Semantic Consistent Topic Discovery with Differential Privacy

Yexuan Shi[1], **Zhiyang Su**[2], **Di Jiang**[3], **Zimu Zhou**[4], **Wenbin Zhang**[1], **Yongxin Tong**[1]
[1] Beihang University, Beijing, China
[2] Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] WeBank, Shenzhen, China
[4] ETH Zurich, Zurich, Switzerland
[1] {skyxuan, zhangwenbin, yxtong}@buaa.edu.cn, [2] zsuab@ust.hk,
[3] dijiang@webank.com, [4] zzhou@tik.ee.ethz.ch

Abstract

General-purpose topic models such as Latent Dirichlet Allocation (LDA) are widely used in industrial applications. However, conventional topic models usually lack data privacy protection mechanisms. This has become a serious issue since industrial data are often sensitive and new policies such as the General Data Protection Regulation (GDPR) enforce protection of sensitive data. To solve this problem, we propose a novel privacy-preserving general-purpose topic model named Private and Consistent Topic Discovery (PC-TD). On the one hand, PC-TD seamlessly integrates differential privacy, a popular privacy preserving technique, to provide privacy guarantees. On the other hand, PC-TD exploits multiple sources of semantic consistency information to retain the accuracy of topic modeling while protecting data privacy. We verify the effectiveness of PC-TD on real-life datasets. Experimental results show its superiority over the state-of-the-art general-purpose topic models.

1 Introduction

Topic modeling is a powerful technique for unsupervised analysis of large document collections. It has been widely applied in tag recommendation, text categorization, opinion mining and statistical language modeling [Krestel *et al.*, 2009; Zhou *et al.*, 2008; Steyvers *et al.*, 2004]. In fact, general-purpose topic models such as Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] have become the de facto in many industrial applications [Das *et al.*, 2007; Liu *et al.*, 2012].

Despite its popularity, topic modeling typically ignores data privacy. Missing data privacy in topic models is a severe problem as industrial data is always sensitive. The enforcement of the General Data Protection Regulation (GDPR) [1] makes it even worse as they can be considered as illegal models. To comply with the data privacy policies in industrial applications, the whole procedure of topic modeling needs to provide certain degree of privacy guarantee.

In this paper, we leverage differential privacy [Dwork *et al.*, 2006; Dwork, 2011; Dwork and Roth, 2014] to enable privacy-preserving topic modeling. Differential privacy is a formal definition of the privacy properties of data analysis algorithms. Yet it is also a lossy mechanism, which may degrade the accuracy of topic models. Hence the challenge is *how to provide a reasonable degree of privacy while retaining the accuracy of topic modeling*.

We propose *Private and Consistent Topic Discovery* (PC-TD), a new privacy-preserving general-purpose topic model. To seamlessly integrate differential privacy into topic modeling, we apply an improved composition method called moments accountant [Abadi *et al.*, 2016] to achieve a reasonable degree of privacy. To retain the accuracy of topic modeling, we rely on two observations. First, general-purpose topic models discover topics solely based on word co-occurrence in document without considering other semantic relations of linguistic phenomena. Thus, we model the linguistic phenomenon as *semantic unit* whose content is generated by a single topic to incorporate the *local sematic consistency* into topic modeling. Second, external knowledge base can improve the topical coherency and interpretability. Thus, a flexible mechanism is proposed to introduce any word relation of external knowledge base into the procedure of topic modeling to ensure the *global semantic consistency*.

The main contributions of this paper are as follows.

- We propose a novel general-purpose topic model named *Private and Consistent Topic Discovery* (PC-TD) which effectively protects data privacy. We prove that PC-TD provides data privacy guarantees. We also design techniques to retain the accuracy of topic modeling by considering global and local semantic consistency.
- We validate the effectiveness of PC-TD on real-life datasets. Extensive experiments demonstrate the superiority of PC-TD.

The rest of this paper is organized as follows. In Section 2, we review the related work. Then we propose the technical details of PC-TD in Section 3. We present the experimental evaluations in Section 4 and finally conclude the paper in Section 5.

[1] https://gdpr-info.eu/

2 Related Work

In this section, we briefly summarize the related work from the following two fields: topic modeling and differential privacy.

2.1 Topic Modeling

Topic modeling is a method designed to discover the abstract "topics" that occur in a collection of documents [Steyvers and Griffiths, 2007; Anthes, 2010]. Within the topic modeling framework, documents can be represented by the topics and the entire corpus can be indexed and organized in terms of this discovered semantic structure.

Research on topic modeling dates back to the Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990] which is an information retrieval model for excavating the latent association between the text and the words. To address the statistical unsoundness of LSA, a generative latent-variable model called Probabilistic Latent Semantic Analysis (PLSA) is proposed [Hofmann, 1999], where the latent variables are topics in documents. As an improvement of PLSA, Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] is a more general Bayesian probabilistic topic model, which models each document as a multi-membership mixture of K corpus-wide topics, and each topic as a multi-membership mixture of the terms in the corpus vocabulary. By applying additional constraints on the basic LDA, more variants and offshoots of LDA have been proposed. [Blei and McAuliffe, 2007; Ramage *et al.*, 2009; Wang *et al.*, 2009; Blei *et al.*, 2010; Yan *et al.*, 2013]

PLSA and LDA have been proved their applicability in industry and have been successfully applied in collaborative filtering for generating personalized recommendations in Google News [Das *et al.*, 2007] and real-time Q&A systems in Baidu [Liu *et al.*, 2012]. However, with the popularity of these two general-purpose topic models in industry, little work has been done to further enhance them by fixing the challenge that is discussed in the introduction.

2.2 Differential Privacy

Differential privacy [Dwork *et al.*, 2006; Dwork, 2011; Dwork and Roth, 2014] is a formal definition of the privacy properties of data analysis algorithms. It is defined in terms of the application-specific concept of adjacent databases. In this paper, the training dataset is a set of documents. Thus, we say that two of these datasets are adjacent if they differ in a single entry, that is, if one word is present in one document in the first dataset and absent in the other.

Definition 1 ((ϵ, δ)-differential privacy)**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for any two adjacent inputs $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subsetneq \mathcal{R}$ it holds that*

$$Pr[\mathcal{M}(d) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(d') \in S] + \delta$$

We use the variant of differential privacy introduced by Dwork [Dwork and Roth, 2014], which allows for the possibility that plain ϵ-differential privacy is broken with probability δ. Intuitively, the definition states that the output probabilities must not change very much when a single individual's data is modified, thereby limiting the amount of information that the algorithm reveals about any one individual.

A common paradigm for approximating a deterministic real-valued function $f : \mathcal{D} \to \mathcal{R}$ with a differentially private mechanism is via additive noise calibrated to f's *sensitivity* S_f, which is defined as the maximum of the Euclidean norm $||f(d) - f(d')||_2$ where d and d' are adjacent inputs. For instance, the Gaussian noise mechanism is defined by

$$\mathcal{M}(d) \triangleq f(d) + \mathcal{N}(0, S_f^2 \cdot \sigma^2) \tag{1}$$

where $\mathcal{N}(0, S_f^2 \cdot \sigma^2)$ is the normal (Gaussian) distribution with mean 0 and standard deviation $S_f \sigma$.

3 Private and Consistent Topic Discovery

We propose a generative model to discover the topics of documents. As shown in Figure. 1, in order to achieve the local semantic consistency, the PC-TD orgnaises the words of documents into semantic units, and use EM algorithm to infer the latent parameters based on the semantic units. When the EM algorithm accesses the statistics of semantic units, a Gaussian noise is added to protect the data privacy of documents. On the other hand, we introduce external knowledge base to help us improving the effectiveness of topic modeling. We get the word similarity matrix via the external knowledge base and integrate it into the M-step to ensure the global semantic consistency.

In this section, we first introduce the assumptions and definition of semantic units of our model in Section 3.1. Then we propose the EM inference method of our model with differential privacy in Section 3.2. Next, we consider the local semantic consistency by introducing the similarity of words in Section 3.3. Finally, the analysis of privacy is given in Section 3.4.

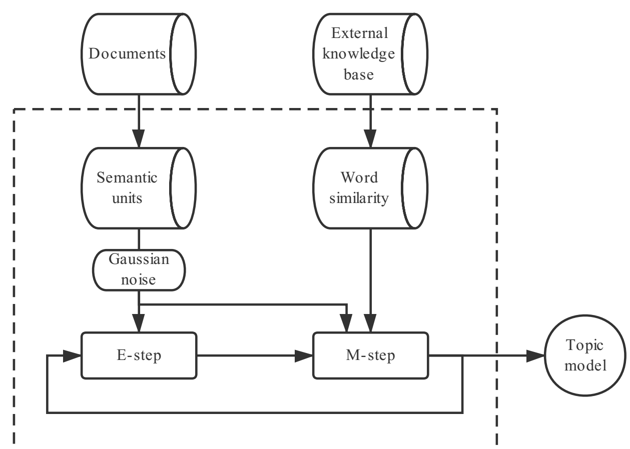

Figure 1: Framework of Private and Consistent Topic Discovery

3.1 Model Assumptions

We utilize d to denote a "document", w a "word" and z a latent topic. Based on these notations, we introduce the following probabilities: $p(d_i)$ is the probability of a particular document d_i , $p(w_j|z_k)$ is the conditional probability of a specific word w_j conditioned on the latent topic variable z_k and $p(z_k|d_i)$ is a document-specific probability distribution over the latent topic z_k. A subtle issue of the assumption of PC-TD is that we need to consider the local linguistic phenomena for the *local semantic consistency*. Therefore, we introduce a concept of *semantic unit*, whose contents are generated by a single topic. Based upon the application scenarios, the semantic unit can be flexibly interpreted as n-gram, sentence, paragraph, etc. We present the generative process of PC-TD as follows:

1. Select a document d_i with probability $p(d_i)$;
2. For each semantic unit s_{ij} in d_i, pick a latent topic z_k with probability $p(z_k|d_i)$;
3. For each position in s_{ij}, generate a word w with probability $p(w|z_k)$.

Translating the generative process into complete data logarithm likelihood results in the following expression:

$$\begin{aligned} L(\mathbf{d},\mathbf{s},\mathbf{z}) &= \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z}\log p(d_i, s_{ij}, z_k) \\ &= \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z}\log\Big(p(d_i)p(z_k|d_i)p(s_{ij}|z_k)\Big) \end{aligned} \quad (2)$$

where D is the number of documents, S_i is the number of semantic units in the ith document and Z is the number of topics. Essentially, to obtain Eq. (2) one has to sum over the possible choices of z_k. Hence, the goal of our model is to identify conditional probability mass functions such that the document-specific word distributions are as faithfully as possible approximated by convex combinations of these topics.

3.2 Private Parameter Inference

We now propose a private EM algorithm to infer the latent parameters of PC-TD. In the E-step, the posterior estimation of the latent topic z_k of semantic unit s_{ij} in document d_i is straightforwardly obtained as follows:

$$p(z_k|d_i, s_{ij}) = \frac{p(z_k|d_i)p(s_{ij}|z_k)}{\sum_{k'=1}^{Z} p(z_{k'}|d_i)p(s_{ij}|z_{k'})} \quad (3)$$

where $p(s_{ij}|z_k) = \prod_{w=1}^{W} p(w|z_k)^{N_{ijw}}$ and N_{ijw} is the number of w in s_{ij}.

In this step, we will access the training data by counting the number of N_{ijw}. Thus, perturbing N_{ijw} leads to perturbing the parameters of interest. To achieve this goal, we add a Gaussian noise to N_{ijw}:

$$\hat{N}_{ijw} = N_{ijw} + \Omega \quad (4)$$

where $\Omega \sim \mathcal{N}(0, (\Delta N)^2\sigma^2)$ and ΔN is the sensitivity.

Since we say two of these datasets are adjacent if one word is present in one document in the first dataset and absent in the other, it is obviously that the sensitivity $\Delta N = 1$.

After we add the noise to statistics N_{ijw}, we can calculate the perturbed posterior estimation:

$$\begin{aligned} \hat{r}_{ijk} &= \hat{p}(z_k|d_i, s_{ij}) \\ &= \frac{p(z_k|d_i)\prod_{w=1}^{W} p(w|z_k)^{\hat{N}_{ijw}}}{\sum_{k'=1}^{Z} p(z_{k'}|d_i)\prod_{w'=1}^{W} p(w'|z'_k)^{\hat{N}_{ijw}}} \end{aligned} \quad (5)$$

Next, we introduce the formulas of inference in the M-step. In this step, we have to maximize the expected logarithm likelihood, which is defined as follows:

$$\begin{aligned} Q &= \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z}\hat{r}_{ijk}\log p(d_i, s_{ij}, z_k) \\ &= \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z}\hat{r}_{ijk}\Big(\log p(z_k|d_i) + \\ &\qquad \log p(s_{ij}|z_k) + \log p(d_i)\Big). \end{aligned} \quad (6)$$

In order to take care of the normalization constraints, Eq. (6) has to be augmented by appropriate Lagrange multipliers. Maximization of the augmented Q with respect to the probability mass functions leads to the following set of stationary equations:

$$p(z_k|d_i) = \frac{\sum_{j=1}^{S_i}\hat{r}_{ijk}}{\sum_{j=1}^{S_i}\sum_{k'=1}^{Z}\hat{r}_{ijk'}}, \quad (7)$$

$$p(w|z_k) = \frac{\sum_{i=1}^{D}\sum_{j=1}^{S_i} N_{ijw}\hat{r}_{ijk}}{\sum_{i=1}^{D}\sum_{j=1}^{S_i} N_{ij}\hat{r}_{ijk}}, \quad (8)$$

where N_{ijw} is the number of w in the semantic unit s_{ij} and N_{ij} is the number of words in the semantic unit s_{ij}.

The same as E-step, we only access training data by counting the number of words in the semantic units of some documents. Thus, we can use the same perturbing method as Eq.(4) in this step and get the perturbed probability $\hat{p}(w|z_k)$:

$$\hat{p}(w|z_k) = \frac{\sum_{i=1}^{D}\sum_{j=1}^{S_i}\hat{N}_{ijw}\hat{r}_{ijk}}{\sum_{i=1}^{D}\sum_{j=1}^{S_i}\hat{r}_{ijk}\sum_{w=1}^{W}\hat{N}_{ijw}}, \quad (9)$$

The resulting algorithm is summarized in Algorithm 1. Note that our algorithms will run for a prespecified number of iterations, and with a prespecified σ; this ensures a certain level of (ϵ, δ) guarantee in the released expected sufficient statistics from Algorithm 1.

3.3 Global Semantic Consistency

The previous subsections illustrate how to model local semantic consistency in PC-TD and infer the parameters via a private EM algorithm. In this subsection, we discuss how to ensure global semantic consistency in PC-TD and present an approach to adapt the EM algorithm presented in the previous section. We refer to *global semantic consistency* as word relations which can be obtained from external sources such as human-engineering ontology [Miller, 1995] and automatically built knowledge base [Dong *et al.*, 2014]. In this paper,

Algorithm 1: Private and Consistent Topic Discovery

Input: Documents $\mathcal{D}$, number of documents D, number of topics K, number of words W, standard deviation σ, total round of EM algorithm T

Output: the probability distribution $\{p(z_k|d_i)\}$, $\{p(w|z_k)\}$

Initialize $p(z_k|d_i), p(w|z_k)$ randomly;
for $t = 1, 2, \cdots, T$ **do**
  $N_{ijw} \leftarrow$ the number of words w in each semantic units s_{ij} from $\mathcal{D}$;
  $\hat{N}_{ijw} \leftarrow$ add gaussian noise $\mathcal{N}(0, \sigma^2)$ to N_{ijw} ;
  /* E-step: */
  Compute $\hat{r}_{ijk}$ as Eq. (5);
  /* M-step: */
  Compute $p(z_k|d_i)$ as Eq. (7);
  Compute $\hat{p}(w|z_k)$ as Eq. (9);
return $\{p(z_k|d_i)\}, \{\hat{p}(w|z_k)\}$;

we use the word embedding as an example to demonstrate how to obtain global semantic information.

Word embedding is a technique of language modeling and feature learning in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. We can use a popular method, Word2vec [Mikolov *et al.*, 2013], to get such a mapping. After we get the vectors of words, the similarity of two words can be calculated as follows.

We denote the similarity of two word vectors v_a and v_b as R_{ab}. It can be calculated by cosine similarity [Singhal, 2001]:

$$R_{ab} = \frac{v_a \cdot v_b}{||v_a||_2||v_b||_2}. \tag{10}$$

We proceed to discuss the strategy of utilizing R in PC-TD. We want the probability $p(w|z_k)$ to be consistent with word relations stored in R. Here we use a quadratic-form influence term with a trade-off factor τ. Formally, for a given R, we adjust the topic-word distribution $P(w|z_k)$ as follows:

$$p'(w|z_k) \leftarrow p(w|z_k) + \tau \frac{p(w|z_k)\sum_{i=1}^{W} R_{iw} p(i|z_k)}{P(\cdot|z_k)^T R P(\cdot|z_k)}. \tag{11}$$

After we get $p'(w|z_k)$, it should be normalized to ensure that $\sum_w p'(w|z_k) = 1$. It is easy to see that the adjusted $p'(w|z_k)$ is influenced by the other words related to w in $\mathbf{R}$. In practice, Eq. (11) is applied after each private EM iteration until convergence is achieved. Since we are only interested in relatively frequent words from the vocabulary, $\mathbf{R}$ will be a sparse matrix and hence computations of R are efficient in practice.

3.4 Privacy Analysis

In this subsection, we present the privacy analysis of PC-TD. Since PC-TD uses EM algorithm to infer the latent parameters, we use the *Moments Accountant* (MA) composition method [Abadi *et al.*, 2016] to account the privacy loss incurred by successive iterations of our EM algorithm.

The moments accountant method provides tighter guarantees than linear strong composition. In moments accountant method, the *log-moments function* of the *privacy loss* random variable is introduced to track the privacy loss incurred by applying mechanisms $\mathcal{M}_1, \cdots, \mathcal{M}_T$ successively to a dataset $\mathcal{D}$.

Specifically, for two neighboring databases $\mathcal{D}, \mathcal{D}'$, it defines the *privacy loss* of a mechanism $\mathcal{M}$ on an outcome $o \in \mathcal{R}$ as

$$L_{\mathcal{M}}(\mathcal{D}, \mathcal{D}', w) = \log \frac{\Pr[\mathcal{M}(\mathcal{D}, w) = o]}{\Pr[\mathcal{M}(\mathcal{D}', w) = o]} \tag{12}$$

In our EM algorithm, each iteration can be regarded as a mechanism M_t and the *log-moments function* $\alpha_{\mathcal{M}_t}$ of a mechanism M_t is defined as:

$$\alpha_{\mathcal{M}_t} = \sup_{\mathcal{D}, \mathcal{D}', w} \log \mathbb{E}[\exp(\lambda L_{\mathcal{M}_t}(\mathcal{D}, \mathcal{D}', w))] \tag{13}$$

[Abadi *et al.*, 2016] shows that if $\mathcal{M}$ is the combination of mechanisms $(\mathcal{M}_1, \cdots, \mathcal{M}_T)$ where each mechanism adds independent noise, then, its log moment generating function $\alpha_{\mathcal{M}}$ has the property of:

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=1}^{T} \alpha_{\mathcal{M}_t}(\lambda) \tag{14}$$

Additionally, given a log moment function $\alpha_{\mathcal{M}}$, [Abadi *et al.*, 2016] shows that the corresponding mechanism $\mathcal{M}$ satisfies a range of privacy parameters (ϵ, δ) with the following equation:

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon) \tag{15}$$

These properties immediately suggest a procedure for tracking privacy loss incurred by a combination of mechanisms $(\mathcal{M}_1, \cdots, \mathcal{M}_T)$ on a dataset.

By using these two properties, we can get our main theorem.

Theorem 1. *For any $\epsilon < \Theta(T)$, PC-TD is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma \geq \Theta\Big(\frac{\sqrt{T\log(1/\delta)}}{\epsilon}\Big)$$

Proof. From the lemma 3 in [Abadi *et al.*, 2016], the log-moments function of the Gaussian Mechanism $\mathcal{M}$ applied to a query with sensitivity $\Delta \leq 1$ is $\alpha_{\mathcal{M}}(\lambda) \leq \frac{\lambda(\lambda+1)}{2\sigma^2}$. Thus, it can be bounded as follows $\alpha(\lambda) \leq T\lambda^2/\sigma^2$. According the two properties, to guarantee Algorithm 1 to be $(\epsilon.\delta)$-differentially private, it suffices that

$$T\lambda^2/\sigma^2 \leq \lambda\epsilon/2$$

$$\exp(-\lambda\epsilon/2) \leq \delta$$

In addition, we need $\lambda \leq \sigma^2 \log(1/\sigma)$.

It is easy to verify that when $\epsilon = \Theta(T)$, we can satisfy all these conditions by setting $\sigma = \Theta(\frac{\sqrt{T\log(1/\delta)}}{\epsilon})$. □

4 Experiments

In this section, we evaluate the performance of PC-TD. In Section 4.1, we describe the experimental setup. In Section 4.2, we demonstrate the impact of privacy. Finally, we demonstrate the effectiveness of PC-TD with quantitative evaluation in Section 4.3.

4.1 Experimental Setup

Dataset. We evaluate our method on a corpus collected from New York Times[2]. We sample 500 documents from the news of June 26th-30th, 2016 as our training dataset. After removing the stopwords, we get 18,286 unique words.

Mertric. We use the perplexity of documents to evaluate the performance of topic models. Perplexity is an information-theoretic measure of the predictive performance of probabilistic models which is commonly used in the context of language modeling. [Jelinek *et al.*, 1977] The perplexity of a topic model on a set of documents is defined as

$$\text{perplexity} = \exp\Big(-\frac{1}{\sum_{i=1}^{D}|d_i|} \sum_{i=1}^{D}\sum_{w\in d_i}\ln(\sum_{k=1}^{Z}p(w|z_k)p(z_k|d_i))\Big) \tag{16}$$

Implementation. For PC-TD, we first split each document into several sentences. Then we consider each three words in these sentences as a semantic unit. To achieve the global semantic consistency, we use a pre-trained Word2vec by Google [Mikolov *et al.*, 2013].

We tune the parameter τ which serves as the weight parameter for global semantic consistency. A small τ tends to diminish the influence of the global semantic consistency while large τ makes the global semantic consistency dominates the other word co-occurrence information. When τ increases from 0.1 to 0.5, the log likelihood of holdout data first increases and then falls. We observe that the best performance is achieved when τ is set to 0.3, showing that 0.3 strikes a good balance for the word co-occurrence and global semantic consistency. Hence τ is set to 0.3 by default in our experiments. The relatively small value of τ indicates that PC-TD primarily relies on the word co-occurrence information in the training data and the word relation information from other sources can achieve a slight improvement.

We compare our algorithm with the typical general-purpose topic model LDA to verify its effectiveness. We use Markov Chain Monte Carlo (MCMC) sampling method to train an LDA model [Steyvers and Griffiths, 2007], with parameter $\alpha = Z/50, \beta = 0.01$.

4.2 Privacy Evaluation

We first demonstrate the impact of privacy. Figure 2 shows the trade-off between ϵ and per-word perplexity on our dataset for the different methods under a variety of conditions, in which we vary the value of $\sigma \in \{0.15, 0.2, 0.25, 0.3\}$.

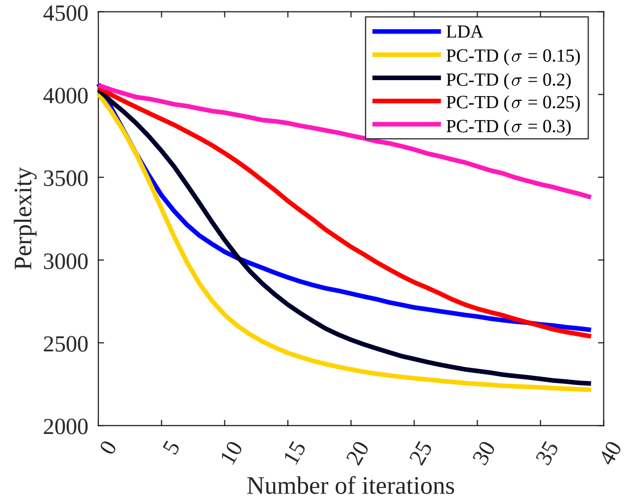

Figure 2: Convergence curves of varying σ

As expected, we observe that the perplexity gradually decreases as the number of iterations increases in all cases. From Figure. 2, we observe that as the deviation of noise increases, PC-TD needs more iterations to converge. When $\sigma = 0.25$, PC-TD converges within 35 iterations on our data set. However, when $\sigma = 0.3$, the convergence is slower. It indicates that the smaller the privacy budget is, the slower the algorithm converges. When $\sigma = 0.15$, PC-TD can even achieve a lower perplexity. When $\sigma = 0.25$, the convergence of PC-TD and LDA are slightly different. Notice that smaller deviation means larger privacy budget and greater risk of privacy disclosure. Thus, we choose $\sigma = 0.25$ for our method as a balance of privacy protection and effectiveness.

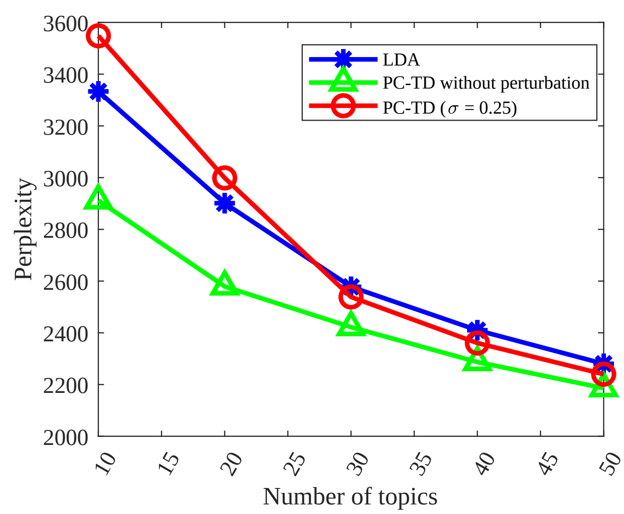

Figure 3: Perplexity of varying number of topics

[2]https://www.kaggle.com/nzalake52/new-york-times-articles

LDA		PC-TD without perturbation		PC-TD ($\sigma = 0.25$)	
topic 1:		topic 1:		topic 1:	
pitch	0.0705596	pitch	0.0199456	pitch	0.0203315
gathered	0.0097323	info	0.0111773	device	0.0041617
curtis	0.0055150	ramp	0.0047826	delivered	0.0036615
corey	0.0047039	smashed	0.0036892	allegation	0.0024893
bunt	0.0038929	gathered	0.0036244	catapult	0.0023844
chilled	0.0038929	indian	0.0034841	landing	0.0023364
grounded	0.0032441	buying	0.0033376	fractured	0.0022788
embraced	0.0025952	strain	0.0031046	micah	0.0020661
son	0.0025952	device	0.0027433	driven	0.0016771
apprehension	0.0024330	objection	0.0027054	responsive	0.0015447
topic 2:		topic 2:		topic 2:	
chad	0.0415856	chad	0.0159180	chad	0.0156159
separation	0.0145096	cookie	0.0067240	cookie	0.0080450
strapped	0.0134732	separation	0.0052944	separation	0.0077629
powerbook	0.0115299	church	0.0052913	lopez	0.0043082
nosed	0.0101049	embraced	0.0043136	church	0.0037838
pinch	0.0097162	lopez	0.0039555	ron	0.0036223
terry	0.0088094	radicalism	0.0028635	radicalism	0.0030566
thompson	0.0063479	segregated	0.0027051	nosed	0.0029864
ron	0.0046638	terry	0.0026754	embraced	0.0026293
gunfight	0.0044047	urbanite	0.0025712	singled	0.0022348
topic 3:		topic 3:		topic 3:	
smashed	0.0215277	smashed	0.0083382	smashed	0.0148687
freestyle	0.0166666	hat	0.0040704	info	0.0051527
indian	0.0127777	difference	0.0033508	raising	0.0038858
grounded	0.0075000	info	0.0032030	strain	0.0030042
hasidic	0.0068055	announce	0.0027637	ramp	0.0027003
designation	0.0054166	ankara	0.0026874	terry	0.0022467
fixed	0.0034722	statute	0.0026862	knotted	0.0019937
adviser	0.0033333	barometer	0.0025822	objection	0.0018980
cry	0.0033333	damned	0.0025310	tempered	0.0012534
conflict	0.0033333	emblematic	0.0025228	skylight	0.0012304

Table 1: Posterior topics from LDA, PC-TD without perturbation and PC-TD ($\sigma = 0.25$)

4.3 Effectiveness Evaluation

In this subsection, we evaluate the effectiveness of PC-TD by perplexity. Lower perplexity indicates better fit for the data. The experimental result of perplexity of training data is shown in Figure. 3. As the amount of topics ranges from 10 to 50, we observe that the perplexity of all the three topic models gradually decreases. This phenomenon indicates that a fairly large number of topics will provide better fit of the data. Among the three compared methods, PC-TD without perturbation performs the best. As the amount of topics increases, the perturbed PC-TD performs gradually better than LDA. For example, when the number of topics is 30, the LDA achieves the perplexity of 2577.8633, while the perplexity of PC-TD ($\sigma = 0.25$) is 2538.6533. This observation supports the idea that integrating global and local semantic consistency provides better fit for the underlying structure of the data. It further illustrates that the PC-TD can achieve similar or even better performance than LDA with privacy protection.

We also choose three topics from each topic models and show the top 10 words in terms of assigned probabilities, as shown in Table. 1. We can find that PC-TD without perturbation results in the most coherent words among all the methods and the PC-TD ($\sigma = 0.25$) also conforms to semantics to a certain extent.

5 Conclusion

In this paper, we propose PC-TD to discover latent topics with semantic consistency and privacy guarantee. PC-TD utilizes the prior knowledge about word relations to implement the global semantic consistency. Meanwhile, in light of the existence of semantic units such as sentences, PC-TD seamlessly integrates such local structures during its generation process. We propose a differential private parameter inference algorithm for PC-TD to ensure the privacy of sensitive documents. Experimental results on a corpus collected from New York Times clearly show that our approach outperforms the conventional LDA in terms of both privacy and perplexity.

References

[Abadi *et al.*, 2016] Martín Abadi, Andy Chu, Ian J. 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, 2016.

[Anthes, 2010] Gary Anthes. Topic models vs. unstructured data. *Communications of the ACM*, 53(12):16–18, 2010.

[Blei and McAuliffe, 2007] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, pages 121–128, 2007.

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

[Blei *et al.*, 2010] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7, 2010.

[Das *et al.*, 2007] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM, 2007.

[Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Dong *et al.*, 2014] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 601–610, 2014.

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

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

[Dwork, 2011] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[Jelinek *et al.*, 1977] Frederick Jelinek, Robert Leroy Mercer, Lalit R Bahl, and J K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62, 1977.

[Krestel *et al.*, 2009] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 61–68, 2009.

[Liu *et al.*, 2012] Qiwen Liu, Tianjian Chen, Jing Cai, and Dianhai Yu. Enlister: baidu's recommender system for the biggest chinese q&a website. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 285–288. ACM, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Ramage *et al.*, 2009] Daniel Ramage, David Leo Wright Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 248–256, 2009.

[Singhal, 2001] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[Steyvers and Griffiths, 2007] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[Steyvers *et al.*, 2004] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas L. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 306–315, 2004.

[Wang *et al.*, 2009] Chong Wang, Bo Thiesson, Christopher Meek, and David M. Blei. Markov topic models. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 583–590, 2009.

[Yan *et al.*, 2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *22nd International World Wide Web Conference, WWW '13*, pages 1445–1456, 2013.

[Zhou *et al.*, 2008] Shibin Zhou, Kan Li, and Yushu Liu. Text categorization based on topic model. In *Rough Sets and Knowledge Technology, Third International Conference, RSKT*, pages 572–579, 2008.