Differentially Private Hypothesis Transfer Learning

Yang Wang(✉)[1], Quanquan Gu[2], and Donald Brown[1]

[1] Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA
{yw3xs, deb}@virginia.edu
[2] Department of Computer Science, University of California, Los Angeles, CA, USA
qgu@cs.ucla.edu

Abstract. In recent years, the focus of machine learning has been shifting to the paradigm of transfer learning where the data distribution in the target domain differs from that in the source domain. This is a prevalent setting in real-world classification problems and numerous well-established theoretical results in the classical supervised learning paradigm will break down under this setting. In addition, the increasing privacy protection awareness restricts access to source domain samples and poses new challenges for the development of privacy-preserving transfer learning algorithms. In this paper, we propose a novel differentially private multiple-source hypothesis transfer learning method for logistic regression. The target learner operates on differentially private hypotheses and importance weighting information from the sources to construct informative Gaussian priors for its logistic regression model. By leveraging a publicly available auxiliary data set, the importance weighting information can be used to determine the relationship between the source domain and the target domain without leaking source data privacy. Our approach provides a robust performance boost even when high quality labeled samples are extremely scarce in the target data set. The extensive experiments on two real-world data sets confirm the performance improvement of our approach over several baselines.

Keywords: Differential privacy · Transfer learning.

1 Introduction

In the era of big data, an abundant supply of high quality labeled data is crucial for the success of modern machine learning. But it is both impractical and unnecessary for a single entity, whether it is a tech giant or a powerful government agency, to collect the massive amounts of data single-handedly and perform the analysis in isolation. Collaboration among data centers or crowd-sourced data sets can be truly beneficial in numerous data-driven research areas and industries nowadays. As modern data sets get bigger and more complex, the traditional practice of centralizing data from multiple data owners turns out to be extremely inefficient. Increasing concerns over data privacy also create barriers to data sharing, making it difficult to coordinate large-scale collaborative studies especially when sensitive human subject data (e.g., medical records, financial information) are involved. Instead of moving raw data, a more efficient approach is to exchange the intermediate computation results obtained from distributed training data sets (e.g., gradients [27], likelihood values in MLE [4]) during the machine learning process. However, even the exchange of intermediate results or summary statistics is potentially privacy-breaching as revealed by recent evidence [15,39]. There is a large body of research work addressing this problem in a rigorous privacy-preserving manner, especially under the notion of differential privacy [34,14,36,40].

In this paper, we consider the differentially private distributed machine learning problem under a more realistic assumption – the multiple training data sets are drawn from different distributions and the learned machine learning hypothesis (i.e. classifier or predictive model) will be tested and employed on a data set drawn from another different yet related distribution. In this setting, the training sets are referred to as the "source domains" and the test set is referred to as the "target domain". High quality labeled data in the target domain is usually costly, if not impossible, to collect, making it necessary and desirable to take advantage of the sources in order to build a reliable hypothesis on the target. The task of "transferring" source knowledge to improve target "learning" is known as transfer learning, one of the fundamental problems in statistical learning that is attracting increasing attention from both academia and industry. This phenomenon is prevalent in real-world applications (e.g., robot manipulation [35], visual object recognition [10], sentiment analysis [3]) and solving the problem is important because numerous solid theoretical justification, especially the generalization capability, in traditional supervised learning paradigm will break down under transfer learning. Plenty of research work has been done in this area, as surveyed in [28]. Meanwhile, the unprecedented emphasis on data privacy today brings up a natural question for the researchers:

How to improve the learning of the hypothesis on the target using knowledge of the sources while protecting the data privacy of the sources?

To our best knowledge, limited research work exists on differentially private multiple-source transfer learning. Two challenges emerge when addressing this problem – what to transfer and how to ensure differential privacy. In the transfer learning literature, there are four major approaches – instance re-weighting [22,17,11], feature representation transfer [32], hypothesis transfer [19,38,20] and relational knowledge transfer [26]. In particular, we focus on hypothesis transfer learning, where hypotheses trained on the source domains are utilized to improve the learning of target hypothesis and no access to source data is allowed [19]. To better incorporate the source hypotheses, we also adapt an importance weighting mechanism in [16] to measure the relationship between sources and target by leveraging a public auxiliary unlabeled data set. Based on the source-target relationship, we can determine how much weight to assign to each source hypothesis in the process of constructing informative Gaussian priors for the parameters of the target hypothesis. Furthermore, the enforcement of differential privacy requires an additional layer of privacy protection because unperturbed source domain knowledge may incur privacy leakage risk. In the following sections, we explicitly apply our methodology to binary logistic regression. Our approach will provide end-to-end differential privacy guarantee by perturbing the importance weighting information and the source hypotheses before transferring them to the target. Extensive empirical evaluations on real-world data demonstrate its utility and significant performance lift over several important baselines.

The main contributions of this work can be summarized as follows.

- We propose a novel multiple-source differentially private hypothesis transfer learning algorithm and focus specifically on its application in binary logistic regression. It addresses the notorious problem of insufficient labeled target data by making use of unlabeled data and provides rigorous differential privacy guarantee.
- Compared with previous work, only one round of direct communication between sources and target is required in our computationally efficient one-shot model aggregation, therefore overcoming some known drawbacks (e.g. avoiding the complicated iterative hypothesis training process and privacy budget composition in [12,41], eliminating the need for a trusted third party in [13]).
- The negative impact of unrelated source domains (i.e. negative transfer) is alleviated as the weight assigned to each source hypothesis will be determined by its relationship with the target.
- The non-private version of our method achieves performance comparable to that of the target hypothesis trained with an abundant supply of labeled data. It provides a new perspective for multiple-source transfer learning when privacy is not a concern.

The rest of the paper is organized as follows: background and related work are introduced in Section 2. We then present the methods and algorithms in Section 3. It is followed by empirical evaluation of the method on two real-world data sets in Section 4. Lastly, we conclude the work in Section 5.

2 Background and Related Work

The privacy notion we use is differential privacy, which is a mathematically rigorous definition of data privacy [8].

Definition 1 (Differential Privacy). *A randomized algorithm $\boldsymbol{M}$ (with output space Ω and well-defined probability density $\mathcal{D}$) is (ϵ, δ)-differentially private if for all adjacent data sets $\boldsymbol{S}, \boldsymbol{S}'$ that differ in a single data point $\boldsymbol{s}$ and for all measurable sets $\omega \in \Omega$:*

$$Pr[\boldsymbol{M}(\boldsymbol{S}) \in \omega] \leq \exp(\epsilon) Pr[\boldsymbol{M}(\boldsymbol{S}') \in \omega] + \delta. \tag{1}$$

Differential privacy essentially implies that the existence of any particular data point $\boldsymbol{s}$ in a private data set $\boldsymbol{S}$ cannot be determined by analyzing the output of a differentially private algorithm $\boldsymbol{M}$ applied on $\boldsymbol{S}$. The parameter ϵ is typically known as the privacy budget which quantifies the privacy loss whenever an algorithm is applied on the private data. A differentially private algorithm with smaller ϵ means that it is more private. The other privacy parameter δ represents the probability of the algorithm failing to satisfy differential privacy and is usually set to zero.

Differential privacy can be applied to address the problem of private machine learning. The final outputs or the intermediate computation results of machine learning algorithms can potentially reveal sensitive information of individuals who contribute data to the training set. A popular approach to achieving differential privacy is output perturbation, derived from the sensitivity method in [8]. It works by adding carefully calibrated noises to the parameters of the learned hypothesis before releasing it. In particular, Chaudhuri et al. [6] adapted the sensitivity method into a differentially private regularized empirical risk minimization (ERM) algorithm of which logistic regression is a special case. They proved that adding Laplace noise with scale inversely proportional to the privacy parameter ϵ and the regularization parameter λ to the learned hypothesis provides $(\epsilon, 0)$-differential privacy. Besides, there has been a large body of systematic theoretical work in the differentially private ERM literature deriving error bounds and efficiency [2] and investigating high-dimensional sparse regression problems [18].

More recently, research efforts have been focused on distributed privacy-preserving machine learning where private data sets are collected by multiple parties. One line of research involves exchanging differentially private information (e.g. gradients) among multiple parties during the iterative hypothesis training process [14,36,37,1,40]. An alternative line of work focuses on privacy-preserving model aggregation techniques. Pathak et al. [31] addressed the hypothesis ensemble problem by simple parameter averaging and publishing the aggregated hypothesis in a differentially private manner. Papernot et al. [30] proposed a different perspective – Private Aggregation of Teacher Ensembles (PATE), where private data set is split into disjoint subsets and different "teacher" hypotheses are trained on all subsets. The private teacher ensemble is used to produce labels on auxiliary unlabeled data by noisy voting and a "student" hypothesis is trained on the auxiliary data and released. Hamm et al. [13] explored a similar strategy of transferring local hypotheses and focused on convex loss functions. Their work involves a "trusted entity" who collects unperturbed local hypotheses trained on multiple private data sets and uses them to generate "soft" labels on an auxiliary unlabeled data set. A global hypothesis is then trained on the soft-labeled auxiliary data and output perturbation is used to ensure its differential privacy.

In the limited literature of differentially private transfer learning, the most recent and relevant work to ours are [12,41] which focus on multi-task learning [5], one of the variants of transfer learning. Gupta et al. [12] proposed a differentially private mutli-task learning algorithm using noisy task relationship matrix and developed a novel attribute-wise noise addition scheme. Xie et al. [41] introduced a privacy-preserving proximal gradient algorithm to asynchronously update the hypothesis parameters of the learning tasks. Both work are iterative solutions and require multiple rounds of information exchange between tasks. However, one of the key drawbacks of iterative differentially private methods is that privacy risks accumulate with each iteration. The composition theorem of differential privacy [9] provides a framework to quantify the privacy budget consumption. Given a certain level of total privacy budget, there is a limit on how many iterations can be performed on a specific private data set, which severely affects the utility-privacy trade-off of iterative solutions.

3 The Proposed Methods

Let us assume that there are K sources in the transfer learning system (as shown in Fig. 1), indexed as $k = 1, ..., K$. For the k-th source, we denote the labeled samples as $\boldsymbol{S}_l^k = \{(\boldsymbol{x}_i^k, y_i^k) : 1 \leq i \leq n_l^k\}$ and the unlabeled samples as $\boldsymbol{S}_{ul}^k = \{\boldsymbol{x}_j^k : 1 \leq j \leq n_{ul}^k\}$, where $\boldsymbol{x}_i^k, \boldsymbol{x}_j^k \in \mathbb{R}^d$ are the feature vectors, y_i^k is the corresponding label, and n_l^k, n_{ul}^k are the sizes of $\boldsymbol{S}_l^k$ and $\boldsymbol{S}_{ul}^k$ respectively. The samples in the k-th source are drawn i.i.d. according to a probability distribution $\mathcal{D}^k$. There is also a target data set $\boldsymbol{T}$ with abundant unlabeled samples and very few labeled samples drawn i.i.d. from a different yet related distribution $\mathcal{D}^T$. Following the setting of [24], $\mathcal{D}^T$ is assumed to be a mixture of the source distributions $\mathcal{D}^k$'s. A public data set $\boldsymbol{P}$ of size n^P (not necessarily labeled) is accessible to both the sources and the target serving as an information intermediary.

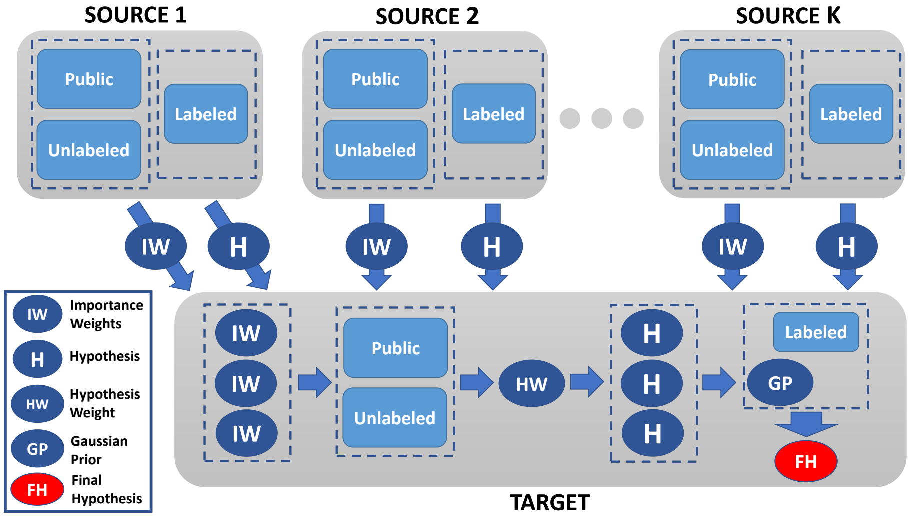


Fig. 1. Diagram of the multiple-source transfer learning system

We present the full details of our method (Algorithm 1) here. Two pieces of knowledge need to be transferred to the target from each source in a differentially private manner. The first piece is the hypothesis locally trained on the labeled source samples $\boldsymbol{S}_l^k$. In order to protect the data privacy, we apply the output perturbation method in [6] to perturb the source hypotheses into $\boldsymbol{\theta}_{priv}^k$ $(k = 1, ..., K)$ before transferring them to the target (Step 1 in Algorithm 1).

The second piece of knowledge is the differentially private "importance weights" vector with respect to samples in the public data set $\boldsymbol{P}$, which is used to measure the relationship between source distribution and target distribution. It is a very difficult task to estimate high-dimensional probability densities and releasing the unperturbed distribution or the histogram can be privacy-breaching in itself [42]. No existing work addresses the differentially private source-target relationship problem explicitly. We leverage a public data set $\boldsymbol{P}$ and propose a novel method based on the importance weighting mechanism in [16] (see Algorithm 2) which was originally developed for the purpose of differentially private data publishing. We take advantage of the "importance weights" vector and successfully quantify the source-target relationship without violating the data privacy of sources. To be more specific, every source k will compute the differentially private "importance weight" for each sample in $\boldsymbol{P}$ using its unlabeled samples $\boldsymbol{S}_{ul}^k$ (Step 2 in Algorithm 1). The "importance weight" of a data point in $\boldsymbol{P}$ is large if it is similar to the samples in $\boldsymbol{S}_{ul}^k$, and small otherwise. The "importance weights" vector $\boldsymbol{w}^k$ is therefore an n^P-dimensional vector with non-negative entries that add up to n^P. By making a multiplicative adjustment to the public data set $\boldsymbol{P}$ using $\boldsymbol{w}^k$, $\boldsymbol{P}$ will look similar to $\boldsymbol{S}_{ul}^k$ in proportion. In other words, $\boldsymbol{w}^k$ is the "recipe" of transforming $\boldsymbol{P}$ into a data set drawn from $\mathcal{D}^k$.

After receiving the "importance weights" vectors $\boldsymbol{w}^k$'s from sources, the target will also compute its own non-private "importance weight" vector $\boldsymbol{w}^T$ (Step 3 in Algorithm 1). These "recipes" bear information about $\mathcal{D}^k$'s and $\mathcal{D}^T$ and are crucial for determining how related each source is to the target. As an intuitive example, the recipe of making regular grape juice out of grapes and water is similar to the recipe of making light grape juice, but very different from that of making wine. By simply comparing the recipes, regular juice can conclude that it is alike light juice but not wine. Given $\boldsymbol{w}^k$'s and $\boldsymbol{w}^T$, the target can further compute the "hypothesis weight" to assign to each source hypothesis by solving an optimization problem that minimizes the divergence between target "importance weights" vector and a linear combination of source "importance weights" vectors (Step 4 in Algorithm 1). The "hypothesis weights" vector $\boldsymbol{w}_H$ is a probability vector of dimension K. The implied assumption here is that hypotheses trained on sources similar to the target should be assigned higher weights in the model aggregation process, whereas the impact of hypotheses trained on source domains unrelated to the target should be reduced. Finally, the target will construct an informative Bayesian prior for its logistic regression model using $\boldsymbol{\theta}_{priv}^k$'s and $\boldsymbol{w}_H$ based on the method in [25] (Step 5-6 in Algorithm 1).

To ensure differential privacy, both pieces of source knowledge transferred to the target need to be perturbed carefully. Note that the private hypothesis is trained on labeled source samples and the private "importance weights" vector is trained on the disjoint unlabeled source samples. Therefore no composition or splitting of privacy budget is involved. The privacy analysis is presented in Theorem 1 below.

Theorem 1. *Algorithm 1 is $(\epsilon, 0)$-differentially private with respect to both the labeled and the unlabeled samples in each source.*

Proof. We sketch the proof here for brevity. First, the perturbed source hypotheses $\boldsymbol{\theta}_{priv}^k$'s are proved to be differentially private with respect to the labeled source samples by [6]. Additionally, the "importance weights" vector $\boldsymbol{w}^k$ is $(\epsilon, 0)$-differentially private with respect to the unlabeled source samples by [16]. Therefore, the final target hypothesis $\boldsymbol{\theta}^T$ built upon $\boldsymbol{\theta}_{priv}^k$'s and $\boldsymbol{w}^k$'s is also $(\epsilon, 0)$-differentially private with respect to the sources by the post-processing guarantee of differential privacy [9].

4 Experiments

In this section we empirically evaluate the effectiveness of our differentially private hypothesis transfer learning method (**DPHTL**) using publicly available real-world data sets. The experiment results show that the hypothesis obtained by our method provides a performance boost over the hypothesis trained on the limited labeled target data while

Algorithm 1 Differentially Private Hypothesis Transfer Learning (DPHTL)

```
Require: $K$ private labeled source data sets $\boldsymbol{S}_l^k$ and unlabeled source data sets $\boldsymbol{S}_{ul}^k$, a public data set $\boldsymbol{P}$, a target data set $\boldsymbol{T}$ with limited labels, privacy parameter $\epsilon$, regularization parameter for importance weighting mechanism $\lambda_{IW}$, regularization parameter for logistic regression model $\lambda_{LR}$,
Ensure: A final target hypothesis $\boldsymbol{\theta}^T \in \mathbb{R}^d$
1: Each source uses its labeled samples $\boldsymbol{S}_l^k$ to train a differentially private logistic regression model $\boldsymbol{\theta}_{priv}^k$ under parameters $\lambda_{LR}$ and $\epsilon$. All the hypotheses $\boldsymbol{\theta}_{priv}^k$ are sent to the target.
2: Each source fetches the public data set $\boldsymbol{P}$, compute the differentially private "importance weights" vector $\boldsymbol{w}^k \leftarrow \textbf{DPIW}(\boldsymbol{S}_{ul}^k, \boldsymbol{P}, \epsilon, \lambda_{IW})$ and sends $\boldsymbol{w}^k$ to the target.
3: The target fetches the public data set $P$ and compute the non-private "importance weights" vector $\boldsymbol{w}^T \leftarrow \textbf{DPIW}(\boldsymbol{T}, \boldsymbol{P}, \infty, \lambda_{IW})$.
4: The target calculates the "hypothesis weights" vector $\boldsymbol{w}_H \in \mathbb{R}^K$ such that the Kullback-Leibler (KL) divergence between $\boldsymbol{w}^T$ and the linear combination of $\boldsymbol{w}^k$ weighted by $\boldsymbol{w}_H$ is minimized:
       $\boldsymbol{w}_H = \underset{\boldsymbol{w} \in \mathbb{R}_{\geq 0}^K, \sum \boldsymbol{w}(k)=1}{\arg\min} \boldsymbol{KL}(\boldsymbol{w}^T, \sum_{k=1}^{K} \boldsymbol{w}(k)\boldsymbol{w}^k) \quad (2)$
5: In order to construct an informative Gaussian prior using $\boldsymbol{w}_H$ and $\boldsymbol{\theta}_{priv}^k$ from the sources, the target first calculates the mean $\boldsymbol{\mu}^T(j)$ and standard deviation $\boldsymbol{\sigma}^T(j)$ of each parameter $\boldsymbol{\theta}^T(j)$ $(j = 1, 2, 3..., d)$ in the target logistic regression hypothesis:
       $\boldsymbol{\mu}^T(j) = \sum_{k=1}^{K} \boldsymbol{w}_H(k)\boldsymbol{\theta}_{priv}^k(j),\ \ \sigma^T(j) = \sqrt{\frac{K}{K-1}\sum_{k=1}^{K} \boldsymbol{w}_H(k)(\boldsymbol{\theta}_{priv}^k(j) - \boldsymbol{\mu}^T(j))^2} \quad (3)$
6: The target trains the Bayesian logistic regression model with the limited labeled target data set and the informative Gaussian prior following the method in [25] and return the posterior parameters $\boldsymbol{\theta}^T$.
```

Algorithm 2 Differentially Private Importance Weights (DPIW) [16]

```
Require: Private data set $\boldsymbol{X}$ of size $N_X$, public data set $\boldsymbol{P}$ of size $N_P$, privacy parameter $\epsilon$, regularization parameter $\lambda$
Ensure: Differentially private importance weights vector $\boldsymbol{w} \in \mathbb{R}^{N_P}$.
1: Label each data point in $\boldsymbol{X}$ as 1 and each data point in $\boldsymbol{P}$ as 0.
2: Fit the regularized logistic regression model on the combination of $\boldsymbol{X}$ and $\boldsymbol{P}$ with the new binary labels:
       $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} - \sum_{\boldsymbol{x} \in \boldsymbol{P}} \frac{\log(p(\boldsymbol{x} \in \boldsymbol{P} | \boldsymbol{x} \in \boldsymbol{X} \cup \boldsymbol{P}))}{N_P}$
       $\qquad - \sum_{\boldsymbol{x} \in \boldsymbol{X}} \frac{\log(p(\boldsymbol{x} \in \boldsymbol{X} | \boldsymbol{x} \in \boldsymbol{X} \cup \boldsymbol{P}))}{N_X} + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \quad (4)$
   where $p(x \in \boldsymbol{X} | \boldsymbol{x} \in \boldsymbol{X} \cup \boldsymbol{P}) = 1 - p(\boldsymbol{x} \in \boldsymbol{P} | \boldsymbol{x} \in \boldsymbol{X} \cup \boldsymbol{P}) = 1/(1 + \exp(-\boldsymbol{\beta}^{\intercal}\boldsymbol{x}))$.
3: Add Laplace noise to $\boldsymbol{\beta}^*$ to get the differentially private $\boldsymbol{\beta}_{priv}$: $\boldsymbol{\beta}_{priv} = \boldsymbol{\beta}^* + \boldsymbol{\delta}$ , where $Pr(\boldsymbol{\delta}) \sim \exp(-\epsilon ||\boldsymbol{\delta}||_2 N_X \lambda / \sqrt{d})$.
4: Output differentially private importance weight $\boldsymbol{w}(\boldsymbol{x}) = \exp(\boldsymbol{\beta}_{priv}^{\intercal}\boldsymbol{x}) N_P / Z$ for each $\boldsymbol{x}$ in $\boldsymbol{P}$, where $Z = \sum_{x \in \boldsymbol{P}} \exp(\boldsymbol{\beta}_{priv}^{\intercal}\boldsymbol{x})$.
```

guaranteeing differential privacy of sources and is also better than the hypotheses obtained by several baselines under various level of privacy requirements.

The first data set is the text classification data set 20NewsGroup (**20NG**)[3] [21], a popular benchmark data set for transfer learning and domain adaptation [33,7,29,23]. It is a collection of approximately 20,000 newsgroup documents with stop words removed and partitioned evenly across 20 different topics. Some topics can be further grouped into a broader subject, for example, computer-related, recreation-related. To construct a binary classification problem, we randomly paired each of the 5 computer-related topics with 2 non-computer-related topics in each source domain. Our purpose is to build a logistic regression model to classify each document as either computer-related or non-computer-related. Specifically, there are 5 source domains with documents sampled from different topic groups as listed in Table 1. The target data set and the public data set are assumed to be mixtures of the source domains.

Table 1. Topics of documents in each source domain for **20NG**

Domain	**Topic group**
Source 1	comp.graphics rec.autos sci.space
Source 2	comp.os.ms-windows.misc talk.politics.guns sci.med
Source 3	comp.sys.ibm.pc.hardware soc.religion.christian talk.politics.mideast
Source 4	comp.sys.mac.hardware misc.forsale rec.sport.baseball
Source 5	comp.windows.x talk.religion.misc sci.crypt

The second data set is the Amazon review sentiment data (**AMAZON**)[4], a famous benchmark multi-domain sentiment data set collected by Blitzer et al [3]. It contains product reviews from 4 different types – *Book*, *DVD*, *Kitchen* and *Electronics*. Reviews with star rating > 3 are labeled positive, and those with star rating < 3 are labeled negative. Each source domain contains reviews from one type and the target domain and public data set are mixtures of the source domains. After initial preprocessing, each product review is represented by its count of unigrams and bigrams in the document and the amounts of positive reviews and negative reviews are balanced. Note that

[3] http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

[4] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

among the four product types, *DVD* and *Book* are similar domains sharing many common sentiment words, whereas *Kitchen* and *Electronics* are more correlated with each other because most of kitchen appliances belong to electronics [43].

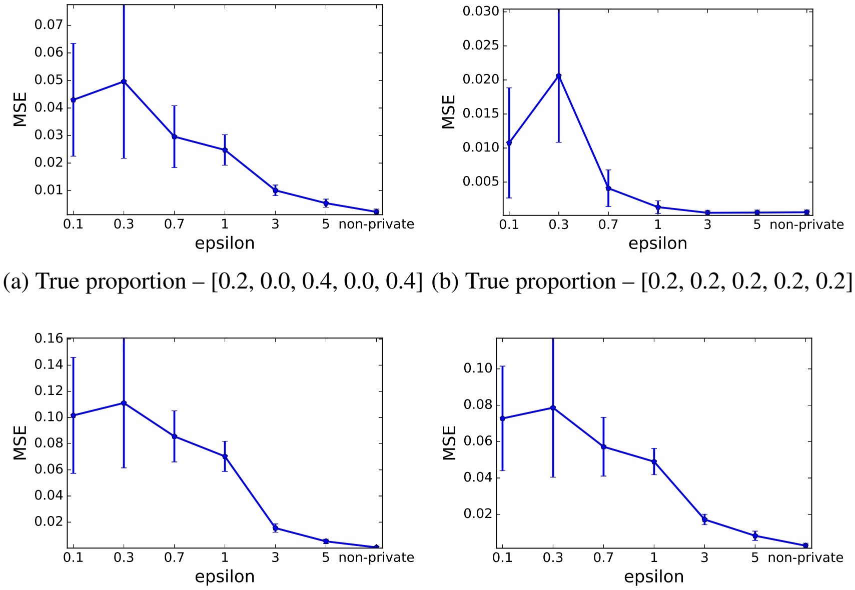
4.1 Data Preprocessing

For the **20NG** data set, we removed the "headers", "footers" and "quotes". When building the vocabulary for each source domain, we ignored terms that have a document frequency strictly higher than 0.5 or strictly lower than 0.005. The total number of terms in the combination of source vocabularies is 5,588. To simulate the transfer learning setup, we sampled 1,000 documents from each source domain as our unlabeled source data set and another 800 documents with labels from each source domain as the labeled source data set. The public data set is an even mixture of the 5 sources with 1,500 documents in total. The target data set is also a mixture of all the source domains with insufficient labels. The final logistic regression model will be tested on a hold-out data set of size 320 sampled from the target domain. For the **AMAZON** data set, we selected the 1,500 most-frequently used unigrams/bigrams from each source domain and combined them to construct the feature set used in the logistic regression model. The total number of features is 2,969. We sampled 1,000 unlabeled reviews from each source domain as the unlabeled source data sets and another 1,000 labeled reviews as the labeled source data sets. The public data set is an even mixture of the 4 product types with 2,000 reviews in total. The counts of features were also transformed to a normalized tf-idf representation. The size of the target data set is 1,500 and the performance will be evaluated on a hold-out test set of size 600 sampled from the target domain. For both data sets, validation sets were reserved for the grid search of regularization parameters. We explored the performance of our method (**DPHTL**) across varying mixtures of sources, varying percentage of target labels and varying privacy requirements.

4.2 Differentially Private Importance Weighting

For the first set of experiments, we study the utility of the differentially private importance weighting mechanism (Step 2 - 4 in Algorithm 1) on both data sets. As shown by the results, our method can indeed reveal the proportion of each source domain in the target domain even under stringent privacy requirements. We set the regularization parameter λ_{IW} to be 0.001 after grid search among several candidates and plot the mean squared error (MSE) between the original proportion vector and the "hypothesis weights" vector determined by our method across varying privacy requirements ϵ. The experiments were repeated 20 times with different samples and the the error bars in the figures represent the standard errors.

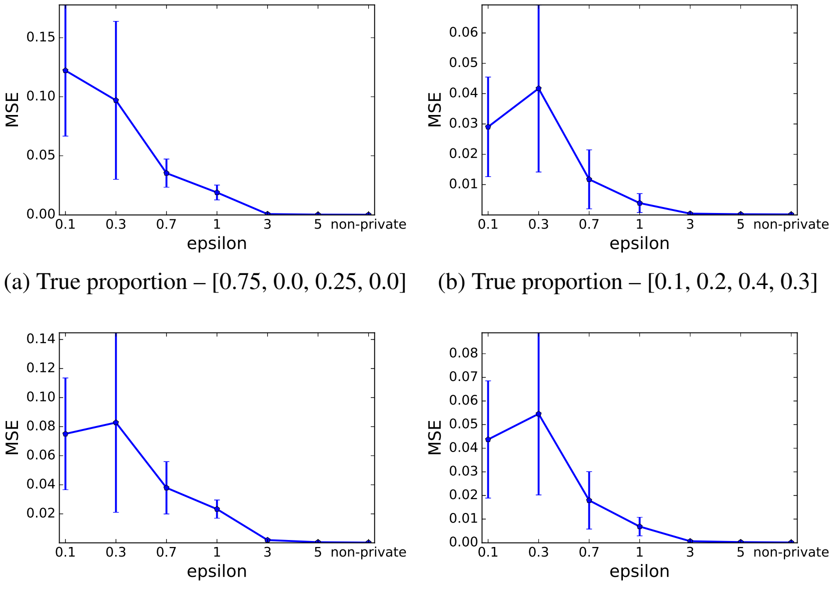
The results (see Fig. 2 and Fig. 3) clearly show that the "hypothesis weights" vector is a good privacy-preserving estimation of the proportion of source domains in the target domain for both **20NG** and **AMAZON**. As the privacy requirement is relaxed (ϵ increases), the MSE will approach zero, reflecting the fact that the weight assigned to each source hypothesis is directly affected by the relationship between the source and the target. It is crucial in avoiding negative transfer as brute-force transfer may actually hurt performance if the sources are too dissimilar to the target [35].



(a) True proportion – [0.2, 0.0, 0.4, 0.0, 0.4] (b) True proportion – [0.2, 0.2, 0.2, 0.2, 0.2]

(c) True proportion – [0.2, 0.0, 0.0, 0.8, 0.0] (d) True proportion – [0.0, 0.0, 0.6, 0.0, 0.4]

Fig. 2. MSE between differentially private hypothesis weight vector and true proportion (**20NG**)



(a) True proportion – [0.75, 0.0, 0.25, 0.0] (b) True proportion – [0.1, 0.2, 0.4, 0.3]

(c) True proportion – [0.0, 0.5, 0.5, 0.0] (d) True proportion – [0.0, 0.2, 0.4, 0.4]

Fig. 3. MSE between differentially private hypothesis weight vector and true proportion (**AMAZON**)

4.3 Differentially Private Hypothesis Transfer Learning

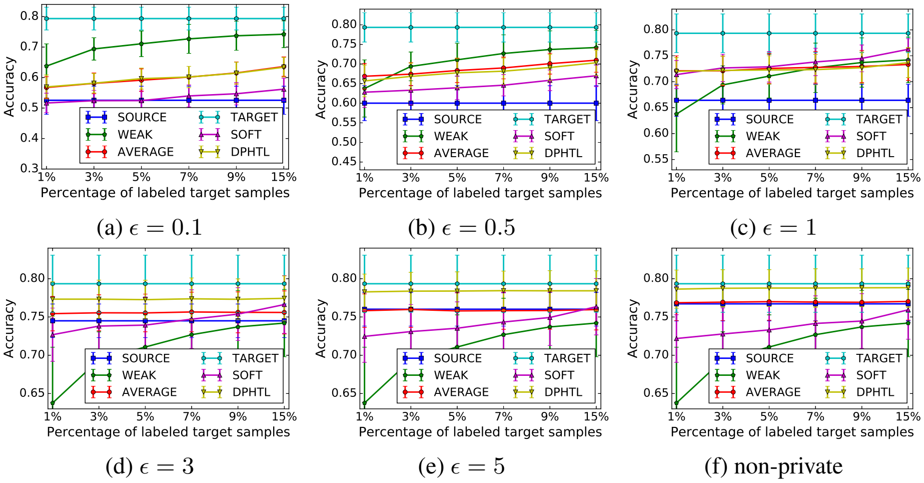
For the second set of experiments, we investigate the privacy-utility trade-off of our **DPHTL** method and compare its performance with those of several baselines. The first baseline **SOURCE** is the best performer on the target test set among all the differentially private source hypotheses. The second baseline **AVERAGE** is the posterior Bayesian logistic regression model which assigns equal weights to all the source hypotheses in constructing the Gaussian prior. The third baseline is referred to as **SOFT**, which represents building the logistic regression model using soft labels. It is an adaptation of the work proposed by Hamm et al. in [13] under our transfer learning setting. More specifically, all the unlabeled samples in the target will be soft-labeled by the fraction $\alpha(\boldsymbol{x})$ of positive votes from all the differentially private source hypotheses (see Algorithm 2 in [13]). For a certain target sample $\boldsymbol{x}$,

$$\alpha(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_{priv}^{k}(\boldsymbol{x}) \tag{5}$$

The minimizer of the loss function of the regularized logistic regression with labeled target samples and soft-labeled target samples will be output as the final hypothesis for the **SOFT** baseline. Similar to our method, Hamm et al. also explored the idea of leveraging auxiliary unlabeled data and proposed a hypothesis ensemble approach. However, there are two key discrepancies. Firstly, no trusted third party is required in our setting so the local hypotheses are perturbed before being transferred to the target. Secondly, the algorithms in [13] were not designed for transfer learning and by assumption the samples are drawn from a distribution common to all parties. Therefore, no mechanism is in place to prevent negative transfer.

In addition, we compared with the non-private hypotheses **WEAK** and **TARGET** trained on the target. The hypothesis **WEAK** is trained using the limited labeled samples in the target data set. The hypothesis **TARGET** is trained using the true labels of all the samples (both labeled and unlabeled) in the target data set and can be considered as the best possible performer since it has access to all the true labels and ignores privacy. The test accuracy of the obtained hypotheses is plotted as a function of the percentage of labeled samples in the target set. The experiments were repeated 20 times and the error bars are standard errors. In order to save space, we illustrate the utility-privacy trade-off by figures for one mixture only.

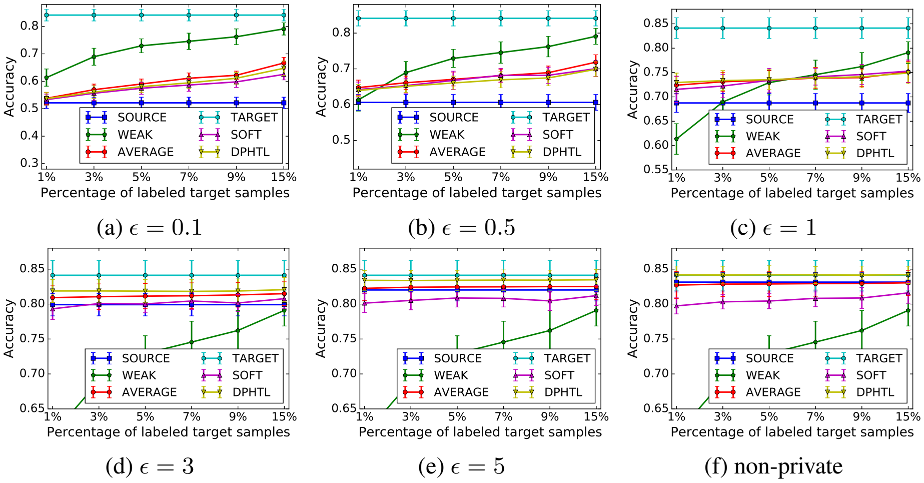
We set the logistic regression regularization parameter at $\lambda_{LR} = 0.003$ for **20NG** and $\lambda_{LR} = 0.005$ for **AMAZON** after preliminary grid search on the validation set. Fig. 4 and 5 show the test accuracy for all the methods across increasing privacy parameter ϵ. **DPHTL** and **AVERAGE** start to emerge as better performers at an intermediate privacy level for both **20NG** and **AMAZON**. When the privacy requirement is further relaxed, **DPHTL** becomes the best performer by a large margin at all range of labeled target sample percentage. Moreover, **DPHTL** is the only method performing as well as **TARGET** when privacy is not a concern. In comparison, the utilities of **AVERAGE** and **SOFT** are hindered by unrelated source domains. More test results are presented in Table 2 when the percentage of labeled target samples is fixed at 5%. In general, **DPHTL** gains a performance improvement over other baselines by taking advantage of



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.5$ (c) $\epsilon = 1$

(d) $\epsilon = 3$ (e) $\epsilon = 5$ (f) non-private

Fig. 4. The test accuracy comparison of **DPHTL** and baselines as a function of percentage of labeled target samples for **20NG**. In the target set, 60% are sampled from source domain 1 and 40% are sampled from source domain 4 .



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.5$ (c) $\epsilon = 1$

(d) $\epsilon = 3$ (e) $\epsilon = 5$ (f) non-private

Fig. 5. The test accuracy comparison of **DPHTL** and baselines as a function of percentage of labeled target samples for **AMAZON**. In the target set, 25% are sampled from source domain 2 and 75% are sampled from source domain 4.

the "hypotheses weights" while preserving the privacy of sources. Besides, it presents a promising solution to the notorious situation where very few labeled data is available at the target domain.

Table 2. The test accuracy on **20NG** and **AMAZON** for different target mixtures when the percentage of labeled target samples is set at 5%. The best performer outside **TARGET** at each privacy level is highlighted.

(a) 20NG

ϵ	WEAK	SOURCE	AVERAGE	SOFT	DPHTL	TARGET
Target mixture 1: $[\mathbf{0.0, 0.6, 0.2, 0.0, 0.2}]$						
0.1	**0.735(0.05)**	0.537(0.05)	0.639(0.03)	0.533(0.03)	0.636(0.03)	0.848(0.02)
0.5	0.735(0.05)	0.651(0.04)	**0.740(0.03)**	0.659(0.03)	0.732(0.04)	0.848(0.02)
1	0.735(0.05)	0.740(0.03)	0.777(0.03)	0.768(0.03)	**0.779(0.03)**	0.848(0.02)
3	0.735(0.05)	0.821(0.03)	0.813(0.02)	0.784(0.03)	**0.832(0.02)**	0.848(0.02)
5	0.735(0.05)	0.832(0.02)	0.819(0.02)	0.777(0.02)	**0.840(0.02)**	0.848(0.02)
∞	0.735(0.05)	0.840(0.02)	0.824(0.02)	0.771(0.03)	**0.853(0.02)**	0.848(0.02)
Target mixture 2: $[\mathbf{0.5, 0.0, 0.0, 0.5, 0.0}]$						
0.1	**0.704(0.04)**	0.540(0.04)	0.588(0.04)	0.533(0.02)	0.585(0.04)	0.792(0.02)
0.5	**0.704(0.04)**	0.602(0.04)	0.671(0.03)	0.628(0.03)	0.664(0.04)	0.792(0.02)
1	0.704(0.04)	0.654(0.04)	0.714(0.02)	**0.725(0.03)**	0.716(0.03)	0.792(0.02)
3	0.704(0.04)	0.745(0.03)	0.747(0.02)	0.734(0.02)	**0.765(0.02)**	0.792(0.02)
5	0.704(0.04)	0.771(0.02)	0.751(0.02)	0.730(0.02)	**0.778(0.01)**	0.792(0.02)
∞	0.704(0.04)	0.781(0.02)	0.764(0.02)	0.728(0.02)	**0.788(0.02)**	0.792(0.02)

(b) AMAZON

ϵ	WEAK	SOURCE	AVERAGE	SOFT	DPHTL	TARGET
Target mixture 1: $[\mathbf{0.25, 0.0, 0.75, 0.0}]$						
0.1	**0.650(0.03)**	0.523(0.03)	0.563(0.03)	0.554(0.02)	0.555(0.03)	0.780(0.02)
0.5	**0.650(0.03)**	0.574(0.03)	0.623(0.02)	0.625(0.03)	0.614(0.03)	0.780(0.02)
1	0.650(0.03)	0.629(0.03)	**0.681(0.02)**	0.676(0.02)	0.670(0.02)	0.780(0.02)
3	0.650(0.03)	0.723(0.02)	0.751(0.02)	0.744(0.02)	**0.755(0.02)**	0.780(0.02)
5	0.650(0.03)	0.753(0.03)	0.767(0.02)	0.754(0.02)	**0.773(0.02)**	0.780(0.02)
∞	0.650(0.03)	0.770(0.02)	0.776(0.02)	0.764(0.02)	**0.788(0.02)**	0.780(0.02)
Target mixture 2: $[\mathbf{0.0, 0.5, 0.0, 0.5}]$						
0.1	**0.718(0.03)**	0.521(0.03)	0.586(0.03)	0.549(0.02)	0.575(0.03)	0.834(0.02)
0.5	**0.718(0.03)**	0.598(0.02)	0.662(0.02)	0.646(0.03)	0.657(0.03)	0.834(0.02)
1	0.718(0.03)	0.672(0.02)	0.729(0.02)	0.726(0.02)	**0.734(0.01)**	0.834(0.02)
3	0.718(0.03)	0.786(0.02)	0.801(0.02)	0.786(0.02)	**0.814(0.02)**	0.834(0.02)
5	0.718(0.03)	0.808(0.02)	0.815(0.01)	0.792(0.02)	**0.829(0.02)**	0.834(0.02)
∞	0.718(0.03)	0.823(0.02)	0.818(0.01)	0.794(0.02)	**0.838(0.02)**	0.834(0.02)

5 Conclusion

This paper proposes a multiple-source hypothesis transfer learning system that protects the differential privacy of sources. By leveraging the relatively abundant supply of unlabeled samples and an auxiliary public data set, we derive the relationship between sources and target in a privacy-preserving manner. Our hypothesis ensemble approach incorporates this relationship information to avoid negative transfer when constructing the Gaussian prior for the target logistic regression model. Moreover, our approach provides a promising and effective solution when the labeled target samples are scarce. Experimental results on benchmark data sets confirm our performance improvement over several baselines from recent work.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. This work is partially supported by a grant from the Army Research Laboratory W911NF-17-2-0110. Quanquan Gu is partly supported by the National Science Foundation SaTC CNS-1717206. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

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. ACM (2016)
2. Bassily, R., Smith, A., Thakurta, A.: Private empirical risk minimization: Efficient algorithms and tight error bounds. In: Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on. pp. 464–473. IEEE (2014)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics. pp. 440–447 (2007)
4. Boker, S.M., Brick, T.R., Pritikin, J.N., Wang, Y., Oertzen, T.v., Brown, D., Lach, J., Estabrook, R., Hunter, M.D., Maes, H.H.: Maintained individual data distributed likelihood estimation (middle). Multivariate behavioral research **50**(6), 706–720 (2015)
5. Caruana, R.: Multitask learning. In: Learning to learn, pp. 95–133. Springer (1998)
6. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**, 1069–1109 (2011)
7. Do, C.B., Ng, A.Y.: Transfer learning for text classification. In: Advances in Neural Information Processing Systems. pp. 299–306 (2006)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography, pp. 265–284. Springer (2006)
9. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2014)
10. Fei-Fei, L.: Knowledge transfer in learning to recognize visual objects classes. In: Proceedings of the International Conference on Development and Learning (ICDL). p. 11 (2006)
11. Garcke, J., Vanck, T.: Importance weighted inductive transfer learning for regression. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 466–481. Springer (2014)

12. Gupta, S.K., Rana, S., Venkatesh, S.: Differentially private multi-task learning. In: Pacific-Asia Workshop on Intelligence and Security Informatics. pp. 101–113. Springer (2016)
13. Hamm, J., Cao, Y., Belkin, M.: Learning privately from multiparty data. In: International Conference on Machine Learning. pp. 555–563 (2016)
14. Hamm, J., Champion, A.C., Chen, G., Belkin, M., Xuan, D.: Crowd-ml: A privacy-preserving learning framework for a crowd of smart devices. In: Distributed Computing Systems (ICDCS), 2015 IEEE 35th International Conference on. pp. 11–20. IEEE (2015)
15. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS genetics **4**(8), e1000167 (2008)
16. Ji, Z., Elkan, C.: Differential privacy based on importance weighting. Machine learning **93**(1), 163–183 (2013)
17. Jiang, J., Zhai, C.: Instance weighting for domain adaptation in nlp. In: Proceedings of the 45th annual meeting of the association of computational linguistics. pp. 264–271 (2007)
18. Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. Journal of Machine Learning Research **1**, 41 (2012)
19. Kuzborskij, I., Orabona, F.: Stability and hypothesis transfer learning. In: Proceedings of The 30th International Conference on Machine Learning. pp. 942–950. ACM (2013)
20. Kuzborskij, I., Orabona, F.: Fast rates by transferring from auxiliary hypotheses. Machine Learning **106**(2), 171–195 (2017)
21. Lang, K.: Newsweeder: Learning to filter netnews. In: Machine Learning Proceedings 1995, pp. 331–339. Elsevier (1995)
22. Liao, X., Xue, Y., Carin, L.: Logistic regression with an auxiliary data source. In: Proceedings of the 22nd International Conference on Machine learning. pp. 505–512. ACM (2005)
23. Lu, Z., Zhu, Y., Pan, S.J., Xiang, E.W., Wang, Y., Yang, Q.: Source free transfer learning for text classification. In: AAAI Conference on Artificial Intelligence. pp. 122–128 (2014)
24. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation with multiple sources. In: Advances in Neural Information Processing Systems. pp. 1041–1048 (2009)
25. Marx, Z., Rosenstein, M.T., Dietterich, T.G., Kaelbling, L.P.: Two algorithms for transfer learning. Inductive transfer: 10 years later (2008)
26. Mihalkova, L., Mooney, R.J.: Transfer learning by mapping with minimal target data. In: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Workshop on Transfer Learning for Complex Tasks (2008)
27. Nedic, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Transactions on Automatic Control **54**(1), 48–61 (2009)
28. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
29. Pan, W., Zhong, E., Yang, Q.: Transfer learning for text mining. In: Mining Text Data, pp. 223–257. Springer (2012)
30. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, U.: Scalable private learning with PATE. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rkZB1XbRZ
31. Pathak, M., Rane, S., Raj, B.: Multiparty differential privacy via aggregation of locally trained classifiers. In: Advances in Neural Information Processing Systems. pp. 1876–1884 (2010)
32. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine learning. pp. 759–766. ACM (2007)

33. Raina, R., Ng, A.Y., Koller, D.: Constructing informative priors using transfer learning. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 713–720. ACM (2006)
34. Rajkumar, A., Agarwal, S.: A differentially private stochastic gradient descent algorithm for multiparty classification. In: International Conference on Artificial Intelligence and Statistics. pp. 933–941 (2012)
35. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer. In: NIPS 2005 workshop on transfer learning. vol. 898, pp. 1–4 (2005)
36. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1310–1321. ACM (2015)
37. Song, S., Chaudhuri, K., Sarwate, A.: Learning from data with heterogeneous noise using sgd. In: Artificial Intelligence and Statistics. pp. 894–902 (2015)
38. Valerio, L., Passarella, A., Conti, M.: Hypothesis transfer learning for efficient data computing in smart cities environments. In: Smart Computing (SMARTCOMP), 2016 IEEE International Conference on. pp. 1–8. IEEE (2016)
39. Wang, R., Li, Y.F., Wang, X., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of the 16th ACM conference on Computer and communications security. pp. 534–544. ACM (2009)
40. Wang, Y., Adams, S., Beling, P., Greenspan, S., Rajagopalan, S., Velez-Rojas, M., Mankovski, S., Boker, S., Brown, D.: Privacy preserving distributed deep learning and its application in credit card fraud detection. In: TrustCom/BigDataSE, 2018. IEEE (2018 (in press))
41. Xie, L., Baytas, I.M., Lin, K., Zhou, J.: Privacy-preserving distributed multi-task learning with asynchronous updates. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1195–1204. ACM (2017)
42. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., Winslett, M.: Differentially private histogram publication. The VLDB Journal **22**(6), 797–822 (2013)
43. Zhang, Y., Yeung, D.Y.: A convex formulation for learning task relationships in multi-task learning. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence. pp. 733–742. AUAI Press (2010)