfer to the student model the knowledge of the ensemble of teachers by using a differentially private aggregation method. This approach considers a setting very close to ours with the notable difference that the aggregation server is trusted. Hence, applying PATE in our scenario makes the teacher models vulnerable. To tackle this issue, our work builds upon PATE idea with two key differences: we let the responsibility of generating the noise to the teachers and we add a layer of homomorphic encryption in order for the overall learning to be kept private. Another difference can also be noted. To derive privacy guarantees, PATE assumes that two databases d and d' are adjacent if only one sample of the personal database d_i of one teacher i changes, with the hypothesis that the personal databases d_i are disjoint. We do not need this hypothesis and we only consider the teacher models, not the personal databases they use to train them. This leads us to a more powerful definition of adjacency: two databases d and d' are adjacent if they differ by one teacher.

Homomorphic Encryption (HE). HE allows to perform computations over encrypted data. In particular, this can be used so that the model can perform both training and prediction without handling cleartext data. In terms of learning, the naive approach would be to have the training sets homomorphically encrypted, sent to a server for training to be done in the encrypted domain and the resulting (encrypted) model sent back to the participants for decryption. However, putting aside many subtleties, even by deploying all the arsenal available in the HE practitioner toolbox (batching, transciphering, etc.) this would be impractical as "classical" learning is both computation and know-how intensive and HE operations are intrinsically costly. As a consequence, there are only very few works that capitalise on HE for private training [24, 25, 32] and inference [21, 27] of machine learning tasks. Moreover, since some attacks can be performed in a black-box setting, the system is still vulnerable to attacks from the end user who has access to the decryption key. In our framework, we do not use HE directly to build the model, we use it as a mean for the aggregation to be kept private. That way, we are protected against potential threats from the aggregation server, which does not have the decryption key, and we keep a manageable computational overhead.

Federated learning. Federated learning approaches gather several users who own data and make them collaborate in an iterative workflow in order to train a global model. The most famous federated learning algorithm is federated averaging [33]

which is a parallelised stochastic gradient descent. In a context of sensitive user data, several works proposed privacy-preserving federated learning or closely related distributed learning that make use of differential privacy [20, 44], cryptographic primitives [7,8,39] or both [12,40,41]. These methods require online communication between the parties whereas our solution takes advantage of homomorphic encryption and the existence of personal trained models to avoid online communication and drastically limit the interactions, that are both bandwidth-consuming and vulnerable to attacks.

Private aggregation. Several approaches have been considered to limit the need for a trusted server when applying differential privacy, for example by considering local differential privacy [15, 28, 29]. In practice it often results in applying too much noise, and maintaining utility can be difficult [29, 47] especially for deep learning applications. In order to recover more accuracy while keeping privacy, some works combined decentralised noise distribution (*a.k.a.* distributed differential privacy [43]) and encryption schemes [2,22,38,43] in the context of aggregation of distributed time-series. Our work contributes to this line of research. However, our framework is the first one to be efficient enough to investigate deep learning applications while combining distributed DP and HE. Another advantage of our solution concerns fault tolerance regarding the added noise. Some works addressed the problem of fault tolerance by making the server generate the noise that some users did not generate [4] while other works assume that the users themselves adapt the noise they generate to the possible failures [11]. In our setting, because of the encryption and the absence of communication between the teachers, we cannot suppose that any honest entity knows if some failures occurred. Moreover, the addition of noise to compensate a failure does not solve the problem of colluding teachers who may still send noise but do not keep it secret. In our protocol, the task of an honest actor (teacher or server) does not depend on the number of failures and we provide privacy guarantees as a function of the number of failures (see Section 5) - it then suffices to assume an upper bound on this number to ensure a privacy guarantee.

Secure Multi-Party Computation (SMPC). Secure Multi-Party Computation is a general approach that enables several parties to collaboratively perform a given computation without revealing to the other parties any more information than the result of this computation. In particular, secure aggregation regroups approaches which use SMPC techniques as one-time pads masking [7,8] or secret-sharing [14] to perform aggregation over sensitive data. Although these approaches are very close in intent to FHE-based ones, as the present one, they achieve different trade-offs. In a nutshell, when FHE is computation-intensive and non-interactive, SMPC puts more stress on protocol interactions. SMPC requires a lot of communication (garbled circuit generation and evaluation, oblivious input key retrieval, secret key sharing), both time-consuming and vulnerable to attacks, and needs in general that *all* teachers play their role in the protocol for it to terminate - or fixing the fault tolerance issue implies additional rounds of communication [7,8]. On the contrary, the FHE approach is more versatile, requires no interaction among the teachers and is robust to temporary teacher unavailability. Still, at the time of writing, it is the authors' opinion that both approaches are worth investigating in their own right (and this paper obviously belongs to the FHE thread of research).

3 Preliminaries

3.1 Differential privacy

Differential privacy [16] is a gold standard concept in privacy preserving data analysis. It provides a guarantee that under a reasonable privacy budget (ϵ, δ), two adjacent databases produce statistically indistinguishable results. In this section, two databases d and d' are said adjacent if they differ by at most one example.

Definition 1 A randomised mechanism $\mathcal{A}$ with output range $\mathcal{R}$ satisfies (ϵ, δ)-*differential privacy* if for any two adjacent databases d, d' and for any subset of outputs $S \subset \mathcal{R}$ one has

$$\mathbb{P}\left[\mathcal{A}(d) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(d') \in S\right] + \delta.$$

Let us also present a famous and widely used differentially private mechanism, known as the *report noisy max* mechanism.

Definition 2 Let $K \in \mathbb{N}^*$, and let $\mathcal{X}$ be a set that can be partitioned into K subsets $\mathcal{X}_1, \ldots, \mathcal{X}_K$. The mechanism that, given a database d of elements of $\mathcal{X}$, reports $\text{argmax}_{k\in[K]}\left[n_k + Y_k\right]$, where $[K] := \{1, \ldots, K\}$, $n_k := |d \cap \mathcal{X}_k|$ and Y_k is a Laplace noise with mean 0 and scale $\frac{1}{\gamma}$, $\gamma \in \mathbb{R}_+^*$, is called *report noisy max*.

Theorem 1 ([17]) *Let $\mathcal{A}$ be the report noisy max as above. Then $\mathcal{A}$ is $(2\gamma, 0)$-differentially private.*

We now define the notion of *infinite divisibility* that we will use to implement distributed differential privacy.

Definition 3 A random variable Y is said to be *infinitely divisible* if, for any $m \in \mathbb{N}^*$, we can find a family $(X_{m,i})_{i\in[m]}$ of independent and identically distributed (i.i.d.) random variables such that Y has the same distribution as $\sum_{i=1}^m X_{m,i}$.

The following proposition from [30] claims that the Laplace distribution is infinitely divisible [1], enabling to distribute its generation among an arbitrary number of agents.

Proposition 1 ([30]) *Let $m \in \mathbb{N}$ and $\gamma \in \mathbb{R}_+^*$. Let $G_p^{(i)}$, for $(i,p) \in [m] \times [2]$, be i.i.d. random variables following the Gamma distribution of shape $\frac{1}{m}$ and scale $\frac{1}{\gamma}$. Then $\sum_{i=1}^m \left(G_1^{(i)} - G_2^{(i)}\right)$ follows the Laplace distribution of mean 0 and scale $\frac{1}{\gamma}$. The Laplace distribution is said to be* infinitely divisible.

Definition 4 Let $\mathcal{A}$ be a randomised mechanism with output range $\mathcal{R}$ and d, d' a pair of adjacent databases. Let aux denote an auxiliary input. For any $o \in \mathcal{R}$, the *privacy loss* at o is defined as

$$c(o; \mathcal{A}, \text{aux}, d, d') := \log\left(\frac{\mathbb{P}[\mathcal{A}(\text{aux}, d) = o]}{\mathbb{P}[\mathcal{A}(\text{aux}, d') = o]}\right).$$

[1] Another well-known example of infinitely divisible probability distribution is the Gaussian distribution which can be seen as the sum of Gaussian distributions of well chosen scale parameter. In a possible further work, we could indeed replace the (distributed) Laplace noise by a (distributed) Gaussian noise.

We define the *privacy loss random variable* $C(\mathcal{A}, \text{aux}, d, d')$ as

$$C(\mathcal{A}, \text{aux}, d, d') := c(\mathcal{A}(d); \mathcal{A}, \text{aux}, d, d')$$

i.e. the random variable defined by evaluating the privacy loss at an outcome sampled from $\mathcal{A}(d)$.

In order to determine the privacy loss of our protocol, we use a traditional two-fold approach. First of all, we determine the privacy loss per query and, in a second step, we compose the privacy losses of each query to get the overall loss. The classical composition theorem (see e.g. [17]) states that the guarantees ϵ of sequential queries add up. Nevertheless, training a deep neural network, even with a collaborative framework as presented in this paper, requires a large amount of calls to the databases, precluding the use of this classical composition. Therefore, to obtain reasonable DP guarantees, we need to keep track of the privacy loss with a more refined tool, namely the moments accountant [1] that we introduce here, deferring the details of the method in Section A.1 of the appendix.

Definition 5 With the same notations as above, the *moments accountant* is defined for any $l \in \mathbb{R}_+^*$ as

$$\alpha_{\mathcal{A}}(l) := \max_{\text{aux},d,d'} \alpha_{\mathcal{A}}(l; \text{aux}, d, d')$$

where the maximum is taken over any auxiliary input aux and any pair of adjacent databases (d, d') and $\alpha_{\mathcal{A}}(l; \text{aux}, d, d') := \log\left(\mathbb{E}\left[\exp(lC(\mathcal{A}, \text{aux}, d, d'))\right]\right)$ is the moment generating function of the privacy loss random variable.

3.2 Homomorphic encryption

Let us consider Λ and Ω which respectively are the set of cleartexts (*a.k.a.* the clear domain) and the set of ciphertexts (*a.k.a.* the encrypted domain). A homomorphic encryption system first consists in two algorithms $\mathsf{Enc}_{\mathsf{pk}} : \Lambda \longrightarrow \Omega$ and $\mathsf{Dec}_{\mathsf{sk}} : \Omega \longrightarrow \Lambda$ where pk and sk are data structures which represent the public encryption key and the private decryption key of the cryptosystem.

Homomorphic encryption systems are by necessity probabilistic, meaning that some randomness has to be involved in the Enc function and that the ciphertexts set Ω is significantly much bigger than the cleartexts set Λ. Any (decent) homomorphic encryption scheme possesses the semantic security property meaning that, given $\mathsf{Enc}(m)$ and polynomially many pairs $(m_i, \mathsf{Enc}(m_i))$ it is hard[2] to gain any information on m with a significant advantage over guessing. Most importantly, a homomorphic encryption scheme offers two other operators $\oplus$ and $\otimes$ where

- $\mathsf{Enc}(m_1) \oplus \mathsf{Enc}(m_2) = \mathsf{Enc}(m_1 + m_2) \in \Omega$
- $\mathsf{Enc}(m_1) \otimes \mathsf{Enc}(m_2) = \mathsf{Enc}(m_1 m_2) \in \Omega$

[2] "Hard" means that it requires solving a reference (conjectured) computationally hard problem on which the security of the cryptosystem hence depends. From a practical viewpoint, given a security target λ, the concrete parameters of a homomorphic scheme are chosen such that the best known (exponential-time) algorithms for solving the underlying reference problem require an order of magnitude of 2^λ nontrivial operations.

When these two operators are supported without restriction by a homomorphic scheme, it is said to be a Fully Homomorphic Encryption (FHE) scheme. A FHE with $\Lambda = \mathbb{Z}_2$ is Turing-complete and, as such, is *in principle* sufficient to perform any computation in the encrypted domain with a computational overhead depending on the security target[3]. *In practice*, though, the $\oplus$ and $\otimes$ are much more computationally costly than their clear domain counterparts which has led to the development of several approaches to HE schemes design each with their pros and cons.

Somewhat HE (SHE). Somewhat homomorphic encryption schemes, such as BGV [10] or BFV [18], provide both operators but with several constraints. Indeed, in these cryptosystems the $\otimes$ operator is much more costly than the $\oplus$ operator and the cost of the former strongly depends on the *multiplicative depth* of the calculation, that is the maximum number of multiplications that have to be chained (although this depth can be optimised [3]). Interestingly, most SHE schemes offer a *batching* capability by which multiple cleartexts can be packed in one ciphertext resulting in (quite massively) parallel homomorphic operations i.e.,

$$\mathsf{Enc}(m_1, ..., m_\kappa) \oplus \mathsf{Enc}(m'_1, ..., m'_\kappa) = \mathsf{Enc}(m_1 + m'_1, ..., m_\kappa + m'_\kappa) \tag{1}$$

(and similarly so for $\otimes$). Typically, several hundreds such slots are available which often allows to significantly speed up encrypted-domain calculations.

Fully HE (FHE). Fully homomorphic encryption schemes offer both the $\oplus$ and $\otimes$ operators without restrictions on multiplicative depth. At the time of writing, only the FHE-over-the-torus approach, instantiated in the TFHE cryptosystem [13], offers practical performances. In this cryptosystem, $\oplus$ and $\otimes$ have the same constant cost. On the downside, TFHE offers no batching capabilities. To get the best of all worlds, the TFHE scheme is often hybridised with SHE by means of operators allowing to homomorphically switch among several ciphertext formats [9, 32] to perform each part of calculation with the most appropriate scheme (see e.g. [51]).

4 SPEED: Secure, Private, and Efficient Deep Learning

4.1 A distributed learning architecture

Let us consider a set of n owners (a.k.a. *teachers*) each holding a personal sensitive model f_i. We assume that we also have an unlabelled public database D. The goal is to label D using the knowledge of the private (teacher) models to train a collaborative model (a.k.a. *student model*) mapping an input space $\mathcal{X}$ to an output space $[K] = \{1, \ldots, K\}$. To do so while keeping the process private, we follow the setting illustrated by Figure 1 relying on a (distrusted) aggregation server:

1. For every sample x of the public database D, the student sends x to the aggregator requesting it to output label for x. The aggregator forwards this request to the n teachers.

[3] Polynomial in λ.

2. Each teacher i labels x using its own private model f_i. Then each teacher adds noise to the label (see Section 4.2) and encrypts the noisy label before sending it to the aggregation server.
3. The aggregator performs a homomorphic aggregation of the noisy labels and returns the result to the student model, namely the most common answered label (see Section 4.3).
4. The student, who owns the decryption key, decrypts the aggregated label and is then able to use the labelled sample to train its model.

Our framework addresses two kinds of threats using two complementary tools. On one hand, differential privacy protects the sensitive data from attacks against the student model. Indeed, some model inversion attacks [49] might disclose the training data of the student model, and especially the labels of database D. But differential privacy ensures that the noise applied to the teachers' answers prevents the aggregated labels from leaking information about the sensitive models f_i [4]. On the other hand, the homomorphic encryption of the teachers' answers prevents the aggregator to learn anything about the sensitive data while enabling it to blindly compute the aggregation.

4.2 Noise generation and threat models

When requested to label a sample x, each owner i uses its model f_i to infer the label of x. In order for the aggregator to compute the most common label in the secret domain, the owner must send a one-hot encoding of the label. That is, rather than sending $f_i(x)$, the i-th teacher sends a K-dimensional vector, say $z^{(i)}$, whose $f_i(x)$-th coordinate is an encryption of 1 while all the other coordinates are encryptions of 0. To guarantee differential privacy (see Section 5 for the formal analysis), the owner adds to this one-hot encoding a noise drawn from $G_1^{(i)} - G_2^{(i)}$ where the $G_1^{(i)}$ and $G_2^{(i)}$ are $2n$ i.i.d. K-dimensional random variables following the Gamma distribution of shape $\frac{1}{n}$ and scale $\frac{1}{\gamma}$, where $\gamma \in \mathbb{R}_+^*$. Then, i sends the (encrypted) noisy one-hot encoded vector whose k-th coordinate corresponds to $z_k^{(i)} + G_{k,1}^{(i)} - G_{k,2}^{(i)}$.

Assuming that the aggregator has access to the student model, distributing the responsibility of adding the noise among all the teachers instead of delegating this task to the aggregator (see paragraph on centralised noise below) is necessary to protect the data against an honest-but-curious aggregator. Indeed, such an aggregator could use the information of the noise it generated to break the differential privacy guarantees and, potentially, recover the sensitive data by model inversion on the student model. Note that such an attack does not break the honest-but-curious assumption since the aggregator still performs its task correctly.

Beyond the honest-but-curious model In a model that would go beyond the honest-but-curious aggregator hypothesis, the capability for the aggregator to add it own noise is even more harmful for the privacy (and of course, the accuracy) than not using noise at all. Indeed it gives the aggregator much more freedom to attack. As an example, think about a malicious aggregator that wants to know a characteristic

[4] Thanks to the DP guarantees, the labels of D could actually be published as well.

χ on a particular teacher, called its victim. Given a query, for all $k \in [K]$, we write $n_k := |\{i : f_i(x) = k\}|$ and call it the number of *votes* for class k. Let us suppose that, for a given query, changing the value of the victim's characteristic χ from χ_0 to χ_1 also changes the victim's vote from a class k_0 to a class k_1. Hence, by denoting $n_{k_0} = \nu_0$ and $n_{k_1} = \nu_1$ if $\chi = \chi_0$ we get $n_{k_0} = \nu_0 - 1$ and $n_{k_1} = \nu_1 + 1$ if $\chi = \chi_1$. Then, if the aggregator knows all the n_k for $k \in [K] \setminus \{k_0, k_1\}$ and knows ν_0 and ν_1 (which are the classical hypotheses in differential privacy), it can add just as much noise as needed for the class k_0 to be the argmax if and only if $\chi = \chi_0$ [5]. The result from the homomorphic argmax would then leak the information about the value of the victim's characteristic χ.

Centralised noise generation In a context in which the student model is kept private and, especially, not available to the aggregator, we can consider a centralised way of generating the noise. If we do not trust the teachers to generate the noise, we can charge the aggregator to do it, since it will not be able to use the knowledge of the noise to attack the sensitive data via the student model. The aggregator only needs to generate a Laplace noise (in the clear domain), and homomorphically add it to the unnoisy encryption of n_k it receives from the teachers. The infinite divisibility of the Laplace distribution (Proposition 1) shows that the resulting noise is the same as in the case presented above in which each teacher generates an individual noise drawn from the difference of two Gamma distributions. The privacy cost of one request is simply the privacy cost of the *report noisy max*, namely 2γ (Theorem 1).

In a nutshell, we can consider the following different threat models:

- honest (H) : the aggregation server performs its tasks properly and do not try to retrieve information from the data it has access to
- honest-but-curious (HBC) : the aggregation server performs its tasks properly but it may compute the available data to get sensitive information
- beyond honest-but-curious (BHBC) : the aggregation server performs the aggregation correctly but cannot be trusted to properly generate the noise necessary to the DP guarantees

Table 1 summarises against which kind of server our protocol is protected, depending on the access the server has to the student model and on the way the noise is generated. As already emphasised, we focus on the case where the student model is public and the noise is distributively generated by the teachers because it is the most general model among the realistic threat models and thus gives the better tradeoff between flexibility and security.

Table 1 Robustness of our framework depending on the availability of the student model and the noise generation

	Private model	Public model
Centralised noise	**HBC**	**H**
Distributed noise	**BHBC**	**BHBC**

[5] For example, add $\nu_0 - \frac{1}{2} - n_k$ to all the classes except k_0 and k_1, $\nu_0 - 1 - \nu_1$ to the class k_1 and nothing to the class k_0.

4.3 Technical details on the homomorphic aggregation

Summing the noisy counts The aggregation server receives the n encrypted noisy labels and sums them up in the secret domain. Due to the infinite divisibility of the Laplace distribution, the server obtains a K-dimensional vector whose k-th ($k \in [K]$) coordinate is an encryption of:

$$\sum_{i=1}^{n} \left(z_k^{(i)} + G_{k,1}^{(i)} - G_{k,2}^{(i)} \right) = n_k + Y_k$$

where $n_k := |\{i : f_i(x) = k\}|$ and Y_k is a Laplace noise with mean 0 and scale $\frac{1}{\gamma}$.

So far, we have only needed homomorphic addition which is a good start. Then an argmax operator must be performed after the summation. However, *efficiently* handling the highly nonlinear argmax function by means of FHE is much more challenging.

Computing the argmax. Most prior work on secure argmax computations use some kind of interaction between a party that holds a sensitive vector of values and a party that wants to obtain the argmax over those values. The non-linearity of the argmax operator presents unique challenges that have mostly been handled by allowing the two interested parties to exchange information. This means increased communication costs and, in some cases, information leakage. This is with the exception of [51]. They provide a fully non-interactive homomorphic argmax computation scheme based on the TFHE encryption. We implemented and parametrised their scheme to fit the specific training problems presented in Section 6. We present here the main idea behind this novel FHE argmax scheme. For more details, see the original paper. The TFHE encryption scheme provides a *bootstrap* operation that can be applied on any scalar ciphertext. Its purpose is threefold: switch the encryption key; reduce the noise; apply a non-linear operation on the underlying plaintext value. This underlying operation can be seen as a function

$$g_{t,a,b}(x) = \begin{cases} a & \text{if} \quad x > t \\ b & \text{if} \quad x < t. \end{cases}$$

One notable application is that of a "sign" bootstrap: we can extract the sign of the input with the underlying function $g_{0,1,0}(x)$. The argmax computation in the ciphertext space is made as follows. For every k, k', $k \neq k'$, we compare the values $n_k + Y_k$ and $n_{k'} + Y_{k'}$ with a subtraction $(n_k + Y_k - n_{k'} - Y_{k'})$ and application of a sign bootstrap operation. This yields $\theta_{k,k'}$, a variable with value 1 if $n_k + Y_k > n_{k'} + Y_{k'}$ and 0 otherwise. Therefore the complexity will be quadratic in the number of classes. For a given k we can then obtain a boolean truth value (0 or 1) for whether $n_k + Y_k$ is the maximum value. To this end, we compute

$$\Theta_k = \sum_{i \neq k} \theta_{k,i}.$$

n_k is the max if and only if, for all i one has $\theta_{k,i} = 1$ *i.e.* $\Theta_k = K - 1$. We can therefore apply another bootstrap operation with $g_{K-\frac{3}{2},1,0}$. If $\Theta_k = K - 1$, the boostrap will return an encryption of 1, and return an encryption of 0 otherwise.

Once decrypted, the position of the only non-zero value is the argmax. Because the underlying function $g_{t,a,b}$ is applied homomorphically, its output is inherently probabilistic. In the FHE scheme used, an error is inserted in all the ciphertexts at encryption time to ensure an appropriate level of security. This means that if two values are too close, then the sign bootstrap operation might return the wrong result over their difference. The exact impact of this approximation on the accuracy is evaluated in Section 6.

Remark. Another solution would be to send the noisy histogram $n_k + Y_k$ of the counts for each class k to the student and let her process the argmax in the clear domain. This could indeed be performed with a plain-old additively-homomorphic cryptosystem such as Paillier or (additive-flavored) ElGamal, avoiding the machinery of the homomorphic argmax. Nevertheless, this approach was put aside because sending the whole histogram instead of the argmax would provide much worse DP guarantees.

5 Differential privacy analysis

In this section, we will give privacy guarantees considering that two databases d and d' are adjacent if they differ by one teacher i.e. there exists $i_0 \in [n]$ such that $f_{i_0} \neq f'_{i_0}$ and, for all $i \in [n] \setminus \{i_0\}$, $f_i = f'_i$. This definition of adjacency is quite conservative and is strictly larger than the definition of adjacency from [35] (indeed, in the assumption whereby the personal teacher databases d_i are disjoint, changing one sample from a personal database changes at most one teacher).

Robustness against colluding teachers. As we have decided not to trust the aggregation server to generate the noise necessary to the privacy guarantees, we may also assume that a subset of teachers might be malicious and collude by communicating their generated noise, which gives the same DP guarantees from the point of view of a colluding teacher as if they would have not generated any noise and, to this extent, our protocol, which addresses this issue, is fault tolerant. The following theorem quantifies the privacy cost of such failures.

In the following, we call $\mathcal{A}$ the aggregation mechanism that outputs the argmax of the noisy counts. $\mathcal{A}(d, Q)$ is the output of $\mathcal{A}$ for the database d and the query Q. Let $\gamma \in \mathbb{R}_+^*$ be the inverse scale parameter of the distributed noise. Considering the DP guarantees from the point of view of an entity $\mathcal{E}$, let $\tau \in (0, 1)$ be the ratio of the teachers whose noise is ignored by $\mathcal{E}$.

Theorem 2 *Let us define* $I \colon v \in \mathbb{R}_+^* \mapsto \int_0^{+\infty} (t+v)^{\tau-1} t^{\tau-1} e^{-2t} dt$ *and* $g \colon t \in \mathbb{R} \mapsto \frac{\int_{\gamma t}^{+\infty} e^{-v} I(v) dv}{\int_{\gamma(t+2)}^{+\infty} e^{-v} I(v) dv}$.

Then, from $\mathcal{E}$'s point of view, $\mathcal{A}$ is $(\epsilon, 0)$-differentially private, with

$$\epsilon = \log\left(1 + 2\frac{\int_0^{\gamma} e^{-v} I(v) dv}{\int_{2\gamma}^{+\infty} e^{-v} I(v) dv}\right).$$

Moreover, if $\tau > \frac{1}{2}$, g *is differentiable in* 0 *and* $\mathcal{A}$ *is* $(\epsilon', 0)$*-differentially private, with*

$$\epsilon' = \min\left[\epsilon, \log\left(g(0) - g'(0)\right)\right]$$

where $g'(0) = \gamma \frac{\frac{\Gamma(\tau)^2}{2} e^{-2\gamma} I(2\gamma) - I(0) \int_{2\gamma}^{+\infty} e^{-v} I(v) dv}{\left(\int_{2\gamma}^{+\infty} e^{-v} I(v) dv\right)^2}$.

Sketch of proof. Adapting the proof of the privacy cost of the report noisy max from [17], we first show that, if we can find a function M of γ and τ such that, for any $t \in \mathbb{R}$, $g(t) \leq M$, then $\mathcal{A}$ is $(\log(M), 0)$-differentially private. This motivates us to find an upper bound of g.

To do so, we prove that g has a maximum on $\mathbb{R}$ and that this maximum is reached on the interval $[-1; 0]$. On one hand, we show that, for all $t \in [-1; 0]$, $g(t) \leq 1 + 2\frac{\int_0^{\gamma} e^{-v} I(v) dv}{\int_{2\gamma}^{+\infty} e^{-v} I(v) dv}$. On the other hand, we prove that, if besides $\tau > \frac{1}{2}$, then g is concave on $[\text{argmax}(g); 0]$ and thus, for all $t \in [-1; 0]$, $g(t) \leq g(0) - g'(0)$ (note that g is not differentiable in 0 if $\tau \leq \frac{1}{2}$). □

Denoting S the subset of teachers who are honest (i.e. do not collude), this theorem allows us to control the privacy cost by the ratio τ of the teachers who kept their noise secret, from the point of view of both:

- a colluding teacher, taking $\tau = \frac{|S|}{n}$
- an honest teacher, taking $\tau = \frac{n-1}{n}$
- any entity who has access to the student model but is not a teacher, taking $\tau = 1$

Note that we can also use Theorem 2 in the hypothesis whereby the colluding teachers publish their noise (to the whole world), adapting τ in consequence [6]. For $\tau = 1$, the privacy guarantee is given by $\lim_{\tau \to 1} \epsilon'$ which, as shown by Proposition 2, is the classical bound of the *report noisy max* with a centralised Laplace noise.

Proposition 2 *For all* $\gamma \in \mathbb{R}_+^*$, $\lim_{\tau \to 1} \left[\log(g(0) - g'(0))\right] = 2\gamma$.

Furthermore, Proposition 3 shows that, naturally, the privacy cost tends to be null when the noise becomes infinitely large (γ approaches 0).

Proposition 3 *For all* $\tau \in (0, 1)$, $\lim_{\gamma \to 0} \left[\log\left(1 + 2\frac{\int_0^{\frac{\gamma}{2}} e^{-v} I(v) dv}{\int_{\gamma}^{+\infty} e^{-v} I(v) dv}\right)\right] = 0$.

Let us also give an upper bound of the probability that the noisy argmax is different from the true argmax.

Proposition 4 *Let* k^* *be the class corresponding to the true argmax.*
If $\tau \in (\frac{1}{2}; 1)$,

$$\mathbb{P}[\mathcal{A}(d; Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma \Delta_k} \left[\frac{1}{2} + \frac{(\gamma \Delta_k)^{2\tau - 1}}{\tau 2^{4\tau - 2} \Gamma(\tau)^2}\right]$$

where $\Delta_k \coloneqq n_{k^*} - n_k$ *for any* $k \in [K]$ *and* $\Gamma : \beta \in \mathbb{R}_+^* \mapsto \int_0^{+\infty} t^{\beta - 1} e^{-t} dt$ *is the gamma function.*

[6] e.g. the privacy guarantee for an honest teacher would be computed with $\tau = \frac{|S|-1}{n}$.

If $\tau \in (0; \frac{1}{2}]$,

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma \Delta_k} \left[\frac{1}{2} + \frac{(\gamma \Delta_k)^{\frac{\tau}{2}}}{\tau 2^{\frac{5}{2}\tau - 1} \Gamma(\tau)^2} \times \left(\frac{3}{2}\tau\right)^{\frac{3}{2}\tau} \left(\frac{2}{\tau} - 3\right)^{1 - \frac{3}{2}\tau} \right].$$

Sketch of proof. The event $(\mathcal{A}(d;Q) \neq k^*)$ is the union of the events $(n_k + Y_k \geq n_{k^*} + Y_{k^*})$, for $k \in [K] \setminus \{k^*\}$, and thus $\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} \mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$. We remark that, for any $k \in [K] \setminus \{k^*\}$,

$$\begin{aligned}\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*}) &= \mathbb{P}(Y_{k^*} \leq Y_k - \Delta_k) \\ &= \int_{-\infty}^{0} f(t) F(t - \Delta_k) dt + \int_{0}^{\Delta_k} f(t) F(t - \Delta_k) dt + \int_{\Delta_k}^{+\infty} f(t) F(t - \Delta_k) dt\end{aligned}$$

where $f \colon u \in \mathbb{R}^* \mapsto \frac{\gamma}{\Gamma(\tau)^2} e^{-\gamma |u|} I(\gamma |u|)$ and $F \colon t \in \mathbb{R} \mapsto \int_{-\infty}^{t} f(u) du$.

We show that $\int_{\Delta_k}^{+\infty} f(t) F(t - \Delta_k) dt \leq \frac{3}{8} e^{-\gamma \Delta_k}$ and $\int_{-\infty}^{0} f(t) F(t - \Delta_k) dt \leq \frac{1}{8} e^{-\gamma \Delta_k}$. Moreover, using Hölder's inequality, we show that, for all $q \in (\frac{1}{1-\tau}; +\infty)$, calling $p := \frac{1}{1 - \frac{1}{q}}$, $\int_{0}^{\Delta_k} f(t) F(t - \Delta_k) dt \leq \frac{e^{-\gamma \Delta_k}}{\tau 2^{4\tau - 2 + \frac{1}{q}} \Gamma(\tau)^2} \times \frac{(\gamma \Delta_k)^{2\tau - 1 + \frac{1}{q}}}{p^{\frac{1}{p}} [q(1-\tau) - 1]^{\frac{1}{q}}}$. For $\tau > \frac{1}{2}$, we take the particular (and classic) case of the limit of the previous bound when q tends to $+\infty$. For $\tau \leq \frac{1}{2}$, we take $q = \frac{1}{1 - \frac{3}{2}\tau}$. □

Theorem 2 and Proposition 4 serve as building blocks to which we apply the following theorem from [35].

Theorem 3 ([35]) *Let $\epsilon, l \in \mathbb{R}_+^*$. Let $\mathcal{A}$ be a $(\epsilon, 0)$-differentially private mechanism and $q \geq \mathbb{P}[\mathcal{A}(d) \neq k^*]$ for some outcome k^*. If $q < \frac{e^{\epsilon} - 1}{e^{2\epsilon} - 1}$, then for any additional information* aux *and any pair (d, d') of adjacent databases, $\mathcal{A}$ satisfies*

$$\alpha_{\mathcal{A}}(l; \text{aux}, d, d') \leq \min \left[\epsilon l, \frac{\epsilon^2 l(l+1)}{2}, \log \left((1-q) \left(\frac{1-q}{1 - e^{\epsilon} q} \right)^l + q e^{\epsilon l} \right) \right].$$

As in [35], Theorem 3 coupled with some properties of the moments accountant (composability and tail bound) allows one to devise the overall privacy budget (ϵ, δ) for the learning procedure (see Section 6 for numerical results). We refer the interested reader to Section A of the appendix for more details and for the extended proofs of our claims.

Influence of the cryptographic layer. One must be aware that the cryptographic layer perturbates the noisy votes because the computation of the homomorphic argmax has a small probability of error. Although this topic deserves further investigations, we make the assumption that these perturbations are negligible and that they do not change the privacy guarantees as they basically constitute an additional noise on the votes. We further discuss this point in Appendix A.3.

6 Experimental results

The experiments presented below enable us to validate the accuracy of our framework on well-known image classification tasks and illustrate the practicality of our method in terms of performance, since the computational overhead due to the homomorphic layer remains reasonable. The source codes necessary to run the following experiments are available on `https://github.com/Arnaud-GS/SPEED`.

HE time overhead. We implemented the homomorphic argmax computation presented in Section 4.3. Without parallelizing, a single argmax query requires just under 4 seconds to compute on an Intel Core i7-6600U CPU. Importantly, this does not depend on the input data. The costliest operation is the computation of θ. Any other part of the scheme is negligible in comparison. Therefore, once the parameters are set, the time performance depends solely on the number of classes (the number of bootstrap comparisons is quadratic in the number of classes). As such, 100 queries require 6.5 minutes and 1000 queries 65 minutes. Of course, the queries can be performed in parallel to decrease the latency allowing for much more challenging applications.

Homomorphic argmax accuracy. As we mention in Section 4.3, the homomorphic computation of the argmax is inherently probabilistic. This is due both to the noise added to any ciphertext at encryption time, and to limitations of the bootstrapping operation in terms of accuracy. On MNIST dataset [31], we evaluate the method with $\tau = 1/0.9/0.7$ and compare the cleartext argmax to our homomorphic argmax. Our implementation of the HE argmax has an average accuracy of 99.4%, meaning that it retrieves the cleartext argmax 99.4% of the time.
To obtain a more general and conservative measure of the inherent accuracy of the HE argmax (which can be applied on any dataset), we make the teachers give uniformly random answers to the queries. In this setting, most counts n_k are likely to be close to one another, which makes even a classical argmax useless. This kind of scenario can be seen as worst-case, since the teacher voting is adversarial to argmax computation. Even in this scenario, and with the same parameters as for MNIST, our implementation of the HE argmax algorithm still produces an average accuracy of 90%. Hence, an accuracy of 90% can be considered a lower bound for any adaptation of this argmax technique to other datasets. Yet in practice a tweaking of the parameters can yield a better accuracy even for this worst-case scenario, at the cost of time efficiency.

Learning setup. To evaluate the performances of our framework, we test our method on MNIST [31] and SVHN [34] datasets. To represent the data holders, we divide the training set in 250 equally distributed and disjoint subsets, keeping the test set for learning and evaluation of the student model. Then we apply the following procedures. We refer the interested reader to Section C of the appendix for more details on the hyper-parameters and learning procedure.

- *Teacher models.* For MNIST, given a dataset, a data holder builds a local model by stacking two convolutional layers with max pooling and a fully connected layer with ReLu activations. Two additional layers have been added for SVHN.
- *Student model.* Following the idea from [35], we train the student in a semi-supervised fashion. Unlabelled inputs are used to estimate a good prior distribution using a GAN-based technique first introduced in [42]. Then we use a limited amount of queries (100 for MNIST, 500 for SVHN) to obtain labelled examples which we use to fine tune the model.

For MNIST experiments, as the student model can substantially vary based on the selected subset of labelled examples, the out-of-sample accuracy has been evaluated 15 times, with 100 labelled examples sampled from a set of 9000 ones.

For each experiment, the remaining 1000 examples have been used to evaluate the student model accuracy. For SVHN, the computations being much more heavy, the out-of-sample accuracy has been evaluated 3 times, with 500 examples sampled from a set of 10000 ones. We used 16032 examples to test the student model accuracy.

Performances on MNIST. Table 2 displays our experimental results for SPEED with MNIST and compares them to a non-private baseline (without DP or HE) and to the framework that we call Trusted which assumes that the server is trusted and thus only involves DP and not HE. Trusted can be considered as PATE framework from [35] with some subtle differences: the noise is generated in a distributed way in Trusted and the notion of adjacency is larger. Even if the inverse noise scale γ we use is greater than the one in [35] (0.1 instead of 0.05), which should lead to a worse DP guarantee, an argmax-specific analysis of the privacy cost per query allowed us to provide a better DP guarantee ($\epsilon = 1.41$ instead of $\epsilon = 2.04$ with $\delta = 10^{-5}$ and 100 queries). To be more conservative in terms of accuracy, the experiments were run considering that the colluding teachers did not generate any noise, which does not change anything in terms of DP. That is why, in spite of the variability of the accuracy, we observe a tradeoff between accuracy and DP. Indeed, even if the reported average accuracy does not vary much across conditions, consistent rankings of the methods have been observed, confirming the expected average rank of the method based on the amount of added noise. As expected, the best DP guarantee ($\epsilon = 1.41$) is obtained when all the teachers generated noise ($\tau = 1$), but this is the case where the accuracy is the lowest. On the contrary, when some teachers failed to generate noise ($\tau = 0.9$ and $\tau = 0.7$), the counts are more precise, leading to a slightly better accuracy but worse DP guarantees. It should also be noted that the variance is high in each condition. It masks the fact that the distribution is highly skewed, with a majority of results in the $97.5\% - 98.5\%$ range, and a few samplings yielding an out-of-sample accuracy around 90%.

Table 2 Results for MNIST dataset with 250 teachers and 100 student queries. We used an inverse noise scale $\gamma = 0.1$. The DP guarantees, computed by composability with the moments accountant method over the 100 queries, are given for $\delta = 10^{-5}$.

Framework	ϵ	Acc. ($\pm$ std) [%]	HE overhead
Non-private	-	96.22 ($\pm$2.27)	-
Trusted	1.41	95.95 ($\pm$2.97)	-
$\tau = 1$	1.41	95.91 ($\pm$2.57)	
$\tau = 0.9$	1.66	96.02 ($\pm$2.92)	6.5 min
$\tau = 0.7$	2.37	96.06 ($\pm$2.61)	

Figure 2 shows the evolution of our DP guarantee as a function of γ, with $\tau = 0.9$ fixed. Note that the privacy cost decreases for $\gamma \geq 2$ which may seem counterintuitive but the reason is thoroughly explained in Section A.4 of the appendix. Anyway, we observed empirically that the privacy cost has a finite limit in $+\infty$ (approximately 2.87) and remains greater than this limit for any $\gamma \geq 2$. The asymptote is shown by a dashed line on Figure 2.

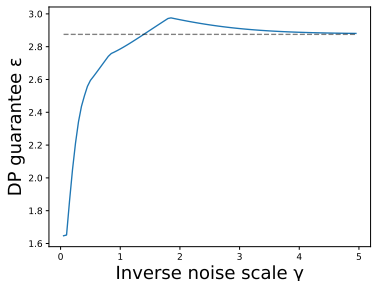

Fig. 2 Differential privacy guarantees for MNIST as a function of γ, with $\tau = 0.9$

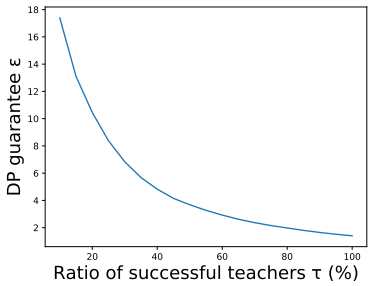

Fig. 3 Differential privacy guarantees for MNIST as a function of τ, with $\gamma = 0.1$

Figure 3 shows the evolution of the DP guarantee as a function of τ, with $\gamma = 0.1$ fixed. As explained before, the greater τ, the better the DP guarantee.

Performances on SVHN. Table 3 presents our experimental results on SVHN dataset [7]. The variance on the accuracy is much smaller than for MNIST dataset because the test set is constituted of 16032 samples. Similarly to the MNIST experiment, the accuracy and the privacy cost increase when less noise is applied because less teachers noised their votes (i.e. when τ is small). The DP guarantees are not as good as for MNIST, this is due to the high amount of queries (500) necessary to obtain a good accuracy because the learning task is more complex.

Table 3 SVHN experimental results for 500 queries, with noise inverse scale $\gamma = 0.1$, $\delta = 10^{-5}$

Framework	ϵ	Acc. [%]	HE overhead
Non-private	-	84.7	-
Trusted	4.73	83.7	-
$\tau = 1$	4.73	83.5	32.5 min
$\tau = 0.9$	5.59	83.8	
$\tau = 0.7$	8.16	84.6	

7 Conclusion and open questions for further works

Our framework allows a group of agents to collaborate and put together their sensitive knowledge while protecting it via two complementary technologies - differential privacy and homomorphic encryption - against any entity contributing to the learning or having access to the final model. Crucially, our experiments showed that our method is practical for deep learning applications, combining high accuracy, mild computational overhead and privacy guarantees adapting to the number of malicious teachers.

[7] Note that our DP guarantee ϵ for Trusted cannot be directly compared with PATE's one since we do not use the same δ.

An interesting further work could investigate the fault tolerance of the privacy guarantees with other noises (e.g. Gaussian noise) or other infinite divisions (Laplace distribution can also be infinitely divided using individual Gaussian noises or individual Laplace noises [23]). A more ambitious direction towards collaborative deep learning with privacy would be to design new aggregation operators, more suitable to FHE performances yet still providing good DP bounds. In particular, a linear or quadratic aggregation operator would be amenable to almost negligible homomorphic computations overhead. This lighter homomorphic layer would enable to extend the applicability of our framework to more complex datasets. Such aggregation operators would also allow to associate homomorphic calculations with verifiable computing techniques (e.g. [19]) whereby the server would provide an encrypted aggregation result along with a formal proof that aggregation was indeed done correctly. These perspectives would then allow to address threats beyond the honest-but-curious model.

8 Declarations

8.1 Funding

The experiments were performed using HPC resources of FactoryIA partially funded by Ile-de-France French region – project SESAME 2017.

8.2 Conflicts of interest/Competing interests

Not applicable.

8.3 Availability of data and material

The MNIST [31] and SVHN [34] datasets can be found respectively at `http://yann.lecun.com/exdb/mnist/` and `http://ufldl.stanford.edu/housenumbers/`.

8.4 Code availability

The source codes used to run the experiments and compute the DP guarantees can be accessed on `https://github.com/Arnaud-GS/SPEED`.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
2. Ács, G., Castelluccia, C.: I have a dream!(differentially private smart metering). In: International Workshop on Information Hiding, pp. 118–132. Springer (2011)
3. Aubry, P., Carpov, S., Sirdey, R.: Faster homomorphic encryption is not enough: improved heuristic for multiplicative depth minimization of boolean circuits. In: CT-RSA, pp. 345–363 (2019)

4. Bao, H., Lu, R.: A new differentially private data aggregation with fault tolerance for smart grid communications. IEEE Internet of Things Journal **2**(3), 248–258 (2015)
5. Beaulieu-Jones, B.K., Yuan, W., Finlayson, S.G., Wu, Z.S.: Privacy-preserving distributed deep learning for clinical data. CoRR **abs/1812.01484** (2018)
6. Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., Rogers, R.: Protection against reconstruction and its applications in private federated learning. arXiv preprint arXiv:1812.00984 (2018)
7. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482 (2016)
8. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191 (2017)
9. Boura, C., Gama, N., Georgieva, M.: Chimera: a unified framework for b/fv, tfhe and heaan fully homomorphic encryption and predictions for deep learning. Cryptology ePrint Archive, Report 2018/758 (2018)
10. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) Fully Homomorphic Encryption Without Bootstrapping. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, pp. 309–325 (2012)
11. Chan, T.H.H., Shi, E., Song, D.: Privacy-preserving stream aggregation with fault tolerance. In: International Conference on Financial Cryptography and Data Security, pp. 200–214. Springer (2012)
12. Chase, M., Gilad-Bachrach, R., Laine, K., Lauter, K.E., Rindal, P.: Private collaborative neural network learning. IACR Cryptology ePrint Archive **2017**, 762 (2017)
13. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In: ASIACRYPT, pp. 3–33 (2016)
14. Danezis, G., Fournet, C., Kohlweiss, M., Zanella-Béguelin, S.: Smart meter aggregation via secret-sharing. In: Proceedings of the first ACM workshop on Smart energy grid security, pp. 75–80 (2013)
15. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pp. 429–438. IEEE (2013)
16. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 486–503. Springer (2006)
17. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
18. Fan, J., Vercauteren, F.: Somewhat practical fully homomorphic encryption. IACR Cryptology ePrint Archive **2012**, 144 (2012)
19. Fiore, D., Gennaro, R., Pastro, V.: Efficiently verifiable computation on encrypted data. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 844–855 (2014)
20. Geyer, R.C., Klein, T., Nabi, M.: Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557 (2017)
21. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning, pp. 201–210 (2016)
22. Goryczka, S., Xiong, L.: A comprehensive comparison of multiparty secure additions with differential privacy. IEEE transactions on dependable and secure computing **14**(5), 463–477 (2015)
23. Goryczka, S., Xiong, L., Sunderam, V.: Secure multiparty aggregation with differential privacy: A comparative study. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, pp. 155–163 (2013)
24. Graepel, T., Lauter, K., Naehrig, M.: Ml confidential: Machine learning on encrypted data. In: International Conference on Information Security and Cryptology, pp. 1–21. Springer (2012)
25. Hesamifard, E., Takabi, H., Ghasemi, M.: Cryptodl: Deep neural networks over encrypted data. arXiv preprint arXiv:1711.05189 (2017)
26. Ishai, Y., Kilian, J., Nissim, K., Petrank, E.: Extending oblivious transfers efficiently. In: Annual International Cryptology Conference, pp. 145–161. Springer (2003)

27. Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: {GAZELLE}: A low latency framework for secure neural network inference. In: 27th {USENIX} Security Symposium ({USENIX} Security 18), pp. 1651–1669 (2018)
28. Kairouz, P., Oh, S., Viswanath, P.: Extremal mechanisms for local differential privacy. The Journal of Machine Learning Research **17**(1), 492–542 (2016)
29. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? SIAM Journal on Computing **40**(3), 793–826 (2011)
30. Kotz, S., Kozubowski, T., Podgorski, K.: The laplace distribution and generalizations (2001)
31. LeCun, Y.: The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (1998)
32. Lou, Q., Feng, B., Charles Fox, G., Jiang, L.: Glyph: Fast and accurately training deep neural networks on encrypted data. Advances in Neural Information Processing Systems **33** (2020)
33. McMahan, H.B., Moore, E., Ramage, D., y Arcas, B.A.: Federated learning of deep networks using model averaging (2016)
34. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
35. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data (2016)
36. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, U.: Scalable private learning with pate (2018)
37. Parliament, E., Council, E.: Regulation (eu) 2016/679 of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. Tech. rep., European Parliament and European Council (2016)
38. Rastogi, V., Nath, S.: Differentially private aggregation of distributed time-series with transformation and encryption. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 735–746 (2010)
39. Ryffel, T., Pointcheval, D., Bach, F.: Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. arXiv preprint arXiv:2006.04593 (2020)
40. Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D., Passerat-Palmbach, J.: A generic framework for privacy preserving deep learning. arXiv preprint arXiv:1811.04017 (2018)
41. Sabater, C., Bellet, A., Ramon, J.: Distributed differentially private averaging with improved utility and robustness to malicious parties. arXiv preprint arXiv:2006.07218 (2020)
42. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. arXiv preprint arXiv:1606.03498 (2016)
43. Shi, E., Chan, T.H., Rieffel, E., Chow, R., Song, D.: Privacy-preserving aggregation of time-series data. In: Proc. NDSS, vol. 2, pp. 1–17. Citeseer (2011)
44. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp. 1310–1321 (2015)
45. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
46. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: 25th {USENIX} Security Symposium ({USENIX} Security 16), pp. 601–618 (2016)
47. Ullman, J.: Tight lower bounds for locally differentially private selection. arXiv preprint arXiv:1802.02638 (2018)
48. Wang, B., Gong, N.Z.: Stealing hyperparameters in machine learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp. 36–52. IEEE (2018)
49. Wu, X., Fredrikson, M., Jha, S., Naughton, J.F.: A methodology for formalizing model-inversion attacks. In: 2016 IEEE 29th Computer Security Foundations Symposium (CSF), pp. 355–370. IEEE (2016)
50. Yan, M., Fletcher, C.W., Torrellas, J.: Cache telepathy: Leveraging shared resource attacks to learn DNN architectures. CoRR **abs/1808.04761** (2018)
51. Zuber, M., Carpov, S., Sirdey, R.: Towards real-time hidden speaker recognition by means of fully homomorphic encryption. In: International Conference on Information and Communications Security, pp. 403–421. Springer (2020)

Springer Machine Learning manuscript No.
(will be inserted by the editor)

SPEED: Secure, PrivatE, and Efficient Deep learning
Appendix

Arnaud Grivet Sébert · Rafaël Pinot · Martin Zuber · Cédric Gouy-Pailler · Renaud Sirdey

Received: date / Accepted: date

A DP analysis of the learning procedure

In this section, we describe the procedure that computes the overall DP guarantees of the student model learning stage. We summarise this procedure in Section A.1, and demonstrate the theorems we use in Sections A.2 and A.4.

We call $\mathcal{A}$ the aggregation mechanism that outputs the argmax of the noisy counts. $\mathcal{A}(d, Q)$ is the output of $\mathcal{A}$ for the database d and the query Q.

Let $\gamma \in \mathbb{R}_+^*$ be the inverse scale parameter of the distributed noise. Considering the DP guarantees from the point of view of an entity $\mathcal{E}$, let $\tau \in (0, 1)$ be the ratio of the teachers whose noise is ignored by $\mathcal{E}$. Typically, from the point of view of a colluding teacher, τ is the ratio of the teachers who do not collude.

A.1 Analysis algorithm

Let us suppose that for every query Q from the student model, we have a privacy guarantee using Theorem 2 and that we can upperbound the probability $\mathbb{P}[\mathcal{A}(d; Q) \neq k^*]$ that $\mathcal{A}$ outputs some specific output k^* (in practice we choose k^* to be the unnoisy argmax). Then, Theorem 3 gives us an upperbound on the moments accountant per query[1]. The computation of these building blocks is detailed in Sections A.2 and A.4, and the procedure is summarised in Algorithm 1.

Let us recall the definition of the moments accountant.

Definition 5 *The* moments accountant *of a mechanism* $\mathcal{M}$ *is defined for any* $l \in \mathbb{R}_+^*$ *as*

$$\alpha_{\mathcal{M}}(l) := \max_{\text{aux}, d, d'} \alpha_{\mathcal{M}}(l; \text{aux}, d, d')$$

Arnaud Grivet Sébert, Rafaël Pinot, Martin Zuber, Cédric Gouy-Pailler, Renaud Sirdey
Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France

Rafaël Pinot
Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France

[1] Note that only the third value over which the minimum is taken in Theorem 3 is data-dependent and, as such, requires this upperbound of $\mathbb{P}[\mathcal{A}(d; Q) \neq k^*]$.

where the maximum is taken over any auxiliary input aux *and any pair of adjacent databases* d, d' *and* $\alpha_{\mathcal{M}}(l; \text{aux}, d, d') := \log\left(\mathbb{E}\left[\exp(lC(\mathcal{M}, \text{aux}, d, d'))\right]\right)$ *is the moment generating function of the privacy loss random variable.*

Theorem 3 ([8]) *Let* $\epsilon, l \in \mathbb{R}_+^*$. *Let* $\mathcal{M}$ *be a* $(\epsilon, 0)$*-differentially private mechanism and* $q \geq \mathbb{P}[\mathcal{M}(d) \neq k^*]$ *for some outcome* k^*. *If* $q < \frac{e^\epsilon - 1}{e^{2\epsilon} - 1}$, *then for any* aux *and any pair* d, d' *of adjacent databases,* $\mathcal{M}$ *satisfies*

$$\alpha_{\mathcal{M}}(l; \text{aux}, d, d') \leq \min\left(\epsilon l, \frac{\epsilon^2 l(l+1)}{2}, \log\left((1-q)\left(\frac{1-q}{1-e^\epsilon q}\right)^l + qe^{\epsilon l}\right)\right).$$

3

Using the moments accountant per query, we evaluate the overall moments accountant by composability, applying the following theorem from [1].

Theorem 4 ([1]) *Let* $p \in \mathbb{N}^*$. *Let us consider a mechanism* $\mathcal{M}$ *defined on a set* $\mathcal{D}$ *that consists of a sequence of adaptive mechanisms* $\mathcal{M}_1, \ldots, \mathcal{M}_p$ *where, for any* $i \in [p]$, $\mathcal{M}_i \colon \prod_{j=1}^{i-1} \mathcal{R}_j \times \mathcal{D} \mapsto \mathcal{R}_i$. *Then, for any* $l \in \mathbb{R}_+^*$,

$$\alpha_{\mathcal{M}}(l) \leq \sum_{i=1}^{p} \alpha_{\mathcal{M}_i}(l).$$

Finally, parameter δ being chosen, the privacy guarantee is derived from the overall moments accountant applying the tail bound property, stated in Theorem 5 from [1].

Theorem 5 ([1]) *For any* $\epsilon \in \mathbb{R}_+^*$, *the mechanism* $\mathcal{M}$ *is* (ϵ, δ)*-differentially private for*

$$\delta = \min_{l \in \mathbb{N}^*} \exp(\alpha_{\mathcal{M}}(l) - l\epsilon).$$

Algorithm 1: Algorithm to determine the overall privacy guarantee of the learning procedure

Input : number of teachers n, number of classes K, ratio τ of teachers with secret noise, set of queries $\mathcal{Q}$, unnoisy teachers' counts n_k, inverse noise scale γ, l_{max} [a], δ

Output: ϵ

```
for l in [l_max] do
    α(l) ← 0
    for query Q in 𝒬 do
        Compute the privacy cost of Q and an upperbound of
          ℙ[𝒜(d; Q) ≠ k*];
        Derive the moments accountant α_Q(l) with Theorem 3;
        α(l) ← α(l) + α_Q(l);
    end
    ε(l) ← (α(l) − δ)/l;
end
ε ← min_{l∈[l_max]} ε(l);
```

[a] To determine the DP guarantees presented in the paper, we took $l_{max} = 25$ because it seems empirically that it captures the best moments accountant in every case.

A.2 DP guarantee per query in the BHBC framework

Preliminaries on the generalised Laplace distribution. For every teacher j who did send noise and whose noise is secret, the noise sent by j is distributed as $G_1^{(j)} - G_2^{(j)}$ where $G_1^{(j)}$ and $G_2^{(j)}$ are two i.i.d. random variables with gamma density $u \mapsto \frac{1}{\left(\frac{1}{\gamma}\right)^{\frac{1}{n}}\Gamma\left(\frac{1}{n}\right)} u^{\frac{1}{n}-1}e^{-\gamma u}$ and characteristic function $t \mapsto \left(\frac{1}{1-i\frac{t}{\gamma}}\right)^{\frac{1}{n}}$ (see [6]). Hence, the characteristic function of $G_1^{(j)} - G_2^{(j)}$ is $\psi\colon t \mapsto \left(\frac{1}{1+\left(\frac{t}{\gamma}\right)^2}\right)^{\frac{1}{n}}$. By summing over all the teachers who did send a secret noise, we get a total noise whose characteristic function is $\psi^{\tau n}\colon t \mapsto \left(\frac{1}{1+\left(\frac{t}{\gamma}\right)^2}\right)^{\tau}$. The corresponding moment generating function is $t \mapsto \left(\frac{1}{1-\left(\frac{t}{\gamma}\right)^2}\right)^{\tau}$. According to [7], this is the moment generating function of a generalised Laplace distribution whose density is

$$f_{\gamma,\tau}\colon u \in \mathbb{R}^* \mapsto \begin{cases} \frac{1}{\left(\frac{1}{\gamma}\right)^{2\tau}\Gamma(\tau)^2} e^{\gamma u} \int_u^{+\infty} t^{\tau-1}\left(t-u\right)^{\tau-1} e^{-2\gamma t} dt & \text{if} \quad u > 0 \\ \frac{1}{\left(\frac{1}{\gamma}\right)^{2\tau}\Gamma(\tau)^2} e^{\gamma u} \int_0^{+\infty} t^{\tau-1}\left(t-u\right)^{\tau-1} e^{-2\gamma t} dt & \text{if} \quad u < 0 \end{cases}$$

which is actually

$$\begin{aligned} u \in \mathbb{R}^* \mapsto & \frac{1}{\left(\frac{1}{\gamma}\right)^{2\tau}\Gamma\left(\tau\right)^2} e^{\gamma|u|} \int_{|u|}^{+\infty} t^{\tau-1}\left(t-|u|\right)^{\tau-1} e^{-2\gamma t} dt \\ &= \frac{\gamma^{2\tau-1}}{\Gamma\left(\tau\right)^2} e^{\gamma|u|} \int_0^{+\infty} \left(\frac{v}{\gamma}+|u|\right)^{\tau-1}\left(\frac{v}{\gamma}\right)^{\tau-1} e^{-2(v+\gamma|u|)} dv \\ &\qquad \text{(by the substitution } v = \gamma(t-|u|)) \\ &= L_{\gamma,\tau} e^{-\gamma|u|} I_\tau(\gamma|u|) \end{aligned}$$

where $I_\tau\colon v \in \mathbb{R}_+^* \mapsto \int_0^{+\infty}(x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx$ and $L_{\gamma,\tau} = \frac{\gamma}{\Gamma(\tau)^2}$.

Let us remark that, since $\tau - 1 \leq 0$, I_τ is decreasing on $\mathbb{R}_+^*$.

As a density function, $f_{\gamma,\tau}$ is integrable on $\mathbb{R}$ (it can also be proved using Lemma 4). We call $F_{\gamma,\tau}$ the associated cumulative distribution function:

$$F_{\gamma,\tau}\colon t \in \mathbb{R} \mapsto \int_{-\infty}^{t} f_{\gamma,\tau}(u)du$$

Note that, $\lim_{+\infty} F_{\gamma,\tau} = 1$ and, since $f_{\gamma,\tau}$ is pair, $F_{\gamma,\tau}(0) = \frac{1}{2}$ and

$$\forall t \in \mathbb{R}, F_{\gamma,\tau}(t) + F_{\gamma,\tau}(-t) = 1. \tag{1}$$

If there is no ambiguity on the parameters γ and τ, we will only write f, F, I and L.

Lemma 1 *Let r be a random variable following the generalised Laplace distribution as defined above. Suppose that we can find a function M of γ and τ such that, for any $t \in \mathbb{R}$, $\frac{\mathbb{P}[r \geq t]}{\mathbb{P}[r \geq t+2]} \leq M$.*

Then $\mathcal{A}$ is $(\log(M), 0)$-differentially private.

Proof We will mimic the proof of the privacy guarantee of the report noisy max from [5] (Claim 3.9), but with two key adaptations.

First of all, let us warn that our definition of the adjacence of two databases is different from the one of [5]. Changing one teacher is analogous to changing one individual in the *counting queries* context. This is why the hypotheses must be adapted. Indeed, d and d' being two adjacent databases (in our sense), since at most one teacher will change its vote between d and d', we have the property $|n_k - n'_k| \leq 1$ for any $k \in [K]$ but we do not have the property of monotonicity of the counts used in [5] [2].

The second difference is that, r being a random variable following the generalised Laplace distribution, we have to substitute the classical upperbound $e^{2\gamma}$ (valid for the Laplace distribution) of $\frac{\mathbb{P}[r \geq t]}{\mathbb{P}[r \geq t+2]}$ by M.

We consider a query Q. Let $k_0 \in [K]$.

For any event E, we write $\mathbb{P}[E|r_{-k_0}]$ the probability of E under the condition that the draw from the $(K-1)$-dimensional generalised Laplace distribution, used for all the noisy counts except the k_0-th count, is equal to r_{-k_0}. We now suppose this draw r_{-k_0} fixed.

We define $r^* = \min\{r_{k_0} | \forall k \in [K] \setminus \{k_0\}, n_{k_0} + r_{k_0} \geq n_k + r_k\}$. Note that, whatever is the tie-breaking policy, r_{-k_0} being fixed, k_0 is the output of $\mathcal{A}$ for database d if $r_{k_0} > r^*$ and k_0 is not the output of $\mathcal{A}$ if $r_{k_0} < r^*$. Since $\mathbb{P}[r_{k_0} = r^*] = 0$, we have $\mathbb{P}[\mathcal{A}(d, Q) = k_0 | r_{-k_0}] = \mathbb{P}[r_{k_0} > r^*] = \mathbb{P}[r_{k_0} \geq r^*]$. Moreover, for all $k \in [K] \setminus \{k_0\}$,

$$\begin{aligned} n'_{k_0} + r^* + 2 &\geq n_{k_0} + r^* + 1 && \text{(because } |n_{k_0} - n'_{k_0}| \leq 1) \\ &\geq n_k + r_k + 1 && \text{(by definition of } r^*) \\ &\geq n'_k + r_k && \text{(because } |n_{k_0} - n'_{k_0}| \leq 1) \end{aligned}$$

We deduce that, if $r_{k_0} > r^* + 2$, then k_0 is the output of $\mathcal{A}$ for database d'. Therefore, $\mathbb{P}[\mathcal{A}(d', Q) = k_0 | r_{-k_0}] \geq \mathbb{P}[r_{k_0} > r^* + 2] = \mathbb{P}[r_{k_0} \geq r^* + 2]$.

Since $\mathbb{P}[r_{k_0} \geq r^*] \leq M\mathbb{P}[r_{k_0} \geq r^* + 2]$ by assumption, we can deduce that $\mathbb{P}[\mathcal{A}(d, Q) = k_0 | r_{-k_0}] \leq M\mathbb{P}[\mathcal{A}(d', Q) = k_0 | r_{-k_0}]$. This being true for any draw r_{-k_0}, the law of total probability gives us $\mathbb{P}[\mathcal{A}(d, Q) = k_0] \leq M\mathbb{P}[\mathcal{A}(d', Q) = k_0]$.

As d and d' play perfectly symmetric roles (unlike in the proof of the report noisy max guarantee from [5]), we also have $\mathbb{P}[\mathcal{A}(d', Q) = k_0] \leq M\mathbb{P}[\mathcal{A}(d, Q) = k_0]$. Since this is true for any query Q, we can conclude that $\mathcal{A}$ is $(\log(M), 0)$-differentially private. □

By definition of F, r being a random variable following the generalised Laplace distribution, for all $t \in \mathbb{R}$,

$$\mathbb{P}[r \geq t] = 1 - F(t).$$

[2] We could have consider a database $\tilde{d}$ such that d is adjacent to $\tilde{d}$ and d' is adjacent to $\tilde{d}$ with Dwork's definition. Then we could have applied twice the result of [5] (using M instead of e^{γ} as upper bound of $\frac{\mathbb{P}[r \geq t]}{\mathbb{P}[r \geq t+2]}$ for $(d, \tilde{d})$ and $(\tilde{d}, d')$). Nevertheless, we performed numerical experimentations that make us believe that it would have given worse privacy guarantees than the present result.

Let $a \in \mathbb{R}_+^*$.

In the following, we exhibit upper bounds of $g\colon t \in \mathbb{R} \mapsto \frac{1-F(t)}{1-F(t+a)}$ (Propositions 5 and 6) to derive privacy guarantees for $\mathcal{A}$ (Theorem 2) taking $a = 2$. Let us first state some useful lemmas.

Lemma 2 *Let $\beta \in \mathbb{R}_+$. The application $h\colon v \in \mathbb{R}_+^* \mapsto \frac{I(v)}{I(v+\beta)}$ is decreasing.*

Proof We will prove that h is differentiable and that its derivative is non-positive.

Let $\phi\colon (v,t) \in (\mathbb{R}_+^*)^2 \mapsto (t+v)^{\tau-1} t^{\tau-1} e^{-2\gamma t}$. ϕ has a partial derivative in the first variable and, for all $(v,t) \in (\mathbb{R}_+^*)^2$, $\frac{\partial \phi}{\partial v}(v,t) = (\tau-1)(t+v)^{\tau-2} t^{\tau-1} e^{-2\gamma t}$. ϕ and $\frac{\partial \phi}{\partial v}$ are continuous in both variables.

Let $b \in \mathbb{R}_+^*$. For all $(v,t) \in [b,+\infty) \times \mathbb{R}_+^*$, $|\frac{\partial \phi}{\partial v}(v,t)| \leq \psi(t)$ where $\psi\colon t \in \mathbb{R}_+^* \mapsto (1-\tau)(t+b)^{\tau-2} t^{\tau-1} e^{-2\gamma t}$. ψ is continuous and integrable on $[b,+\infty)$. Applying Leibniz's theorem, we deduce that I is differentiable on $[b,+\infty)$ and that, for all $v \in [b,+\infty)$, $I'(v) = \int_0^{+\infty} (\tau-1)(t+v)^{\tau-2} t^{\tau-1} e^{-2\gamma t} dt$. Since this is true for all $b \in \mathbb{R}_+^*$, we know that I is differentiable on $\mathbb{R}_+^*$ and that, for all $v \in \mathbb{R}_+^*$, $I'(v) = \int_0^{+\infty} (\tau-1)(t+v)^{\tau-2} t^{\tau-1} e^{-2\gamma t} dt$. As a consequence, h is differentiable on $\mathbb{R}_+^*$ and, for all $v \in \mathbb{R}_+^*$, $h'(v) = \frac{I(v+\beta)I'(v) - I(v)I'(v+\beta)}{I(v+\beta)^2}$.

Let $v \in \mathbb{R}_+^*$.

$$
\begin{aligned}
&I(v+\beta)I'(v) - I(v)I'(v+\beta) \\
&= \int_0^{+\infty} (x+v+\beta)^{\tau-1} x^{\tau-1} e^{-2x} dx \times \int_0^{+\infty} (\tau-1)(y+v)^{\tau-2} y^{\tau-1} e^{-2y} dy \\
&\quad - \int_0^{+\infty} (y+v)^{\tau-1} y^{\tau-1} e^{-2y} dy \times \int_0^{+\infty} (\tau-1)(x+v+\beta)^{\tau-2} x^{\tau-1} e^{-2x} dx \\
&= (\tau-1)\left[\int_0^{+\infty} (x+v+\beta)^{\tau-1} x^{\tau-1} e^{-2x} \int_0^{+\infty} (y+v)^{\tau-2} y^{\tau-1} e^{-2y} dy dx \right. \\
&\quad \left. - \int_0^{+\infty} (x+v+\beta)^{\tau-2} x^{\tau-1} e^{-2x} \int_0^{+\infty} (y+v)^{\tau-1} y^{\tau-1} e^{-2y} dy dx\right] \\
&= (\tau-1)\left[\int_0^{+\infty}\int_0^{+\infty} (x+v+\beta)^{\tau-1} (y+v)^{\tau-2} (xy)^{\tau-1} e^{-2(x+y)} dy dx \right. \\
&\quad \left. - \int_0^{+\infty}\int_0^{+\infty} (x+v+\beta)^{\tau-2} (y+v)^{\tau-1} (xy)^{\tau-1} e^{-2(x+y)} dy dx\right] \\
&= (\tau-1)\int_0^{+\infty}\int_0^{+\infty} (xy)^{\tau-1} e^{-2(x+y)} \\
&\qquad \times \left[(x+v+\beta)^{\tau-1}(y+v)^{\tau-2} - (x+v+\beta)^{\tau-2}(y+v)^{\tau-1}\right] dy dx \\
&= (\tau-1)\int_0^{+\infty}\int_0^{+\infty} (x+v+\beta)^{\tau-2}(y+v)^{\tau-2}(xy)^{\tau-1} e^{-2(x+y)} \\
&\qquad \times \left[(x+v+\beta) - (y+v)\right] dy dx \\
&= (\tau-1)\int_0^{+\infty}\int_0^{+\infty} (x+v+\beta)^{\tau-2}(y+v)^{\tau-2}(xy)^{\tau-1} e^{-2(x+y)} \\
&\qquad \times (x+\beta-y)\, dy dx
\end{aligned}
$$

$$\leq (\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x+v+\beta)^{\tau-2}(y+v)^{\tau-2}(xy)^{\tau-1}e^{-2(x+y)}(x-y)dydx \tag{2}$$

(because $\tau-1\leq 0$ and $\beta\geq 0$)

Similarly, we show that

$$\begin{aligned}
&I(v+\beta)I'(v)-I(v)I'(v+\beta)\\
&=\int_0^{+\infty}(y+v+\beta)^{\tau-1}y^{\tau-1}e^{-2y}dy\times\int_0^{+\infty}(\tau-1)(x+v)^{\tau-2}x^{\tau-1}e^{-2x}dx\\
&\quad-\int_0^{+\infty}(x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx\times\int_0^{+\infty}(\tau-1)(y+v+\beta)^{\tau-2}y^{\tau-1}e^{-2y}dy\\
&=(\tau-1)\left[\int_0^{+\infty}(x+v)^{\tau-2}x^{\tau-1}e^{-2x}\int_0^{+\infty}(y+v+\beta)^{\tau-1}y^{\tau-1}e^{-2y}dydx\right.\\
&\quad\left.-\int_0^{+\infty}(x+v)^{\tau-1}x^{\tau-1}e^{-2x}\int_0^{+\infty}(y+v+\beta)^{\tau-2}y^{\tau-1}e^{-2y}dydx\right]\\
&=(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(xy)^{\tau-1}e^{-2(x+y)}\\
&\qquad\times\left[(x+v)^{\tau-2}(y+v+\beta)^{\tau-1}-(x+v)^{\tau-1}(y+v+\beta)^{\tau-2}\right]dydx\\
&=(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x+v)^{\tau-2}(y+v+\beta)^{\tau-2}(xy)^{\tau-1}e^{-2(x+y)}\\
&\qquad\times(y+\beta-x)dydx\\
&\leq(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x+v)^{\tau-2}(y+v+\beta)^{\tau-2}(xy)^{\tau-1}e^{-2(x+y)}(y-x)dydx
\end{aligned} \tag{3}$$

Alternatively, we can use 2 to deduce 3 directly using Fubini's theorem and exchanging the roles of x and y.

From 2 and 3, we get:

$$\begin{aligned}
&2\times\left[I(v+\beta)I'(v)-I(v)I'(v+\beta)\right]\\
&\leq(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x+v+\beta)^{\tau-2}(y+v)^{\tau-2}(x-y)(xy)^{\tau-1}e^{-2(x+y)}dydx\\
&\quad+(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x+v)^{\tau-2}(y+v+\beta)^{\tau-2}(y-x)(xy)^{\tau-1}e^{-2(x+y)}dydx\\
&=(\tau-1)\int_0^{+\infty}\int_0^{+\infty}(x-y)(xy)^{\tau-1}e^{-2(x+y)}\\
&\qquad\times\left[(x+v+\beta)^{\tau-2}(y+v)^{\tau-2}-(x+v)^{\tau-2}(y+v+\beta)^{\tau-2}\right]dydx
\end{aligned}$$

Let $(x,y)\in(\mathbb{R}_+^*)^2$.

Note that $(x+v+\beta)(y+v)-(x+v)(y+v+\beta)=\beta(y-x)$ and then

$$(x+v+\beta)^{\tau-2}(y+v)^{\tau-2}\geq(x+v)^{\tau-2}(y+v+\beta)^{\tau-2}$$

$$\Leftrightarrow (x+v+\beta)(y+v) \leq (x+v)(y+v+\beta) \qquad \text{(because } \tau-2<0\text{)}$$
$$\Leftrightarrow x \geq y.$$

We deduce that

$$\left[(x+v+\beta)^{\tau-2}(y+v)^{\tau-2} - (x+v)^{\tau-2}(y+v+\beta)^{\tau-2}\right](x-y) \geq 0.$$

This inequality being true for all $(x,y) \in (\mathbb{R}_+^*)^2$ and, since $\tau - 1 \leq 0$, we have:

$$(\tau-1)\int_0^{+\infty}\int_0^{+\infty}\left[(x+v+\beta)^{\tau-2}(y+v)^{\tau-2} - (x+v)^{\tau-2}(y+v+\beta)^{\tau-2}\right] \times (x-y)(xy)^{\tau-1}e^{-2(x+y)}dydx \leq 0$$

Finally, $I(v+\beta)I'(v) - I(v)I'(v+\beta) \leq 0$ and $h'(v) \leq 0$.

Since this is true for any $v \in \mathbb{R}_+^*$, we can conclude that h is decreasing on $\mathbb{R}_+^*$. □

Lemma 3 *The function g has a maximum on $\mathbb{R}$, and this maximum is reached in the interval $[-\frac{a}{2}; 0]$.*

Proof Since f is defined on $\mathbb{R}^*$, F is differentiable on $\mathbb{R}^*$. Thus g is differentiable on $\mathbb{R}^* \setminus \{-a\}$ and, for all $t \in \mathbb{R}^* \setminus \{-a\}$,

$$g'(t) = \frac{(1-F(t))f(t+a) - (1-F(t+a))f(t)}{(1-F(t+a))^2}.$$

First of all, let us prove that g is increasing on $(-\infty; -\frac{a}{2})$. For all $t \in (-\infty; -a)$, $|t| = -t \geq -t-a = |t+a|$ and, for all $t \in (-a; -\frac{a}{2})$, $|t| = -t \geq t+a = |t+a|$. Let $t \in (-\infty; -a) \cup (-a; -\frac{a}{2})$. Then, since $x \mapsto e^{-\gamma x}I(\gamma x)$ is decreasing on $\mathbb{R}_+^*$, $e^{-\gamma|t|}I(\gamma|t|) \leq e^{-\gamma|t+a|}I(\gamma|t+a|)$ which means $f(t) \leq f(t+a)$. Besides, F is increasing then, since $a \geq 0$, $1-F(t+a) \leq 1-F(t)$. Since $f(t)$, $f(t+a)$, $1-F(t)$ and $1-F(t)$ are all positive quantities, we deduce that $g'(t) \geq 0$. Then, g is increasing on $(-\infty; -a)$ and on $(-a; -\frac{a}{2})$ and since g is defined and continuous in $-a$, g is increasing on $(-\infty; -\frac{a}{2})$.

Let us now prove that g is decreasing on $\mathbb{R}_+$. Let $t \in \mathbb{R}_+^*$.

$$\begin{aligned}
&\frac{(1-F(t+a))^2}{L^2}g'(t) \\
&= \frac{1}{L^2}\left[(1-F(t))f(t+a) - (1-F(t+a))f(t)\right] \\
&= e^{-\gamma|t+a|}I(\gamma|t+a|)\int_t^{+\infty} e^{-\gamma|u|}I(\gamma|u|)du \\
&\quad - e^{-\gamma|t|}I(\gamma|t|)\int_{t+a}^{+\infty} e^{-\gamma|u|}I(\gamma|u|)du \\
&= e^{-\gamma(t+a)}I(\gamma(t+a))\int_t^{+\infty} e^{-\gamma u}I(\gamma u)du - e^{-\gamma t}I(\gamma t)\int_{t+a}^{+\infty} e^{-\gamma u}I(\gamma u)du \\
&= e^{-\gamma(t+a)}I(\gamma(t+a))\int_t^{+\infty} e^{-\gamma u}I(\gamma u)du
\end{aligned}$$

$$
\begin{aligned}
&- e^{-\gamma t} I(\gamma t) \int_t^{+\infty} e^{-\gamma(v+a)} I(\gamma(v+a)) dv \\
&\qquad \text{(by the substitution } v = u - a\text{)} \\
&= e^{-\gamma(t+a)} \left[\int_t^{+\infty} e^{-\gamma u} \left[I(\gamma(t+a)) I(\gamma u) - I(\gamma t) I(\gamma(u+a)) \right] du \right]
\end{aligned}
$$

For any $u \in [t; +\infty)$, Lemma 2 with $\beta = \gamma a$ tells us that $\frac{I(\gamma u)}{I(\gamma(u+a))} \leq \frac{I(\gamma t)}{I(\gamma(t+a))}$ which means $I(\gamma(t+a))I(\gamma u) - I(\gamma t)I(\gamma(u+a)) \leq 0$.

Therefore, $\int_t^{+\infty} e^{-\gamma u} \left[I(\gamma(t+a))I(\gamma u) - I(\gamma t)I(\gamma(u+a)) \right] du \leq 0$ and finally $g'(t) \leq 0$. This being valid for all $t \in \mathbb{R}_+^*$ and g being continuous in 0, we deduce that g is decreasing on $\mathbb{R}_+$.

From the two previous discussions and from the fact that g is continuous on $[-\frac{a}{2}; 0]$, we conclude that g has a maximum on $\mathbb{R}$ and that this maximum is reached in $[-\frac{a}{2}; 0]$. □

Proposition 5 *For all* $t \in [-\frac{a}{2}; 0]$,

$$
g(t) \leq 1 + 2 \frac{\int_0^{\frac{\gamma a}{2}} e^{-v} I(v) dv}{\int_{\gamma a}^{+\infty} e^{-v} I(v) dv}.
$$

Proof For all $t \in [-\frac{a}{2}; 0]$,

$$
g(t) = 1 + \frac{F(t+a) - F(t)}{1 - F(t+a)}.
$$

Calling $\phi \colon t \in [-\frac{a}{2}; 0] \mapsto F(t+a) - F(t)$, we know that ϕ is differentiable on $[-\frac{a}{2}; 0)$ and that $\phi' \colon t \in [-\frac{a}{2}; 0) \mapsto f(t+a) - f(t)$. Since $x \in \mathbb{R}_+^* \mapsto e^{-x} I(x)$ is decreasing, we have, for all $t \in [-\frac{a}{2}; 0)$,

$$
\begin{aligned}
\phi'(t) \geq 0 &\Leftrightarrow e^{-\gamma|t+a|} I(\gamma|t+a|) \geq e^{-\gamma|t|} I(\gamma|t|) \\
&\Leftrightarrow |t+a| \leq |t| \\
&\Leftrightarrow t + a \leq -t \qquad \text{(because } t + a \geq 0 \text{ and } t \leq 0\text{)} \\
&\Leftrightarrow t \leq -\frac{a}{2}
\end{aligned}
$$

Since ϕ is continuous in 0, we deduce that ϕ is decreasing on $[-\frac{a}{2}; 0]$ and then, for all $t \in [-\frac{a}{2}; 0]$, $F(t+a) - F(t) \leq F(\frac{a}{2}) - F(-\frac{a}{2})$. Moreover, since F is increasing, for all $t \in [-\frac{a}{2}; 0]$, $1 - F(t+a) \geq 1 - F(a)$.

Finally, for all $t \in [-\frac{a}{2}; 0]$,

$$
\begin{aligned}
g(t) &\leq 1 + \frac{F(\frac{a}{2}) - F(-\frac{a}{2})}{1 - F(a)} \\
&= 1 + \frac{L \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-\gamma|u|} I(\gamma|u|) du}{L \int_a^{+\infty} e^{-\gamma|u|} I(\gamma|u|) du} \\
&= 1 + \frac{\frac{L}{\gamma} \int_{-\frac{\gamma a}{2}}^{\frac{\gamma a}{2}} e^{-|v|} I(|v|) dv}{\frac{L}{\gamma} \int_{\gamma a}^{+\infty} e^{-|v|} I(|v|) dv} \qquad \text{(by the substitutions } v = \gamma u\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&= 1 + \frac{\int_{-\frac{\gamma a}{2}}^{0} e^{-|v|} I(|v|)dv + \int_{0}^{\frac{\gamma a}{2}} e^{-|v|} I(|v|)dv}{\int_{\gamma a}^{+\infty} e^{-|v|} I(|v|)dv} \\
&= 1 + \frac{\int_{0}^{\frac{\gamma a}{2}} e^{-|v'|} I(|v'|)dv' + \int_{0}^{\frac{\gamma a}{2}} e^{-|v|} I(|v|)dv}{\int_{\gamma a}^{+\infty} e^{-|v|} I(|v|)dv} \quad \text{(by the substitution } v' = -v\text{)} \\
&= 1 + \frac{2\int_{0}^{\frac{\gamma a}{2}} e^{-|v|} I(|v|)dv}{\int_{\gamma a}^{+\infty} e^{-|v|} I(|v|)dv} \\
&= 1 + 2\frac{\int_{0}^{\frac{\gamma a}{2}} e^{-v} I(v)dv}{\int_{\gamma a}^{+\infty} e^{-v} I(v)dv}
\end{aligned}
$$

□

Proposition 6 *Let us suppose that* $\tau > \frac{1}{2}$.

For all $t \in [-\frac{a}{2}; 0]$,

$$g(t) \leq g(0) - \frac{a}{2} g'(0).$$

with

$$g'(0) = \gamma \frac{\frac{\Gamma(\tau)^2}{2} e^{-\gamma a} I(\gamma a) - I(0) \int_{\gamma a}^{+\infty} e^{-v} I(v)dv}{\left(\int_{\gamma a}^{+\infty} e^{-v} I(v)dv\right)^2}.$$

Proof The result basically comes from the fact that g is concave on $[\text{argmax}(g); 0]$ which we prove hereafter.

From the proof of Lemma 3 we know that g is differentiable on $[-\frac{a}{2}; 0)$ and $g' : t \mapsto \frac{(1-F(t))f(t+a)-(1-F(t+a))f(t)}{(1-F(t+a))^2} = \frac{g(t)f(t+a)-f(t)}{1-F(t+a)}$. In the proof of Lemma 2, we saw that I is differentiable on $\mathbb{R}_+^*$ and thus f is differentiable on $\mathbb{R}_+^*$. Finally, we get that g' is differentiable on $(-a; 0)$ and, for all $t \in (-a; 0)$,

$$
\begin{aligned}
g''(t) = \frac{1}{(1-F(t+a))^2} &\left[(1-F(t+a))[g'(t)f(t+a) + g(t)f'(t+a) - f'(t)]\right. \\
&\left. + f(t+a)[g(t)f(t+a) - f(t)]\right] \\
= \frac{1}{(1-F(t+a))^2} &\left[(1-F(t+a))[g'(t)f(t+a) + g(t)f'(t+a) - f'(t)]\right. \\
&\left. + (1-F(t+a))f(t+a)g'(t)\right] \\
= 2g'(t)\frac{f(t+a)}{1-F(t+a)} &+ \frac{(1-F(t+a))[g(t)f'(t+a) - f'(t)]}{(1-F(t+a))^2} \\
= 2g'(t)\frac{f(t+a)}{1-F(t+a)} &+ \frac{(1-F(t))f'(t+a) - (1-F(t+a))f'(t)}{(1-F(t+a))^2}.
\end{aligned}
$$

Since I' is strictly negative on $\mathbb{R}_+^*$, for all $u < 0$, $f'(u) = L\gamma[e^{\gamma u} I(-\gamma u) - e^{\gamma u} I'(-\gamma u)] > 0$ and, for all $u > 0$, $f'(u) = L\gamma[-e^{-\gamma u} I(\gamma u) + e^{-\gamma u} I'(\gamma u)] < 0$. Then, for all $t \in (-a; 0)$, $f'(t) > 0$ and $f'(t+a) < 0$ and, since $1 - F(t) > 0$ and

$1 - F(t+a) > 0$, $(1-F(t))f'(t+a) < 0$ and $(1-F(t+a))f'(t) > 0$. We deduce that, for all $t \in (-a; 0)$,

$$g''(t) < 2g'(t)\frac{f(t+a)}{1-F(t+a)} + \frac{(1-F(t))f'(t+a)}{(1-F(t+a))^2} \tag{4}$$

where $2\frac{f(t+a)}{1-F(t+a)} > 0$ and $\frac{(1-F(t))f'(t+a)}{(1-F(t+a))^2} < 0$.

According to Lemma 3, g has a maximum, which is reached on $[-\frac{a}{2}; 0]$. Let $t_{max} = \text{argmax}(g)$. If $t_{max} \neq 0$, we can argue that $g'(t_{max}) = 0$ and then, from Inequation 4, g'' is strictly negative on a neighbourhood of t_{max}. This implies that g' is decreasing on a neighbourhood of $(t_{max})^+$ and then strictly negative on a neighbourhood of $(t_{max})^+$.

Removing the assumption that $t_{max} \neq 0$, we need to be slightly more subtle since g' is not differentiable in 0 (because I is not differentiable in 0).

Since $\tau > \frac{1}{2}$, $v \mapsto v^{2\tau-2}e^{-2v}$ is integrable on $\mathbb{R}_+^*$ and we can extend the definition of I to $\mathbb{R}_+$. This implies in particular that F and then g are differentiable on the whole interval $(-a; +\infty)$ (with $g'(0) = \frac{(1-F(0))f(a)-(1-F(a))f(0)}{(1-F(a))^2}$). Then $g'(t_{max}) = 0$ and, from Inequation 4, $\lim\limits_{(t_{max})^+} g'' < \frac{(1-F(t_{max}))f'(t_{max}+a)}{(1-F(t_{max}+a))^2} < 0$. Thus g'' (not defined in 0) is strictly negative on a neighbourhood of $(t_{max})^+$. Then g' is strictly decreasing on a neighbourhood of $(t_{max})^+$ and, by continuity in t_{max}, strictly negative on a neighbourhood of $(t_{max})^+$.

Let us suppose that $g''(t) \geq 0$ for a t in $[t_{max}; 0)$ (trivially false if $t_{max} = 0$ since $[t_{max}; 0)$ is empty in this case). We fix such a t and call it t_0. Then, from Inequation 4, $g'(t_0) > 0$ and we can fix $t_1 = \inf\{t \in [t_{max}; t_0] | g'(t) \geq 0\}$. g' is non-negative on a neighbourhood of $(t_1)^+$ thus $t_1 > t_{max}$. We also know that g' is non-positive on $[t_{max}; t_1)$ by definition of t_1. This implies $g'(t_1) = 0$. Since $g'(t_1) = 0$, from Inequation 4, we know that $g''(t_1) < 0$ and then g' is strictly negative on a neighbourhood of $(t_1)^+$. We get a contradiction so $g''(t) < 0$ for all $t \in [t_{max}; 0)$. We deduce that g' is decreasing on $[t_{max}; 0)$.

Thus, for all $t \in [t_{max}; 0)$, $g'(t) \geq g'(0)$. As a consequence, since $t_{max} \leq 0$, $g(t_{max}) \leq g(0) + t_{max}g'(0)$. Besides, $t_{max} \geq -\frac{a}{2}$ and $g'(0) \leq g'(t_{max}) = 0$, thus $g(t_{max}) \leq g(0) - \frac{a}{2}g'(0)$. Finally, by definition of t_{max}, for all $t \in \mathbb{R}$,

$$g(t) \leq g(0) - \frac{a}{2}g'(0)$$

with

$$\begin{aligned}
g'(0) &= \frac{(1-F(0))f(a) - (1-F(a))f(0)}{(1-F(a))^2} \\
&= \frac{\frac{1}{2}Le^{-\gamma a}I(\gamma a) - L^2 I(0)\int_a^{+\infty} e^{-\gamma u}I(\gamma u)du}{\left(L\int_a^{+\infty} e^{-\gamma u}I(\gamma u)du\right)^2} \\
&= \frac{\frac{1}{2L}e^{-\gamma a}I(\gamma a) - I(0)\int_a^{+\infty} e^{-\gamma u}I(\gamma u)du}{\left(\int_a^{+\infty} e^{-\gamma u}I(\gamma u)du\right)^2} \\
&= \frac{\frac{\Gamma(\tau)^2}{2\gamma}e^{-\gamma a}I(\gamma a) - \frac{1}{\gamma}I(0)\int_{\gamma a}^{+\infty} e^{-v}I(v)dv}{\left(\frac{1}{\gamma}\int_{\gamma a}^{+\infty} e^{-v}I(v)dv\right)^2} \qquad \text{(by the substitutions } v{=}\gamma u\text{)}
\end{aligned}$$

$$= \gamma \frac{\frac{\Gamma(\tau)^2}{2} e^{-\gamma a} I(\gamma a) - I(0) \int_{\gamma a}^{+\infty} e^{-v} I(v) dv}{\left(\int_{\gamma a}^{+\infty} e^{-v} I(v) dv \right)^2}.$$

□

Theorem 2 *The aggregation mechanism $\mathcal{A}$ is $(\epsilon, 0)$-differentially private, with*

$$\epsilon = \log \left(1 + 2 \frac{\int_0^{\gamma} e^{-v} I(v) dv}{\int_{2\gamma}^{+\infty} e^{-v} I(v) dv} \right).$$

Moreover, if $\tau > \frac{1}{2}$, g is differentiable in 0 and $\mathcal{A}$ is $(\epsilon', 0)$-differentially private, with

$$\epsilon' = \min \left[\epsilon, \log \left(g(0) - g'(0) \right) \right].$$

Proof Thanks to Lemma 3, we can use Propositions 5 and 6 to upper bound g, for $a = 2$. We then just have to apply Lemma 1 to conclude. □

Lemma 4 *For all $v \in \mathbb{R}_+^*$, $I(v) \leq v^{\tau - 1} \frac{\Gamma(\tau)}{2^\tau}$.*

Proof Let $v \in \mathbb{R}_+^*$.

$$\begin{aligned}
I(v) &= \int_0^{+\infty} (t+v)^{\tau-1} t^{\tau-1} e^{-2t} dt \\
&\leq v^{\tau-1} \int_0^{+\infty} t^{\tau-1} e^{-2t} dt && \text{(because } \tau - 1 \leq 0\text{)} \\
&= v^{\tau-1} \int_0^{+\infty} \left(\frac{u}{2} \right)^{\tau-1} e^{-u} \frac{du}{2} && \text{(by the substitution } u = 2t\text{)} \\
&= v^{\tau-1} \frac{\Gamma(\tau)}{2^\tau}
\end{aligned}$$

□

Proposition 3 *For all $\tau \in (0,1)$, $\lim_{\gamma \to 0} \left[\log \left(1 + 2 \frac{\int_0^{\gamma} e^{-v} I(v) dv}{\int_{2\gamma}^{+\infty} e^{-v} I(v) dv} \right) \right] = 0$.*

Proof For all $v \in \mathbb{R}_+^*$, $e^{-v} I(v) > 0$ thus, supposing $\gamma \in (0, 1]$, $\int_{2\gamma}^{+\infty} e^{-v} I(v) dv \geq \int_2^{+\infty} e^{-v} I(v) dv > 0$. Therefore, it suffices to prove that $\lim_{\gamma \to 0} \left[\int_0^{\gamma} e^{-v} I(v) dv \right] = 0$ to deduce the announced result.

Applying Lemma 4, we get

$$\begin{aligned}
\int_0^{\gamma} e^{-v} I(v) dv &\leq \frac{\Gamma(\tau)}{2^\tau} \int_0^{\gamma} e^{-v} v^{\tau-1} dv \\
&\leq \frac{\Gamma(\tau)}{2^\tau} \int_0^{\gamma} v^{\tau-1} dv \\
&= \frac{\Gamma(\tau)}{2^\tau} \frac{\gamma^\tau}{\tau}
\end{aligned}$$

which gives $\lim_{\gamma \to 0} \left[\int_0^{\gamma} e^{-v} I(v) dv \right] = 0$. □

Proposition 2 *For all* $\gamma \in \mathbb{R}_+^*$, $\lim_{\tau \to 1} \left[\log\left(g(0) - g'(0)\right)\right] = 2\gamma$.

Proof We use the dominated convergence theorem to determine the limit of f and F when τ approaches 1. Let us suppose in the following that $\tau \in (\frac{3}{4}; 1)$.

First of all, we determine the limit of I and deduce the one of f. Let $v \in \mathbb{R}_+$.

For all $x \in (0; 1]$, $(x+v)^{\tau-1}x^{\tau-1}e^{-2x} \leq x^{2\tau-2}e^{-2x} \leq x^{-\frac{1}{2}}e^{-2x}$. As $x \mapsto x^{-\frac{1}{2}}e^{-2x}$ is integrable on $(0; 1]$, and, for all $x \in (0; 1]$, $\lim_{\tau \to 1}\left[(x+v)^{\tau-1}x^{\tau-1}e^{-2x}\right] = e^{-2x}$, by the dominated convergence theorem we get that $\lim_{\tau \to 1}\left[\int_0^1 (x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx\right] = \int_0^1 e^{-2x}dx$.

Similarly, as, for all $x \in [1; +\infty)$, $(x+v)^{\tau-1}x^{\tau-1}e^{-2x} \leq e^{-2x}$ and $\lim_{\tau \to 1}\left[(x+v)^{\tau-1}x^{\tau-1}e^{-2x}\right] = e^{-2x}$, by the dominated convergence theorem, $\lim_{\tau \to 1}\left[\int_1^{+\infty} (x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx\right] = \int_1^{+\infty} e^{-2x}dx$.

From the two points above, we deduce that

$$\begin{aligned}\lim_{\tau \to 1} I(v) &= \lim_{\tau \to 1}\left[\int_0^1 (x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx + \int_1^{+\infty} (x+v)^{\tau-1}x^{\tau-1}e^{-2x}dx\right] \\ &= \int_0^1 e^{-2x}dx + \int_1^{+\infty} e^{-2x}dx \\ &= \int_0^{+\infty} e^{-2x}dx \\ &= \frac{1}{2}\end{aligned}$$

and, for any $u \in \mathbb{R}$, $\lim_{\tau \to 1} f(u) = \lim_{\tau \to 1}\left[\frac{\gamma}{\Gamma(\tau)^2}e^{-\gamma|u|}I(\gamma|u|)\right] = \frac{1}{2}\gamma e^{-\gamma|u|}$.

Let us now determine the limit of F.

Let $u_0 \in [0; \frac{1}{\gamma}]$ and $u_1 \in [0; \frac{1}{\gamma}]$ such that $u_0 < u_1$. According to Lemma 4, for all $u \in (u_0; u_1]$, $e^{-\gamma u}I(\gamma u) \leq e^{-\gamma u}(\gamma u)^{\tau-1}\frac{\Gamma(\tau)}{2^\tau} \leq e^{-\gamma u}(\gamma u)^{-\frac{1}{4}}\frac{\Gamma(\frac{3}{4})}{2^{\frac{3}{4}}}$ because $\gamma u \leq 1$ and Γ is decreasing on $(0; 1]$. Since $u \mapsto e^{-\gamma u}(\gamma u)^{-\frac{1}{4}}\frac{\Gamma(\frac{3}{4})}{2^{\frac{3}{4}}}$ is integrable on $(u_0; u_1]$ and, for all $u \in (u_0; u_1]$, $\lim_{\tau \to 1}\left[e^{-\gamma u}I(\gamma u)\right] = \frac{e^{-\gamma u}}{2}$, by the dominated convergence theorem, $\lim_{\tau \to 1}\left[\int_{u_0}^{u_1} e^{-\gamma u}I(\gamma u)du\right] = \int_{u_0}^{u_1}\frac{e^{-\gamma u}}{2}du$.

Let $u_0 \in [\frac{1}{\gamma}; +\infty)$ and $u_1 \in [\frac{1}{\gamma}; +\infty) \cup \{+\infty\}$ such that $u_0 < u_1$. Similarly, as, for all $u \in [u_0; u_1)$, $e^{-\gamma u}I(\gamma u) \leq e^{-\gamma u}(\gamma u)^{\tau-1}\frac{\Gamma(\tau)}{2^\tau} \leq e^{-\gamma u}\frac{\Gamma(\frac{3}{4})}{2^{\frac{3}{4}}}$. Since $u \mapsto e^{-\gamma u}\frac{\Gamma(\frac{3}{4})}{2^{\frac{3}{4}}}$ is integrable on $[u_0; u_1)$ and, for all $u \in [u_0; u_1)$, $\lim_{\tau \to 1}\left[e^{-\gamma u}I(\gamma u)\right] = \frac{e^{-\gamma u}}{2}$, by the dominated convergence theorem, $\lim_{\tau \to 1}\left[\int_{u_0}^{u_1} e^{-\gamma u}I(\gamma u)du\right] = \int_{u_0}^{u_1}\frac{e^{-\gamma u}}{2}du$.

We deduce that, whatever are the bounds $u_0 \in [0; +\infty)$ and $u_1 \in [0; +\infty) \cup \{+\infty\}$ with $u_0 < u_1$, $\lim_{\tau \to 1}\left[\int_{u_0}^{u_1} e^{-\gamma u}I(\gamma u)du\right] = \int_{u_0}^{u_1}\frac{e^{-\gamma u}}{2}du$. By substitution, we also have $\lim_{\tau \to 1}\left[\int_{u_0}^{u_1} e^{\gamma u}I(-\gamma u)du\right] = \int_{u_0}^{u_1}\frac{e^{\gamma u}}{2}du$ for any $u_0 \in (-\infty; 0] \cup \{-\infty\}$ and $u_1 \in (-\infty; 0]$ with $u_0 < u_1$.

Finally, for any $u_0 \in (-\infty; 0] \cup \{-\infty\}$ and $u_1 \in [0; +\infty) \cup \{+\infty\}$ such that $u_0 < u_1$, we have $\lim_{\tau \to 1} \left[\int_{u_0}^{u_1} e^{-\gamma|u|} I(\gamma|u|) du\right] = \int_{u_0}^{u_1} \frac{e^{-\gamma|u|}}{2} du$. In particular, for all $z \in \mathbb{R}$,

$$\begin{aligned}
\lim_{\tau \to 1} F(z) &= \lim_{\tau \to 1} (L) \times \int_{-\infty}^{z} \frac{e^{-\gamma|u|}}{2} du \\
&= \gamma \int_{-\infty}^{z} \frac{e^{-\gamma|u|}}{2} du \\
&= \begin{cases} \frac{1}{2} e^{\gamma z} \text{ if } z < 0 \\ 1 - \frac{1}{2} e^{-\gamma z} \text{ if } z \geq 0 \end{cases}
\end{aligned}$$

which is actually the expression of the Laplace cumulative distribution function.

From what precedes we can conclude that, with $a = 2$,

$$\begin{aligned}
&\lim_{\tau \to 1} \left[g(0) - g'(0)\right] \\
&\quad = \lim_{\tau \to 1} \left[\frac{1 - F(0)}{1 - F(2)} - \frac{(1 - F(0)) f(2) - (1 - F(2)) f(0)}{(1 - F(2))^2}\right] \\
&\quad = \frac{\frac{1}{2}}{\frac{1}{2} e^{-2\gamma}} - \frac{1}{2} \frac{\frac{1}{2} \times \frac{1}{2} \gamma e^{-2\gamma} - \frac{1}{2} e^{-2\gamma} \times \frac{1}{2} \gamma}{(\frac{1}{2} e^{-2\gamma})^2} \\
&\quad = e^{2\gamma}
\end{aligned}$$

□

A.3 Influence of the HE layer on the DP guarantee per query

The computation of the homomorphic argmax induces some perturbations on the noisy counts and, as such, could harm the DP guarantees that we just gave. The three kinds of perturbations due to the HE layer are:

- the addition of (Gaussian) noise at the time of TFHE encryption which is inherently probabilistic
- the addition of a constant value A on the noisy counts to ensure that all the noisy counts are positive (with high probability) (see Section B)
- a possible mistake on the argmax if two noisy counts are too close (see Section 6 of the main paper)

While these perturbations can be seen as some postprocessing applied on the clear noisy histogram, they cannot be seen as a postprocessing on the clear noisy argmax on which we showed DP guarantees in Section A.2. Nevertheless, if we can prove that these perturbations consist of an addition of noise on the clear histogram, the upper bound on $\frac{\mathbb{P}[r \geq t]}{\mathbb{P}[r \geq t+2]}$, r being the total noise (generalised Laplace noise and HE perturbations) applied to the histogram of the n_k's, would still hold, leading to the same DP guarantees. The additions of Gaussian noise and constant A at encryption have, by commutativity, the same effect as the addition of a sum of Gaussian noises and nA after summation and they will anyway change the output of the homomorphic argmax with very low probability. However, some further work needs to be done in order to check whether the third kind of perturbation can be simulated as a noise addition on the histogram.

A.4 Upper bound of the probability of a report noisy max mistake

In this subsection, we give an upper bound of the probability that $\mathcal{A}$ outputs a wrong argmax because of the added noise following the generalised Laplace distribution.

Lemma 5 *Let $u_0 \in \mathbb{R}_+$. Let $q \in \left(\frac{1}{1-\tau}; +\infty\right)$ and $p := \frac{1}{1-\frac{1}{q}}$.*

We have

$$\int_{u_0}^{+\infty} e^{-\gamma u} I(\gamma u) du \leq \frac{\Gamma(\tau)}{2^\tau \gamma} \frac{e^{-\gamma u_0}}{p^{\frac{1}{p}}} \frac{(\gamma u_0)^{\tau - 1 + \frac{1}{q}}}{[q(1-\tau)-1]^{\frac{1}{q}}}.$$

Proof Let $u_0 \in \mathbb{R}_+$. Let $(p, q) \in (\mathbb{R}_+^*)^2$ such that $\frac{1}{p} + \frac{1}{q} = 1$ and $q > \frac{1}{1-\tau}$.

$$\begin{aligned}
&\int_{u_0}^{+\infty} e^{-\gamma u} I(\gamma u) du \\
&\quad \leq \frac{\Gamma(\tau)}{2^\tau} \int_{u_0}^{+\infty} e^{-\gamma u} (\gamma u)^{\tau - 1} du && \text{(according to Lemma 4)} \\
&\quad = \frac{\Gamma(\tau)}{2^\tau \gamma} \int_{\gamma u_0}^{+\infty} e^{-v} v^{\tau - 1} dv && \text{(by the substitution } v = \gamma u)
\end{aligned}$$

By assumption, $q > \frac{1}{1-\tau}$ so, since $\tau < 1$, $q(\tau - 1) < -1$ and then $v \in \mathbb{R}_+^* \mapsto v^{q(\tau-1)}$ is integrable in the neighbourhood of $+\infty$. Then we can apply Hölder's inequality in the following manner:

$$\begin{aligned}
&\int_{u_0}^{+\infty} e^{-\gamma u} I(\gamma u) du \\
&\quad \leq \frac{\Gamma(\tau)}{2^\tau \gamma} \left(\int_{\gamma u_0}^{+\infty} e^{-pv} dv \right)^{\frac{1}{p}} \left(\int_{\gamma u_0}^{+\infty} v^{q(\tau-1)} dv \right)^{\frac{1}{q}} \\
&\quad = \frac{\Gamma(\tau)}{2^\tau \gamma} \times \left(\frac{e^{-p\gamma u_0}}{p} \right)^{\frac{1}{p}} \times \left(\frac{-(\gamma u_0)^{q(\tau-1)+1}}{(q(\tau-1)+1)} \right)^{\frac{1}{q}} \\
&\quad = \frac{\Gamma(\tau)}{2^\tau \gamma} \times \frac{e^{-\gamma u_0}}{p^{\frac{1}{p}}} \times \frac{(\gamma u_0)^{\tau - 1 + \frac{1}{q}}}{[q(1-\tau)-1]^{\frac{1}{q}}}
\end{aligned}$$

□

Lemma 6 *Let us consider a query Q. Let $k^* \in [K]$ be the unnoisy argmax (for all $k \in [K]$, $n_{k^*} \geq n_k$). For all $k \in [K]$, we define $\Delta_k := n_{k^*} - n_k \geq 0$. Then, for all $q \in (\frac{1}{1-\tau}; +\infty)$, calling $p := \frac{1}{1-\frac{1}{q}}$,*

$$\mathbb{P}[\mathcal{A}(d, Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma \Delta_k} \left[\frac{1}{2} + \frac{1}{\tau 2^{4\tau - 2 + \frac{1}{q}} \Gamma(\tau)^2} \times \frac{(\gamma \Delta_k)^{2\tau - 1 + \frac{1}{q}}}{p^{\frac{1}{p}} [q(1-\tau)-1]^{\frac{1}{q}}} \right].$$

Proof In the following, we will assume that $\Delta_k > 0$ and the upper bound for $\Delta_k = 0$ is obtained by continuity.

For any $k \in [K]$, let us denote Y_k the random variable following the generalised Laplace distribution generated by the sum of the τn individual noises.

Let $k \in [K]$.

$$\begin{aligned}
&\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*}) \\
&\quad = \mathbb{P}(Y_{k^*} \leq Y_k - \Delta_k) \\
&\quad = \int_{-\infty}^{+\infty} f(t)F(t-\Delta_k)dt \\
&\quad = \int_{-\infty}^{0} f(t)F(t-\Delta_k)dt + \int_{0}^{\Delta_k} f(t)F(t-\Delta_k)dt \\
&\qquad + \int_{\Delta_k}^{+\infty} f(t)F(t-\Delta_k)dt
\end{aligned} \tag{5}$$

We will now upper bound each one of the three above integrals separately. The two extreme integrals can be nicely bounded by decreasing exponentials in Δ_k:

$$\begin{aligned}
&\int_{\Delta_k}^{+\infty} f(t)F(t-\Delta_k)dt \\
&\quad = \int_{0}^{+\infty} f(v+\Delta_k)F(v)dv && \text{(by the substitution } v = t - \Delta_k\text{)} \\
&\quad = L\int_{0}^{+\infty} e^{-\gamma|v+\Delta_k|} I(\gamma|v+\Delta_k|)F(v)dv \\
&\quad = L\int_{0}^{+\infty} e^{-\gamma(v+\Delta_k)} I(\gamma(v+\Delta_k))F(v)dv \\
&\quad = Le^{-\gamma\Delta_k}\int_{0}^{+\infty} e^{-\gamma v} I(\gamma(v+\Delta_k))F(v)dv \\
&\quad \leq Le^{-\gamma\Delta_k}\int_{0}^{+\infty} e^{-\gamma v} I(\gamma v)F(v)dv && \text{(because } I \text{ is decreasing)} \\
&\quad = Le^{-\gamma\Delta_k}\int_{0}^{+\infty} e^{-\gamma|v|} I(\gamma|v|)F(v)dv \\
&\quad = e^{-\gamma\Delta_k}\int_{0}^{+\infty} f(v)F(v)dv \\
&\quad = e^{-\gamma\Delta_k} \times \frac{\lim_{+\infty} F^2 - F(0)^2}{2} \\
&\quad = e^{-\gamma\Delta_k} \times \frac{1-\frac{1}{4}}{2} \\
&\quad = \frac{3}{8}e^{-\gamma\Delta_k}
\end{aligned} \tag{6}$$

and

$$\int_{-\infty}^{0} f(t)F(t-\Delta_k)dt$$

$$
\begin{aligned}
&= L\int_{-\infty}^{0} f(t)\int_{-\infty}^{t-\Delta_k} e^{-\gamma|u|}I(\gamma|u|)dudt \\
&= L\int_{-\infty}^{0} f(t)\int_{-\infty}^{t-\Delta_k} e^{\gamma u}I(-\gamma u)dudt \\
&= L\int_{-\infty}^{0} f(t)\int_{-\infty}^{t} e^{\gamma(v-\Delta_k)}I(\gamma(\Delta_k-v))dudt \\
&\qquad\qquad \text{(by the substitution } v = u+\Delta_k\text{)} \\
&= Le^{-\gamma\Delta_k}\int_{-\infty}^{0} f(t)\int_{-\infty}^{t} e^{\gamma v}I(\gamma(\Delta_k-v))dudt \\
&\leq Le^{-\gamma\Delta_k}\int_{-\infty}^{0} f(t)\int_{-\infty}^{t} e^{\gamma v}I(-\gamma v)dudt \qquad \text{(because } I \text{ is decreasing)} \\
&= Le^{-\gamma\Delta_k}\int_{-\infty}^{0} f(t)\int_{-\infty}^{t} e^{-\gamma|v|}I(\gamma|v|)dudt \\
&= e^{-\gamma\Delta_k}\int_{-\infty}^{0} f(t)F(t)dt \\
&= e^{-\gamma\Delta_k}\times\frac{F(0)^2-\lim_{-\infty}F^2}{2} \\
&= \frac{1}{8}e^{-\gamma\Delta_k}. \qquad (7)
\end{aligned}
$$

As for the middle integral, we have

$$
\begin{aligned}
&\int_0^{\Delta_k} f(t)F(t-\Delta_k)dt \\
&\quad = L\int_0^{\Delta_k} f(t)\int_{-\infty}^{t-\Delta_k} e^{-\gamma|u|}I(\gamma|u|)dudt \\
&\quad = L\int_0^{\Delta_k} f(t)\int_{\Delta_k-t}^{+\infty} e^{-\gamma|v|}I(\gamma|v|)dvdt \qquad \text{(by the substitution } v=-u\text{)} \\
&\quad = L\int_0^{\Delta_k} f(t)\int_{\Delta_k-t}^{+\infty} e^{-\gamma v}I(\gamma v)dvdt
\end{aligned}
$$

Since, for all $t\in[0;\Delta_k]$, $0\leq\Delta_k-t$, we can apply Lemma 5. Let $q\in\left(\frac{1}{1-\tau};+\infty\right)$ and $p=\frac{1}{1-\frac{1}{q}}$. We have, for all $t\in(0;\Delta_k)$, $\int_{\Delta_k-t}^{+\infty}e^{-\gamma v}I(\gamma v)dv\leq\frac{\Gamma(\tau)}{2^\tau\gamma}\times\frac{1}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times e^{-\gamma(\Delta_k-t)}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}$. Since $\tau-1+\frac{1}{q}>-1$, $t\mapsto[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}$ is integrable on a neighbourhood of $(\Delta_k)^-$ and then, since $t\mapsto f(t)e^{-\gamma(\Delta_k-t)}$ is bounded on a neighbourhood of Δ_k, $t\mapsto f(t)e^{-\gamma(\Delta_k-t)}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}$ is integrable on a neighbourhood of $(\Delta_k)^-$.

Thus, we can write

$$\int_0^{\Delta_k} f(t)F(t-\Delta_k)dt$$

$$
\begin{aligned}
&\leq L\frac{\Gamma(\tau)}{2^\tau\gamma}\times\frac{1}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times\int_0^{\Delta_k} f(t)e^{-\gamma(\Delta_k-t)}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}dt\\
&= L^2\frac{\Gamma(\tau)}{2^\tau\gamma}\times\frac{1}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\\
&\qquad\qquad\times\int_0^{\Delta_k} e^{-\gamma|t|}I(\gamma|t|)e^{-\gamma(\Delta_k-t)}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}dt\\
&= \frac{\gamma}{2^\tau\Gamma(\tau)^3}\times\frac{1}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\\
&\qquad\qquad\times\int_0^{\Delta_k} e^{-\gamma t}I(\gamma t)e^{-\gamma(\Delta_k-t)}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}dt\\
&= \frac{e^{-\gamma\Delta_k}}{2^\tau\Gamma(\tau)^3}\times\frac{\gamma}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times\int_0^{\Delta_k} I(\gamma t)[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}dt
\end{aligned}
$$

$t\mapsto(\gamma t)^{\tau-1}$ is integrable on a neighbourhood of 0^+ because $\tau-1>-1$. Therefore, $t\mapsto(\gamma t)^{\tau-1}\frac{\Gamma(\tau)}{2^\tau}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}$ is integrable on $(0;\Delta_k)$ so we can apply Lemma 4:

$$
\begin{aligned}
&\int_0^{\Delta_k} f(t)F(t-\Delta_k)dt\\
&\leq \frac{e^{-\gamma\Delta_k}}{2^\tau\Gamma(\tau)^3}\times\frac{\gamma}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times\int_0^{\Delta_k}(\gamma t)^{\tau-1}\frac{\Gamma(\tau)}{2^\tau}[\gamma(\Delta_k-t)]^{\tau-1+\frac{1}{q}}dt\\
&= \frac{e^{-\gamma\Delta_k}}{2^{2\tau}\Gamma(\tau)^2}\times\frac{\gamma\Delta_k}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times\int_0^1(\gamma\Delta_k u)^{\tau-1}[\gamma(\Delta_k-\Delta_k u)]^{\tau-1+\frac{1}{q}}du\\
&\qquad\qquad\text{(by the substitution } u=\frac{t}{\Delta_k})\\
&= \frac{e^{-\gamma\Delta_k}}{2^{2\tau}\Gamma(\tau)^2}\times\frac{(\gamma\Delta_k)^{2\tau-1+\frac{1}{q}}}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}\times\int_0^1 u^{\tau-1}(1-u)^{\tau-1+\frac{1}{q}}du
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\int_0^1 u^{\tau-1}(1-u)^{\tau-1+\frac{1}{q}}du\\
&= \int_0^{\frac{1}{2}} u^{\tau-1}(1-u)^{\tau-1+\frac{1}{q}}du+\int_{\frac{1}{2}}^1 u^{\tau-1}(1-u)^{\tau-1+\frac{1}{q}}du\\
&\leq \int_0^{\frac{1}{2}} u^{\tau-1}\frac{1}{2^{\tau-1+\frac{1}{q}}}du+\int_{\frac{1}{2}}^1\frac{1}{2^{\tau-1}}(1-u)^{\tau-1+\frac{1}{q}}du\\
&\qquad\qquad\text{(because } \tau-1+\frac{1}{q}<0 \text{ and } \tau-1<0)\\
&= \frac{1}{2^{\tau-1+\frac{1}{q}}}\int_0^{\frac{1}{2}} u^{\tau-1}du+\frac{1}{2^{\tau-1}}\int_0^{\frac{1}{2}} v^{\tau-1+\frac{1}{q}}dv\\
&\qquad\qquad\text{(by the substitution } v=1-u)
\end{aligned}
$$

$$= \frac{1}{2^{\tau-1+\frac{1}{q}}} \times \frac{1}{\tau 2^{\tau}} + \frac{1}{2^{\tau-1}} \times \frac{1}{(\tau+\frac{1}{q})2^{\tau+\frac{1}{q}}}$$
$$= \frac{1}{2^{2\tau-1+\frac{1}{q}}} \left(\frac{1}{\tau} + \frac{1}{\tau+\frac{1}{q}} \right)$$
$$\leq \frac{1}{\tau 2^{2\tau-2+\frac{1}{q}}}$$

Therefore

$$\int_0^{\Delta_k} f(t)F(t-\Delta_k)dt \leq \frac{e^{-\gamma\Delta_k}}{\tau 2^{4\tau-2+\frac{1}{q}}\Gamma(\tau)^2} \times \frac{(\gamma\Delta_k)^{2\tau-1+\frac{1}{q}}}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}}. \tag{8}$$

Using 5, 6, 7 and 8, we get

$$\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*}) \leq e^{-\gamma\Delta_k} \left[\frac{1}{2} + \frac{1}{\tau 2^{4\tau-2+\frac{1}{q}}\Gamma(\tau)^2} \times \frac{(\gamma\Delta_k)^{2\tau-1+\frac{1}{q}}}{p^{\frac{1}{p}}[q(1-\tau)-1]^{\frac{1}{q}}} \right].$$

The overall upper bound for $\mathbb{P}[\mathcal{A}(d;Q) \neq k^*]$ is obtained using the fact that the event $(\mathcal{A}(d;Q) \neq k^*)$ is the union of the events $(n_k + Y_k \geq n_{k^*} + Y_{k^*})$, for $k \in [K] \setminus \{k^*\}$, and then $\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} \mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$. □

Proposition 4 *If* $\tau \in (\frac{1}{2}; 1)$,

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma\Delta_k} \left[\frac{1}{2} + \frac{(\gamma\Delta_k)^{2\tau-1}}{\tau 2^{4\tau-2}\Gamma(\tau)^2} \right].$$

If $\tau \in (0; \frac{1}{2}]$,

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma\Delta_k} \left[\frac{1}{2} + \frac{(\gamma\Delta_k)^{\frac{\tau}{2}}}{\tau 2^{\frac{5}{2}\tau-1}\Gamma(\tau)^2} \times \left(\frac{3}{2}\tau\right)^{\frac{3}{2}\tau} \left(\frac{2}{\tau} - 3\right)^{1-\frac{3}{2}\tau} \right].$$

Proof Let us distinct two cases according to the value of τ.

First case: $\tau > \frac{1}{2}$

Taking the limit when q approaches $+\infty$ in 8 (which actually amounts to substitute $v^{\tau-1}$ by its upper bound $(\gamma u_0)^{\tau-1}$ in the integral $\int_{\gamma u_0}^{+\infty} e^{-v} v^{\tau-1} dv$ of the proof of Lemma 5, without needing Hölder's inequality), we get

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma\Delta_k} \left[\frac{1}{2} + \frac{(\gamma\Delta_k)^{2\tau-1}}{\tau 2^{4\tau-2}\Gamma(\tau)^2} \right]$$

Second case: $\tau \leq \frac{1}{2}$

By convention, if $\tau = \frac{1}{2}$, we have $\frac{1}{1-2\tau} = +\infty$.

We take $q < \frac{1}{1-2\tau}$ (it is possible since $\frac{1}{1-2\tau} > \frac{1}{1-\tau}$) and write $q = \frac{1}{1-2\tau+\epsilon}$, with $0 < \epsilon < \tau$. Then, $\frac{1}{p} = 1 - \frac{1}{q} = 2\tau - \epsilon$ and we get

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*]$$
$$\leq \sum_{k \neq k^*} e^{-\gamma\Delta_k} \left[\frac{1}{2} + \frac{(2\tau-\epsilon)^{2\tau-\epsilon}}{\tau 2^{2\tau-1+\epsilon}\Gamma(\tau)^2} \times \left(\frac{1-2\tau+\epsilon}{\tau-\epsilon}\right)^{1-2\tau+\epsilon} \times (\gamma\Delta_k)^{\epsilon} \right]$$

For example, with $\epsilon = \frac{\tau}{2}$ (i.e. $q = \frac{1}{1-\frac{3}{2}\tau}$), we have

$$\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} e^{-\gamma \Delta_k} \left[\frac{1}{2} + \frac{1}{\tau 2^{\frac{5}{2}\tau - 1} \Gamma(\tau)^2} \times \left(\frac{3}{2}\tau\right)^{\frac{3}{2}\tau} \left(\frac{2}{\tau} - 3\right)^{1-\frac{3}{2}\tau} \times (\gamma \Delta_k)^{\frac{\tau}{2}} \right]$$

□

Note that, whatever is the value of $\tau \in (0;1)$, our upper bound of $\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$ tends to 0 when Δ_k approaches $+\infty$ which follows the intuition that $\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$ tends to 0 when the true argmax k^* has a much higher count than k. The upper bound tends to $\frac{1}{2}$ when Δ_k approaches 0, which is consistent with the actual value of the probability $\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$ when the counts n_{k^*} and n_k are equal.

Similarly, the upper bound tends to 0 when γ tends to $+\infty$ and to $\frac{1}{2}$ when γ approaches 0. These are the expected values of the probability $\mathbb{P}(n_k + Y_k \geq n_{k^*} + Y_{k^*})$ when there is no noise or an infinitely wide noise respectively.

Finally, let us remark that we recover the upper bound $\mathbb{P}[\mathcal{A}(d;Q) \neq k^*] \leq \sum_{k \neq k^*} \frac{2+\gamma\Delta_k}{4e^{\gamma\Delta_k}}$ from [8] (obtained with a centralised Laplace noise) when we consider the limit when τ tends to 1.

Remark. The data-dependent bound $\alpha_{\mathcal{A}}(l; \text{aux}, d, d') \leq \log\left((1-q)\left(\frac{1-q}{1-e^{\epsilon}q}\right)^l + qe^{\epsilon l}\right)$ from Theorem 3 is non-monotonic in γ. This may appear counter-intuitive since a smaller noise (greater γ) usually gives worse privacy guarantees and, as one would expect, a bigger moments accountant. Nevertheless, a smaller noise means that the probability of outputting the true (unnoisy) argmax is closer to 1, which may lower the moments accountant. Indeed, two adjacent databases will both output the true argmax with high probability, giving less chance to an adversary to distinguish them. This non-monotonicity of the data-dependent bound induces the non-monotonicity of the overall privacy cost ϵ. This is illustrated in Figure 2 of the paper on which we can see, however, that choosing a small γ still gives better guarantees.

B FHE argmax implementation details

We implemented the FHE argmax algorithm using the C++ TFHE library [4]. Table 1 presents all of the parameters needed to reproduce our results and build a fully homomorphic argmax scheme using the TFHE library. The first two lines present our values for the standard TFHE parameters: the first line for initial ciphertext encryption; the second line for the two bootstrapping keys we use. Given the parameters that we use here, we achieve a security parameter of 110. We base the security of our scheme on the `lwe-estimator`[3] script. The estimator is based on the work presented in [2] and is consistently kept up to date.

The third line presents parameters that are specific to our implementation. Because of the use of Gamma distributions, the values sent by the teachers can

[3] `https://bitbucket.org/malb/lwe-estimator/raw/HEAD/estimator.py`

Table 1 Parameters for our implementation. The top line presents the overall security (λ), and the parameters for the initial encryption: σ is the Gaussian noise parameter and N is the size of polynomials. In the TFHE encryption scheme, there is a parameter k (different from the one used in this paper) which, in our case, is always equal to 1. The second line presents the parameters needed to create the two bootstrapping keys we are using. For these two lines, we used the notations from [10] and [3]. The third line presents parameters specific to our implementation given the specificities of the data to process. A is the value to add to the ciphertexts before subtracting $n_k + Y_k - n_{k'} - Y_{k'}$ as per the notations in Section 4.3 of the paper. b_i is the modulus with which the values are rescaled at encryption time to obtain values in $[0, 1]$ and to allow for a correct result of the θ computation. $b_\theta^{(1)}$ is the output modulus of the first bootstrapping operation creating the θ values. $b_\theta^{(2)}$ is the output modulus of the second and final bootstrapping operation.

N	σ		
1024	1e−9		
N_b	σ_b	B_g	ℓ
1024	1e-9	64	6
A	b_i	$b_\theta^{(1)}$	$b_\theta^{(2)}$
900	4102	36	4

be negative. This can be an important issue: if a value is negative, then it will be interpreted in the ciphertext space as a very high positive value and the resulting argmax will be wrong. Therefore, after summing the ciphertexts from the teachers, we add a constant value (we can add a clear value to a ciphertext value) A to ensure that the $n_k + Y_k + A$ are all positive before subtraction. We evaluated that, given the parameters of the Gamma distributions used, choosing $A = 900$ gives us less than a 2^{-64} probability of failure: with Y_k following a Laplace distribution (as seen in Section 4 of the paper), then we have $\mathbb{P}(Y_k < -A) < 2^{-64}$. The b_i variable corresponds to the value by which we rescale the cleartexts before encryption. Indeed, the cleartext and ciphertext spaces of the TFHE encryption scheme are both $\mathbb{T} = ([0, 1], +)$. Additionally, for a correct θ computation, we need to have $|\frac{n_k+Y_k-n_{k'}-Y_{k'}}{b_i}| < \frac{1}{2}$, which is true if, for all $k \in [K]$, $\frac{n_k+Y_k+A}{b_i} \in [0, \frac{1}{2})$. Since $\mathbb{P}(Y_k \geq A) < 2^{-64}$ by symmetry, $b_i = 2(n + 2A) = 4100$ (with n the number of teachers) is sufficient to have $|\frac{n_k+Y_k-n_{k'}-Y_{k'}}{b_i}| < \frac{1}{2}$ with high probability. $b_\theta^{(1)}$ is the output modulus of the first bootstrapping operation. It needs to be chosen so that we have $\Theta_k > \frac{1}{2}$ for one and only one k. That k will then be considered the argmax. $b_\theta^{(2)}$ is the modulus for the final bootstrapping operation.

C Detailed experimental settings

In this section, we provide the reader with additional details regarding experimental settings. In order to reproduce experimental results, all necessary source codes are available on `https://github.com/Arnaud-GS/SPEED`.

C.1 Experimental settings for MNIST

Following PATE experimental conditions, we built our framework based on the code repositories[4] accompanying [8]. The teacher models are based on two convolutional layers with max-pooling and one fully connected layer with ReLUs. Code modifications have been performed on the initial repository, and are available on `https://github.com/Arnaud-GS/SPEED`. The execution environment consists in Python 3 and Tensorflow 1.15.0. The batch size, learning rate and max steps parameters have been respectively set to 128, 0.01 and 5000. As stated in [8], this yields an aggregate test-error rate of 93%. A semi-supervised technique proposed in [9] has been used[5], in an execution environment consisting of Python 3 and Theano 0.7. Besides modifications available on `https://github.com/Arnaud-GS/SPEED`, the learning rate and number of epochs have been set to 0.001 and 500 respectively.

C.2 Experimental settings for SVHN

For SVHN, two additional layers have been added to the teacher models which were learned using a node with 8 NVIDIA v100. The batch size, learning rate and max steps parameters have been respectively set to 64, 0.08 and 2000. The student model also uses the improved GAN semi-supervised model, relying on Python 3 and Theano 0.8.2. The learning rate and number of epochs have been set to 0.0003 and 600 respectively.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
2. Albrecht, M.R., Player, R., Scott, S.: On the concrete hardness of learning with errors. Journal of Mathematical Cryptology **9**(3), 169 – 203 (2015)
3. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In: ASIACRYPT, pp. 3–33 (2016)
4. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: TFHE: Fast fully homomorphic encryption library (August 2016). Https://tfhe.github.io/tfhe/
5. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
6. Kotz, S., Kozubowski, T., Podgorski, K.: The laplace distribution and generalizations (2001)
7. Mathai, A.: On noncentral generalized laplacianness of quadratic forms in normal variables. Journal of multivariate analysis **45**(2), 239–246 (1993)
8. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data (2016)
9. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. arXiv preprint arXiv:1606.03498 (2016)
10. Zuber, M., Carpov, S., Sirdey, R.: Towards real-time hidden speaker recognition by means of fully homomorphic encryption. In: International Conference on Information and Communications Security, pp. 403–421. Springer (2020)

[4] https://github.com/tensorflow/privacy/tree/master/research/pate_2017

[5] https://github.com/openai/improved-gan

Noname manuscript No.
(will be inserted by the editor)

SPEED: Secure, PrivatE, and Efficient Deep learning (information sheet)

Arnaud Grivet Sébert · Rafaël Pinot · Martin Zuber · Cédric Gouy-Pailler · Renaud Sirdey

Received: date / Accepted: date

Main claim. This paper proposes a deep learning framework able to deal with strong privacy constraints. Our protocol allows accurate collaborative learning while protecting the sensitive knowledge of each agent (teacher) both from the aggregation server and from any user of the student model. Moreover, differential privacy guarantees are provided even in the case that some of the teachers are malicious in the sense that they collude and share the noise they generate.
Evidence. On a theoretical point of view, we formally proved the upper bound of the privacy cost, as a function of the ratio of teachers who do not collude, and detailed the algorithm used to compute it. To support our theoretical guarantees, we performed experiments on MNIST and SVHN datasets to determine the accuracy and the privacy cost of our framework.
Most related contributions. As mentioned in the paper, our work was inspired by PATE framework from *Semi-supervised knowledge transfer for deep learning from private training data*, Papernot et al. [5] and *Scalable private learning with PATE*, Papernot et al. [6]. Nevertheless, while PATE needs to trust the aggregation server, SPEED addresses the issue of threats from an honest but curious aggregation server - or even beyond - thanks to the HE computation of the aggregation and the argmax and to the use of *distributed* differential privacy. Our notion of adjacence is also larger and leads to more general differential privacy guarantees.

In *Practical secure aggregation for privacy-preserving machine learning* [1], Bonawitz et al. consider the problem of privacy-preserving secure aggregation to apply it to federated learning frameworks. The secure aggregation is performed using one-time pads masking, which, contrary to homomorphic encryption, requires communication and is vulnerable to agents' failures. Fault tolerance is nevertheless achieved, at the price of heavy additional communication and computation. The possibility of using differential privacy is mentioned in the appendix of [1] but the problem of a server which has access to the final model is not treated.

Arnaud Grivet Sébert, Rafaël Pinot, Martin Zuber, Cédric Gouy-Pailler, Renaud Sirdey
Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France

Rafaël Pinot
Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France

Private collaborative neural network learning, Chase et al. [2], makes use of both differential privacy and secure aggregation to perform collaborative gradient descent which requires much more computation and communication than our vote-aggregation framework which takes advantage of the existence of personal models. The employed cryptographic techniques (secret sharing), which do not include homomorphic encryption, increase the communication load between the agents. Moreover, the problem of colluding agents is not addressed.

Appropriate reviewers. We believe that relevant reviewers for our paper would be Slawomir Goryczka, Nicolas Papernot, Keith Bonawitz and Théo Ryffel. As the first author of two comparative studies on secure aggregation using distributed differential privacy [3, 4], Slawomir Goryczka would give very interesting reviews on our work. Since SPEED's architecture is close to PATE's one and our differential privacy analysis was inspired from the one from [5], Nicolas Papernot would review our paper with a deep understanding of the problems we tackled. Due to his attention to the fault tolerance problem in secure aggregation and his interest on the differential privacy answer to distrusted servers, we believe that Keith Bonawitz [1] would also be a very appropriate reviewer. Specialised in privacy-preserving machine learning and cryptography, Théo Ryffel [7, 8] would give more cryptography-oriented and complementary insights on our work.

References

1. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191 (2017)
2. Chase, M., Gilad-Bachrach, R., Laine, K., Lauter, K.E., Rindal, P.: Private collaborative neural network learning. IACR Cryptology ePrint Archive **2017**, 762 (2017)
3. Goryczka, S., Xiong, L.: A comprehensive comparison of multiparty secure additions with differential privacy. IEEE transactions on dependable and secure computing **14**(5), 463–477 (2015)
4. Goryczka, S., Xiong, L., Sunderam, V.: Secure multiparty aggregation with differential privacy: A comparative study. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, pp. 155–163 (2013)
5. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data (2016)
6. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, U.: Scalable private learning with pate (2018)
7. Ryffel, T., Pointcheval, D., Bach, F.: Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. arXiv preprint arXiv:2006.04593 (2020)
8. Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D., Passerat-Palmbach, J.: A generic framework for privacy preserving deep learning. arXiv preprint arXiv:1811.04017 (2018)