On Privacy Protection of Latent Dirichlet Allocation Model Training

Fangyuan Zhao[1,2,3], **Xuebin Ren**[1,2], **Shusen Yang**[2,3], **Xinyu Yang**[1,2]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]National Engineering Laboratory for Big Data Analytics, Xi'an Jiaotong University, China
[3]Ministry of Education Key Lab For Intelligent Networks and Network Security, Xi'an Jiaotong University, China
zfy1454236335@stu.xjtu.edu, xuebinren@mail.xjtu.edu, shusenyang@mail.xjtu.edu, yxyphd@mail.xjtu.edu

Abstract

Latent Dirichlet Allocation (LDA) is a popular topic modeling technique for discovery of hidden semantic architecture of text datasets, and plays a fundamental role in many machine learning applications. However, like many other machine learning algorithms, the process of training a LDA model may leak the sensitive information of the training datasets and bring significant privacy risks. To mitigate the privacy issues in LDA, we focus on studying privacy-preserving algorithms of LDA model training in this paper. In particular, we first develop a privacy monitoring algorithm to investigate the privacy guarantee obtained from the inherent randomness of the Collapsed Gibbs Sampling (CGS) process in a typical LDA training algorithm on centralized curated datasets. Then, we further propose a locally private LDA training algorithm on crowdsourced data to provide local differential privacy for individual data contributors. The experimental results on real-world datasets demonstrate the effectiveness of our proposed algorithms.

1 Introduction

Massive text data have arisen in the sustained and rapid development of Internet. Mining and analyzing of text data can help us gain a vast amount of knowledge, thus benefiting the whole society. As a fundamental model for text mining, Latent Dirichlet Allocation(LDA) [Blei *et al.*, 2003] can be used for discovering the main features of the sparse text datasets by identifying their hidden semantic architecture. Particularly, LDA can map the high-dimensional text data to a low-dimensional topic space while retaining the implicit semantics, which has been an effective machine learning technique for clustering or classification. Many enterprises such as Yahoo [Smola and Narayanamurthy, 2010], Tencent [Wang *et al.*, 2014][Yut *et al.*, 2017], and Microsoft [Yuan *et al.*, 2015] have all built LDA platforms for supporting big data analysis and training machine learning models on various text data.

Similar to other machine learning models, LDA may be trained on the datasets that contain some sensitive information of individuals and will inevitably memorize some knowledge about the datasets. Unfortunately, aiming at this characteristic, some attacks have been proposed to extract the private information of the training data from machine learning models. For example, membership inference attacks (MIA)[Shokri *et al.*, 2017] can be launched to infer the membership information of an individual. Model inversion attacks [Fredrikson *et al.*, 2014] have been proved to be able to extract training data from observed model predictions. Therefore, despite the popularity and effectiveness, the naive LDA model may also suffer from these attacks and lead to great privacy risks.

Differential privacy proposed by Dwork [Dwork *et al.*, 2006] has been the de-facto standard of privacy protection with a rigorous mathematical proof. Due to its strong privacy guarantee, DP has also been exploited in many fields such as data publication [Ren *et al.*, 2018][Li *et al.*, 2019]and machine learning [Chaudhuri *et al.*, 2011][Abadi *et al.*, 2016]as well as LDA training [Park *et al.*, 2016][Zhu *et al.*, 2016]. For example, Park et al. [Park *et al.*, 2016] proposed to obtain privacy guarantee for LDA models by perturbing the expected sufficient statistics in each iteration of the variational Bayesian method, which is a parameter estimation algorithm for LDA. Zhu et. al. [Zhu *et al.*, 2016] presented a differentially private LDA algorithm by perturbing the sampling distribution in the collapsed Gibbs sampling(CGS) process, which is a typical training algorithm for LDA.

Both the above algorithms achieve DP by injecting extra noise to the training process of LDA regarding to centralized training datasets. However, as a typical sampling algorithm with inherent randomness, CGS possesses uncertainty in its execution and naturally provides some level of privacy guarantee, which has been indicated in [Wang *et al.*, 2015][Foulds *et al.*, 2016]. In particular, Wang et al. [Wang *et al.*, 2015] proved that posterior sampling and the stochastic gradient Markov chain Monte Carlo techniques possess some inherent privacy guarantee. Foulds et al. [Foulds *et al.*, 2016] further extended this conclusion to the general MCMC methods. Besides the inherent privacy, both existing algorithms consider the LDA model training on centralized datasets owned by a trustworthy data curator. Nevertheless, due to privacy concerns, individual data contributors may be reluctant to directly share their sensitive data but prefer to send the locally sanitized data to the model trainer.

Therefore, aiming to provide strong privacy guarantee for LDA model training, this paper not only investigates to utilize the inherent privacy of CGS in LDA training on centralized datasets, but also proposes a locally private version of LDA that can be trained on crowd-sourced datasets with local sanitations. The contributions are summarized as follows:

- We develop a privacy monitoring algorithm to measure the inherent privacy guarantee of CGS algorithm in LDA. In particular, we first define two different levels of privacy: document level and word level, and present the corresponding lower bound of privacy guarantee after a given number of iterations.
- We propose LP-LDA, a novel mechanism that supports training a LDA model on crowd-sourced datasets with local sanitation, which can provide the guarantee of local privacy for individual data contributors.
- We conduct experiments on several real-world datasets to demonstrate the effectiveness of our proposed algorithms. Particularly, experimental results show that our LP-LDA can achieve a high model training accuracy while providing sufficient local privacy guarantee.

The rest of paper is organized as follows. Section 2 reviews the preliminaries. Section 3 describes our algorithms in detail. The experiments are presented in Section 4. Finally, we conclude the paper in Section 5.

2 Preliminaries

2.1 LDA and Collapsed Gibbs Sampling

LDA model was first proposed by David Blei in 2003 for analyzing the implicit semantic architecture of a corpus. In LDA model, any document m in a corpus D can be described by different distributions on K latent topics, where each topic k can be represented by a distribution on all words. LDA assumes the generative process of the documents in the corpus D as follows:

1. for each topic k, draw a "topic-word" distribution ϕ_k on all words t from Dirichlet($\beta\vec{1}$), where β is the hyperparameter for the Dirichlet priors and can be interpreted as the prior observation for the "topic-word" count.
2. for each document m, draw a "document-topic" distribution θ_m from Dirichlet($\alpha\vec{1}$), where α is a hyperparameter similar to β and represents the prior observation for the "document-topic" count.
3. for each word w in a document m, first draw a topic k from θ_m, and then draw a word t from ϕ_k.

The essence of training a LDA model is to estimate the parameters ϕ_k for a given corpus D.The collapsed Gibbs sampling is such an effective parameter estimation algorithm. It iterates over each word w_i and samples new topic z_i for w_i based on this full conditional distribution

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n_k^t + \beta}{\sum_{t=1}^{V}(n_k^t + \beta)} \cdot \frac{n_m^k + \alpha}{\sum_{k=1}^{K}(n_m^k + \alpha)} \quad (1)$$

where $\neg i$ denotes the whole words except word w_i, n_k^t denotes the count of topic k assigned to word t and n_m^k denotes the count of topic k appeared in document m which are maintained in matrices N_k^t and N_m^k respectively.

After multiple rounds of sampling over the whole corpus, the topic sample of each word can be obtained. And the parameter ϕ_k can be estimated by its posterior expectation.

$$\mathbf{E}[\phi_k^t|\vec{z}, \vec{w}] = \frac{n_k^t + \beta}{\sum_{t=1}^{V}(n_k^t + \beta)}$$

The detailed procedures of CGS can be referred to [Heinrich, 2005].

2.2 Differential Privacy

Differential privacy proposed by Dwork [Dwork *et al.*, 2006] has been the de-facto standard of privacy protection with a rigorous mathematical proof. The rationale of DP guarantee is that negligible information can be gained by manipulating the output of a query on neighboring datasets.

Definition 1. *[Dwork* et al.*, 2006] (Differential Privacy) A randomized mechanism $M : D \to Y$ is ε-differential private if for any neighboring datasets $D, D^{'}$ that satisfying $|D \Delta D'| = 1$ and any output $S \subseteq Y$:*

$$Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D^{'}) \in S]$$

2.3 Local Privacy

Differential privacy implicitly assumes a centralized dataset owned by a trustworthy curator and does not ensure the privacy guarantee for individual data contributors. Recently, local (differential) privacy has been proposed to provide data sanitization at the individual users' side instead of the central server side.

Definition 2. *[Dwork* et al.*, 2014](Local Privacy)* A randomized function f satisfies ε-local privacy if and only if for any two input tuples t and t' in the domain of f, and for any output t^* of f, there is:

$$Pr[f(t) = t^*] \leq e^{\varepsilon} Pr[f(t') = t^*]$$

3 Our Approach

In this section, we first investigate the inherent privacy of CGS process in LDA for a non-sanitized dataset owned by a trustworthy curator. Then, as a complement to the privacy guarantee of the data acquisition period, a locally private mechanism LP-LDA is presented to realize LDA model training on a sanitized dataset by local users.

3.1 Privacy Monitoring Algorithm

Inherent Privacy of CGS

Generally, DP is achieved on most machine learning algorithms by introducing extra noise or randomness, which will inevitably cause a utility loss of the trained model. However, it has been shown in [Foulds *et al.*, 2016] that some degree of inherent DP can be obtained on Gibbs sampling algorithm for free. This is because each sampling process in Gibbs sampling works in a way the same as an exponential mechanism, which is a classic method to achieve DP. Obviously, as one version of Gibbs sampling, Collapsed Gibbs Sampling naturally inherits this property. Furthermore, such a property can also provide privacy for free in the CGS-based LDA training process. Therefore, aiming to utilize the inherent privacy, we develop a privacy monitoring algorithm to quantify the privacy guarantee of CGS in the LDA training process. In particular, the rationale behind the privacy monitoring algorithm is to find an adequate exponential mechanism for each sampling process in CGS and then accumulate the total privacy guarantee of all exponential mechanisms according to the composition theorem of DP.

Document-level Privacy and Word-level Privacy

This paper considers to provide DP for the individual words and documents in the training corpus for LDA, respectively.

Word-level privacy: Let $D = \{w_1, w_2, ...w_W\}$ denote a corpus with $|D| = W$ words w_i $(i = 1, 2, ..., W)$. Then, its neighboring dataset D' satisfying $|D \Delta D'| = 1$ differs from D by a single word w. Word-level privacy prevents membership inference of individual words of the training corpus from the trained LDA model.

Document-level privacy: Let $D = \{m_1, m_2, ...m_M\}$ denote a corpus with $|D| = M$ documents m_i $(i = 1, 2, ..., M)$. Then, its neighboring dataset D' satisfying $|D \Delta D'| = 1$ differs from D by a single document m. In order to bound the sensitivity, we assume that a single document includes at most N_{max} words. Document-level privacy prevents re-identification of individual documents in the training dataset of LDA, which may be contributed by and associated with individual users.

Inherent Privacy in Each Sampling

To begin with, we show the essence of the intrinsic privacy guarantee in each sampling of CGS in terms of exponential mechanism. Consider the sampling process for word w_i in the nth iteration. Suppose its sampling distribution on K topics is given by $\mathbf{P} = (p_1, p_2, ..., p_K)^\top$, where p_k denotes the probability that topic k is assigned to w_i in this sampling. Then we can rewrite p_k as

$$p_k = e^{\frac{(2\Delta \ln p_k) \ln p_k}{2\Delta \ln p_k}},$$

which could be understood as an output probability of an exponential mechanism $M_E(w_i, u, \mathcal{K})$ that selects the topic $k \in \mathcal{K}$ with probability of p_k. The utility function of $M_E(w_i, u, \mathcal{K})$ is $u(w_i, k) = \ln p_k$ and its sensitivity is $\Delta \ln p_k$. Obviously, $\varepsilon = 2\Delta \ln p_k$ is the intrinsic privacy guarantee of the exponential mechanism $M_E(w_i, u, \mathcal{K})$.

Privacy Monitoring for Each Sampling

Unfortunately, it's intractable to specify an exact value of $2\Delta \ln p_k$ in the execution process of CGS algorithm in LDA due to the complicated architecture of training corpus, hence we attempt to find an upper bound of $2\Delta \ln p_k$ to quantify the privacy guarantee ε.

According to Equation (1), the sampling distribution $\mathbf{P}$ for word $w_i = t$ in D in the n-th iteration could be computed by

$$p_k \propto r_k = \frac{n_k^t + \beta}{\sum_{t=1}^V (n_k^t + \beta)} \cdot \frac{n_m^k + \alpha}{\sum_{k=1}^K (n_m^k + \alpha)}$$

Suppose that $\mathbf{P}' = (p'_1, p'_2, ..., p'_K)^\top$ is the corresponding distribution on D', which is the neighboring dataset of D, then

$$p'_k \propto r'_k = \frac{n_k^t + \beta}{\sum_{t=1}^V (n_k^t + \beta) - N_k} \cdot \frac{n_m^k + \alpha}{\sum_{k=1}^K (n_m^k + \alpha)} \quad (2)$$

where N_k denotes the count of topic k assigned in the $D - D'$ where $k \in \{1, 2, ..., K\}$. We refer to $\{N_1, N_2, ..., N_k\}$ as a topic partition on $D - D'$ and $\sum N_k = |D - D'|$.

Given a topic partition $\gamma = \{N_1, N_2, ..., N_k\}$, the privacy guarantee in this sampling process could be measured by

$$\varepsilon_\gamma = \max_{k \in \{1,2,...,K\}} \{2\xi_k\} = \max_{k \in \{1,2,...,K\}} \{2 \ln \frac{p'_k}{p_k}\}$$

where ξ_k denotes the sensitivity of $\ln p_k$. However, there are $\binom{N+K-1}{K-1}$ partitions in total. So, it is computational prohibitive to find the maximal ε_γ among all partitions. In the following, we consider how to reduce the searching space of partitions.

For simplicity, we first consider a special case, in which there exists some topic *i* with $N_i = 0$ in a given partition.

Theorem 1. *Suppose that there exists some $N_k = 0$ in a given partition $\gamma = \{N_1, N_2, ..., N_K\}$, then the privacy guarantee*

$$\varepsilon_\gamma = 2\xi_k = 2 \max\{\xi_1 \, \xi_2 \, ..., \xi_K\} = 2 \ln \frac{\sum_k r'_k}{\sum_k r_k}, \quad (3)$$

if and only if for any $j \neq k$

$$\ln \frac{\sum_{t=1}^V (n_k^t + \beta)}{\sum_{t=1}^V (n_k^t + \beta) - N_j} < 2 \ln \frac{\sum_k r'_k}{\sum_k r_k} \quad (4)$$

Proof. Appendix A of full version [Zhao *et al.*, 2019]. □

Corollary 1. *Suppose that there exists a topic set $\mathcal{T} = \{k, ..., j\}$ with $\{N_j \neq 0, \forall j \in \mathcal{T}\}$ in a given partition γ, and it holds that*

$$\ln \frac{\sum_{t=1}^V (n_k^t + \beta)}{\sum_{t=1}^V (n_k^t + \beta) - N_k} > 2 \ln \frac{\sum_k r'_k}{\sum_k r_k}$$

for some $k \in \mathcal{T}$, then the privacy guarantee

$$\varepsilon_\gamma = 2 \max_{k \in \mathcal{T}} \{\ln(\frac{\sum_k r'_k}{\sum_k r_k} \cdot \frac{r_k}{r'_k})\} > 2 \ln \frac{\sum_k r'_k}{\sum_k r_k} \quad (5)$$

Proof. This proof follows from the result of Theorem 1. □

Theorem 1 and corollary 1 illustrate a special case to find the privacy ε_γ. The following lemma and theorem further demonstrate that among all the partitions, the one with the largest privacy guarantee belongs to a partitions set $\mathcal{P} = \{\gamma | \exists k, s.t.\ N_k = N, N_j = 0, \forall j \neq k\}$.

Lemma 1. *There exists a partition γ^* in $\mathcal{P}$ such that*

$$\sum_{\gamma^*} r'_k = \max_\gamma \{\sum_k r'_k | \gamma \in \Gamma\}$$

where Γ denotes the set consisting of all the partitions.

Proof. Appendix B of full version [Zhao *et al.*, 2019]. □

Algorithm 1 Privacy Monitoring for Each Sampling

Input: word count matrices N_k^t and N_m^k, $N = |D - D'|$ ($D = 1$ for word-level privacy or $D = N_{max}$ for document-level privacy)
Parameter: hyper parameters α, β
Output: privacy guarantee ε

```
compute the sampling distribution P with p_k ∝ r_k = (n_k^t+β)/(Σ_{t=1}^V (n_k^t+β)) · (n_m^k+α)/(Σ_{k=1}^K (n_m^k+α))
compute pseudo sampling distribution q with q_k = (n_k^t+β)/(Σ_{t=1}^V (n_k^t+β) − N) · (n_m^k+α)/(Σ_{k=1}^K (n_m^k+α))
for each component q_k of q do
    compute ξ_k = ln( (Σ_k r'_k)/(Σ_k r_k) · r_k/q_k )
end for
find index k such that |r_k − q_k| = ||r − q||_∞
compute ξ = ln( (Σ_{j≠k} r_j + q_k)/(Σ_j r_j) ) with q_k
return ε = 2 max{ξ_1, ξ_2, ..., ξ_K, ξ}
```

Definition 3. *(Pseudo sampling distribution) Suppose that given a vector* $\mathbf{q}$ *with length K, each component*

$$q_k = \frac{n_k^t + \beta}{\sum_{t=1}^{V}(n_k^t + \beta) - N} \cdot \frac{n_m^k + \alpha}{\sum_{k=1}^{K}(n_m^k + \alpha)} \qquad (6)$$

Then $\mathbf{q}$ *is the pseudo sampling distribution in this sampling.*

Theorem 2. *Among all the partitions, there must exist a partition* γ' *in*

$$\mathcal{P} = \{\gamma | \exists k, s.t N_k = N, N_j = 0, \forall j \neq k\}$$

such that

$$\varepsilon_{\gamma'} = \max_{\gamma}\{\varepsilon | \gamma \in \Gamma\} = 2\ln(\frac{\sum_{j\neq k} r_j + q_k}{\sum_j r_j}) \qquad (7)$$

if condition (4) holds. k *is the topic index such that* $|r_k - q_k| = \|\mathbf{r} - \mathbf{q}\|_\infty$, $\mathbf{q}$ *is the pseudo sampling distribution.*

Proof. Appendix C of full version [Zhao *et al.*, 2019]. □

Theorem 2 indicates that only the partitions in $\mathcal{P}$ need to be considered for computing the privacy ε in the each sampling processing, which greatly reduce the searching scope. In particular, if condition (4) holds for all partitions in $\mathcal{P}$, the privacy guarantee could be computed directly by Equation (8), which is the first case to consider. If not, for any partition γ in $\mathcal{P}$ not satisfying condition (4), the privacy guarantee could be computed by Equation (5). Due to the arbitrariness of γ, we have another K cases to consider since there are K partitions in $\mathcal{P}$. Furthermore, since whether condition (4) holds is unknown, we have to enumerate all these $K+1$ cases to find the privacy guarantee bound. Algorithm 1 presents the searching-based algorithm for monitoring the privacy guarantee of each sampling for each word.

Privacy Monitoring for LDA

So far, the privacy guarantee ε_w^i of the sampling process for word w in the ith iteration can be measured by Algorithm 1.

Algorithm 2 Privacy Monitoring for CGS in LDA

Input: document corpus D
Parameter: iteration number n
Output:privacy guarantee ε

```
while not finished do
    for each document m in D do
        for each word w in m do
            compute the sampling distribution p
            call algorithm 1 to compute ε_w^i
            sample a topic and update matrices N_k^t and N_m^k
        end for
    end for
end while
return max_w{Σ_{i=1}^n ε_w^i}
```

Since the sampling process of the whole CGS algorithm is iteratively performed for each word but alternatively among all the words in the corpus, the total privacy guarantee of the whole CGS process in LDA training could be computed according to the composition theorems of DP.

Theorem 3. *Given a corpus* D*, suppose the CGS algorithm performed on word* w *at the* i*-th iteration satisfies* ε_w^i*-DP, then after* n *iterations, the whole CGS algorithm performed on* D *satisfies* $\max_w\{\sum_{i=1}^n \varepsilon_w^i\}$*-DP.*

Proof. For any word w, after n iterations of sampling, it will be accessed to by the whole CGS process n times, according to the sequential composition theorem [Li *et al.*, 2016], the total privacy guarantee for word w in the CGS algorithm is $\varepsilon_w = \sum_{i=1}^n \varepsilon_w^i$. While, according to the Equation (1), each iteration of CGS in LDA only accesses to each word once to perform the sampling, then according to the parallel composition theorem [Li *et al.*, 2016], the total privacy guarantee for the copus(all words) should be the maximum privacy guarantee of CGS among all words, that is $\max_w\{\sum_{i=1}^n \varepsilon_w^i\}$. □

Based on this observation, Algorithm 2 shows the privacy monitoring algorithm for the whole CGS process in LDA.

3.2 LP-LDA

As analyzed above, CGS algorithm can intrinsically guarantee the privacy of individual documents for the LDA model trained on a plain-text dataset, which is owned by a trustworthy curator. However, in many distributed applications, data servers are not always privacy-reliable and data owners may not be willing to directly contribute their sensitive data. In this case, we further propose a hidden-data based LDA mechanism LP-LDA that can perform the training process on a sanitized dataset. In particular, the LP-LDA mechanism mainly consists of two components: local perturbation at the user side and training on reconstructed dataset at the server side.

Local Perturbation

The local perturbation at the user side includes the following steps:

- *Step 1.* Each document m is encoded as a binary vector $\mathbf{V}_m$, in which each bit $\mathbf{V}_m[j]$ represents the presence of the j-th word in the word bag of the corpus.

- *Step 2*. Each bit $\mathbf{V}_m[j]$ of the binary vector $\mathbf{V}_m$ is then randomly flipped according to the following randomized response rule:

$$\hat{\mathbf{V}}_m[i] = \begin{cases} \mathbf{V}_m[j], & \text{with probability of } 1-f \\ 1, & \text{with probability of } f/2 \\ 0, & \text{with probability of } f/2 \end{cases}$$

  where $f \in [0,1]$ is a parameter that specifies the randomness of flipping and adjusts the local privacy level.

- *Step 3*. Then the noisy binary vector $\hat{\mathbf{V}}_m[j]$ is sent to the central server by each user. Obviously, $\hat{\mathbf{V}}_m[j]$ is locally sanitized without concerning user's privacy.

Training on Reconstructed Dataset

After receiving the flipped binary vectors from a large number of data contributors, the central server can aggregate the vectors, reconstruct the dataset and then perform training on the reconstructed dataset. The rationale behind this is that the training result of topic-word distribution is insensitive to the document partitions and only depends on the total word counts in the corpus.

- *Step 1*. For each bit in the noisy binary vectors, the server counts the number of $1'$s as $n_t = \sum_{i=1}^{M} \hat{\mathbf{V}}_m[t]$.
- *Step 2*. The server then estimates the true count N_t of each bit in the original binary vectors $\mathbf{V}_m$ as $\hat{N}_t = (2n_t - fM)/2(1-f)$.
- *Step 3*. For each bit, the server first computes the difference $\delta_t = \hat{N}_t - n_t$.
- *Step 4*. For each bit t, if $\delta_t > 0$, the server randomly samples δ_t binary vectors with the t-th bit as 0 and sets the t-th bit as 1; if $\delta_t < 0$, then the server randomly samples $|\delta_t|$ binary vectors with the t-th bit as 1 and sets the t-th bit as 0; otherwise, keeps the noisy bit vectors as received.
- *Step 5*. Based on the noisy bit vectors, the server reconstructs a dataset and performs the CGS process on it.

Privacy Analysis of LP-LDA

Theorem 4. *The LP-LDA satisfies ε-differential privacy for each document contributor where $\varepsilon = \ln \frac{1-f/2}{f/2}$.*

Proof. Suppose a word t appears in a noisy bit vector, then the probability of it being kept from the original bit vector is $Pr(\hat{\mathbf{V}}_m[t] = 1|\mathbf{V}_m[t] = 1) = 1 - f/2$ and the probability of it being flipped from the original bit vector is $Pr(\hat{\mathbf{V}}_m[t] = 1|\mathbf{V}_m[t] = 0) = f/2$. Then, according to the definition of DP, it guarantees the privacy of

$$\varepsilon = \left| \ln \frac{Pr(\hat{\mathbf{V}}_m[t] = 1|\mathbf{V}_m[t] = 1)}{Pr(\hat{\mathbf{V}}_m[t] = 1|\mathbf{V}_m[t] = 0)} \right| = \ln \frac{1-f/2}{f/2}.$$

The analysis also holds for any bit t that $\hat{\mathbf{V}}_m[t] = 0$. □

Since the reconstruction and training process are essentially post-processes on the noisy bit vectors, the local privacy remains unchanged for all the documents.

Dataset	#. words	#. training docs	#. test docs
KOS	209169	3000	430
NIPS	410753	1349	150
Enron	356363	8000	2000

Table 1: Details about the real-world datasets

Utility Analysis of LP-LDA

Theorem 5. *Let N_t and n_t denote the counts of word t in the original and perturbed datasets, respectively, then*

$$\hat{N}_t = \frac{2n_t - fM}{2(1-f)} \tag{8}$$

is an unbiased estimator of N_t with the variance of

$$D(\hat{N}_t) = \frac{(2-f)fM}{4(1-f)^2}. \tag{9}$$

Proof. See proof in full version [Zhao *et al.*, 2019]. □

4 Experiment

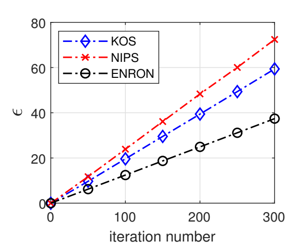
In this section, we evaluate the effectiveness of our proposed privacy monitoring algorithm and locally private LDA algorithm LP-LDA on real-world datasets.

The datasets used in our experiment are: KOS[1]: contains 3430 blog entries from dailykos website. NIPS[2]: contains 1740 research papers from NIPS conference. Enron[3]: contains 0.5 million email messages from about 150 users.

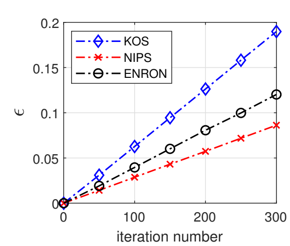
We extracted part of these datasets as our training datasets and the rest as the testsets. For simplicity, we setup a preprocessing phase on these dataset before running our experiments. For example, all stop words were removed and 1000 most frequent words in each dataset were chosen as the corresponding vocabulary list. Details about these datasets after pre-processing can be found in Table 1.

In our experiments, for all datasets, the topic number is set as 50, the maximum iteration number of CGS process in LDA model training is set as 300, which is sufficient for convergence on all three datasets. The hyper parameters α and β are set as 0.1, 0.01, respectively.

4.1 Inherent Privacy of CGS in LDA



(a) Document-level privacy



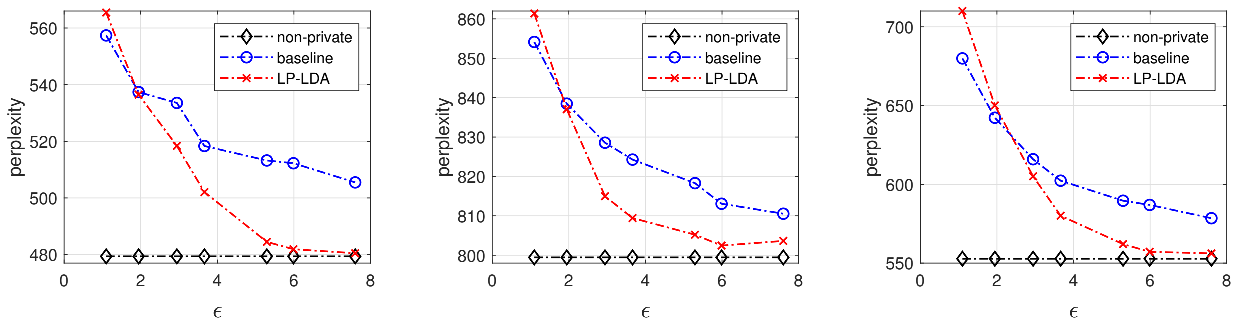
(b) Word-level privacy

Figure 1: privacy guarantee vs. iteration number of CGS in LDA

[1]http://archive.ics.uci.edu/ml/

[2]http://nips.djvuzone.org/txt.html

[3]www.cs.cmu.edu/ enron



(a) KOS (b) NIPS (c) Enron

Figure 2: Perplexity of LDA model vs. privacy

Figure 1 illustrates the inherent privacy guarantee of CGS algorithm in LDA measured by our proposed privacy monitoring algorithm on three datasets for both document-level and word-level privacy. It should be noted that a larger privacy parameter ε in the figures means less privacy guarantee.

As we can see in both subfigures, both word-level and document-level privacy parameter ε of CGS in LDA increase approximately linearly with the number of sampling iterations. This is because the privacy bound in each iteration of sampling is very close, and the total privacy parameter will accumulate with the number of iterations according to the sequential composition theorem.

Although CGS on all datasets can obtain privacy guarantee for free, the inherent privacy varies on different datasets. For document level, the privacy guarantee achieved on NIPS is the weakest while that on Enron is the strongest. That is because the documents in NIPS averagely contain the most words, which also means it is the most difficult to be effectively hidden. For word-level privacy, the LDA model trained on NIPS has the strongest privacy guarantee because it contains largest number of words and the sampling probability for each unique word will be the lowest. On the contrary, with the fixed length of vocabulary list, KOS contains the fewest words in total and results the weakest word-level privacy after same number of iteration.

4.2 Local Mechanism

Figure 2 depicts the simulation performance of our proposed LP-LDA mechanism in terms of different level of privacy. The flipping probability f in LP-LDA varies from 0.5 to 0.001, and the corresponding privacy level varies from 1.089 to 7.6004. The utility of LDA model training is measured by the perplexity on test sets. Perplexity is an information-theoretical measure commonly used to evaluate the prediction performance of LDA model and generally smaller perplexity on a test set means better prediction accuracy. In particular, we compared LP-LDA with a baseline privacy-preserving LDA mechanism based on Laplace mechanism, in which the sufficient statistics of the likelihood, i.e., word count matrices N_k^t and N_m^k are privatized at the beginning of the CGS algorithm with the sensitivity of 1 and privacy of ε [Foulds *et al.*, 2016].

As shown, both the perplexity of LP-LDA and baseline algorithm decrease with the increase of ε, which shows the trade-off between the privacy and utility. For stronger privacy regime with smaller ε, the perplexity of LP-LDA is larger than that of the baseline algorithm. That is because the Laplace mechanism baseline algorithm incurs less noise than randomized response in LP-LDA for the statistics of word count N_t. While for weaker privacy regime with larger ε, the perplexity of LP-LDA is far less than that of the baseline algorithm and shows greater LDA model training utility. These utility comparison results can be also explained by the variance difference of the word count N_t in two mechanisms. In baseline mechanism based on Laplace noise, the noise variance is $D(N_t') = 2K^2/\varepsilon^2$, while the variance $D(\hat{N}_t)$ in our proposed LP-LDA is shown in Equation (10). In particular, for larger ε on all three datasets, the baseline method always has a higher noise variance than LP-LDA.

5 Conclusion and Future Work

In this work, we investigate the privacy protection of LDA model training. We first present that the CGS algorithm in LDA can possess some inherent privacy in each sampling process and then propose a efficient searching-based privacy monitoring algorithm to identify the privacy guarantee bound in the iterative CGS process of LDA. In addition, besides training on a trustworthy data server, we also propose a locally private solution of LP-LDA to achieve LDA training on a sanitized dataset by individual local users, which is applicable to many scenarios. The experiments on real-world datasets validate our proposed approaches. Future work will center on finding tighter bound of the inherent privacy guarantee in LDA model training.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant 61572398, Grant 61772410, Grant 61802298, Grant 11690011 and Grant U1811461, in part by the Fundamental Research Funds for the Central Universities under Grant xjj2018237, in part by the China Postdoctoral Science Foundation under Grant 2017M623177, and in part by the National Key Research and Development Program of China under Grant 2017YFB1010004.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Chaudhuri *et al.*, 2011] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

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

[Foulds *et al.*, 2016] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

[Fredrikson *et al.*, 2014] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32, 2014.

[Heinrich, 2005] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Technical report, 2005.

[Li *et al.*, 2016] Ninghui Li, Min Lyu, Dong Su, and Weining Yang. Differential privacy: From theory to practice. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(4):1–138, 2016.

[Li *et al.*, 2019] Yanan Li, Xuebin Ren, Shusen Yang, and Xinyu Yang. Impact of prior knowledge and data correlation on privacy leakage: A unified analysis. *IEEE Transactions on Information Forensics and Security*, 2019.

[Park *et al.*, 2016] Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Variational bayes in private settings (vips). *arXiv preprint arXiv:1611.00340*, 2016.

[Ren *et al.*, 2018] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A McCann, and S Yu Philip. Lopub: High-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 13(9):2151–2166, 2018.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[Smola and Narayanamurthy, 2010] Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

[Wang *et al.*, 2014] Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Jia Zeng, Qiang Yang, et al. Towards topic modeling for big data. *arXiv preprint arXiv:1405.4402*, 2014.

[Wang *et al.*, 2015] Yu-Xiang Wang, Stephen E Fienberg, and Alexander J Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *ICML*, volume 15, pages 2493–2502, 2015.

[Yuan *et al.*, 2015] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee, 2015.

[Yut *et al.*, 2017] Lele Yut, Ce Zhang, Yingxia Shao, and Bin Cui. Lda*: a robust and large-scale topic modeling system. *Proceedings of the VLDB Endowment*, 10(11):1406–1417, 2017.

[Zhao *et al.*, 2019] Fangyuan Zhao, Xuebin Ren, Shusen Yang, and Xinyu Yang. On privacy protection of latent dirichlet allocation model training. *full version of this paper, available online at https://arxiv.org/pdf/1906.01178v2.pdf*, 2019.

[Zhu *et al.*, 2016] Tianqing Zhu, Gang Li, Wanlei Zhou, Ping Xiong, and Cao Yuan. Privacy-preserving topic model for tagging recommender systems. *Knowledge and information systems*, 46(1):33–58, 2016.