Federated Topic Discovery: A Semantic Consistent Approach

Yexuan Shi and Yongxin Tong, *Beihang University, Beijing, 100191, China*
Zhiyang Su, *Hong Kong University of Science and Technology, Hong Kong, 999077, China*
Di Jiang, *WeBank, Zhenzhen, 440305, China*
Zimu Zhou, *Singapore Management University, 188065, Singapore*
Wenbin Zhang, *Beihang University, Beijing, 100191, China*

General-purpose topic models have widespread industrial applications. Yet high-quality topic modeling is becoming increasingly challenging because accurate models require large amounts of training data typically owned by multiple parties, who are often unwilling to share their sensitive data for collaborative training without guarantees on their data privacy. To enable effective privacy-preserving multiparty topic modeling, we propose a novel federated general-purpose topic model named private and consistent topic discovery (PC-TD). On the one hand, PC-TD seamlessly integrates differential privacy in topic modeling to provide privacy guarantees on sensitive data of different parties. 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 demonstrate its superiority over the state-of-the-art general-purpose topic models.

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. In fact, general-purpose topic models such as latent Dirichlet allocation (LDA)[1] have become the de facto in many industrial applications.[2]

Despite its widespread adoption, topic modeling faces a new challenge in the era of big data. Learning an accurate generic-purpose topic model requires large amounts of training data, which is typically owned by multiple industrial parties. For example, several hospitals need to categorize their medical records by topic modeling. Since these data may contain sensitive information, data owners are usually reluctant to share their data for collaborative topic model learning without guarantees on their data privacy. The enforcement of the General Data Protection Regulation (https://gdpr-info.eu/) further sharpens the need for privacy-preserving multiparty topic modeling, since collaborative modeling without privacy protection may now even be considered illegal.

A conceptual solution to privacy-preserving multiparty machine learning is federated learning,[3–5] which aims to provide quantified privacy guarantees such as differential privacy,[6] while still allowing effective collaborative model training among multiple parties. The principle to ensure differential privacy is to add controlled noise to the raw data, which may impair the accuracy of model learning. Hence, remedies to recover model accuracy are also necessary. Despite the generic concept, it needs dedicated technical design to realize federated learning of topic models. This is because there are no universal data perturbation mechanism and model accuracy recovery methods. Hence, new techniques tailored for topic models are compulsory.

In this article, we propose Private and Consistent Topic Discovery (PC-TD), a new federated general-purpose topic model. To protect data privacy, we

1541-1672 © 2020 IEEE
Digital Object Identifier 10.1109/MIS.2020.3033459
Date of publication 26 October 2020; date of current version 22 October 2021.

devise a data perturbation mechanism that ensures differential privacy and can be seamlessly integrated into topic modeling. To retain model accuracy, 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 semantic 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 article are as follows.

- We propose a novel federated general-purpose topic model named PC-TD, which effectively protects data privacy with proven guarantees.
- We design techniques to retain the accuracy of topic modeling by considering global and local semantic consistency.
- We conduct extensive experiments on real-world datasets to evaluate the proposed methods. The results demonstrate the validity and superiority of PC-TD.

Compared to our preliminary version, this article makes the following new contributions. 1) We study the federated scenario of topic modeling, where documents are owned by multiple industrial parties. 2) We extend the standalone topic modeling method to a federated framework. 3) We conduct new evaluations on real-world dataset. The rest of this article is organized as follows. In the "Related Work" section, we review the related work. Then, we elaborate the technical details of PC-TD in the "Private and Consistent Topic Discovery" section. We present the experimental evaluations in the "Experiments" section and, finally, conclude this article.

RELATED WORK

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

Federated Learning

Federated learning is a privacy-preserving collaborative learning paradigm, which can co-construct the model with multiple participants. During the training process, the data privacy of participants can be held. Federated learning is proposed by Google[7] for training models collaboratively on Android mobile phones and extended by Yang *et al.* in 2019.[3] Because of the reasonable privacy preserving property, federated learning has been applied gradually to industrial applications such as language modeling of mobile keyboards.

Topic Modeling

Topic modeling[8] aims to find a series of abstract "topics" in a set of documents. Within the topic modeling framework, we can represent each document by the topics and cluster these documents according to their respective topic distributions.

Research on topic modeling dates back to the latent semantic analysis (LSA)[9] which is a 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,[10] where the latent variables are topics in documents. As an improvement of PLSA, latent Dirichlet allocation (LDA)[1] is a more general Bayesian probabilistic topic model, which models each document as a multimembership mixture of K corpus-wide topics, and each topic as a multimembership mixture of the terms in the corpus vocabulary. By applying additional constraints on the basic LDA, more variants of LDA such as Sentence LDA[11] and Labeled LDA[12] have been proposed. These topic modeling methods have been proved their applicability in industry and have been successfully applied in collaborative filtering for generating personalized recommendations in Google News[2] and real-time Q&A systems in Baidu.[13] 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.

Differential Privacy

Differential privacy[6] 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 article, 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 $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for any two

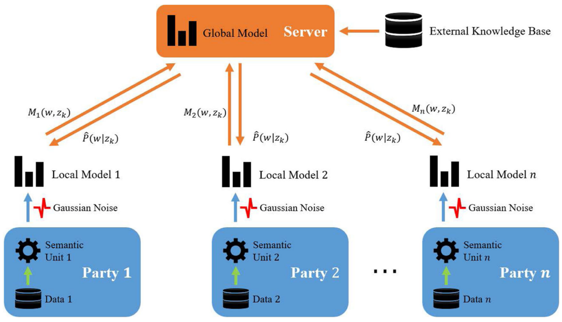

FIGURE 1. Federated framework of PC-TD.

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[6], 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 $f : \mathcal{D} \rightarrow \mathcal{R}$ 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 distribution with mean 0 and standard deviation $S_f\sigma$.

PRIVATE AND CONSISTENT TOPIC DISCOVERY

We propose a generative model to discover the topics of documents. As shown in Figure 1, to achieve the local semantic consistency, the PC-TD organizes the words of documents into semantic units. We use a federated framework to infer the latent parameters based on the semantic units and protect the data privacy of documents with Gaussian noise. 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 the "Model Assumptions" section. Then, we propose the federated inference method of our model with differential privacy in the "Federated Inference Framework" section. Next, we consider the global semantic consistency by introducing the similarity of words in the "Global Semantic Consistency" section. Finally, the analysis of privacy is provided in the "Privacy Analysis" section.

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:

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

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 (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.

Federated Inference Framework

We now propose a federated EM algorithm to infer the latent parameters of PC-TD.

E-step

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:

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

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}

$$\hat{N}_{ijw} = N_{ijw} + \Omega \tag{4}$$

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

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

The E-step can be done in each party locally.

M-step

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:

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

In order to take care of the normalization constraints, (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:

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

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

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}.

From (7), we can find that $p(z_k|d_i)$ can also be calculated in each party locally. All terms needed in $p(z_k|d_i)$ can be obtained in their own party. Thus, we focus on the second formula, $p(w|z_k)$.

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 (4) in this step and get the perturbed probability $\hat{p}(w|z_k)$

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

We can find that the calculation of $\hat{p}(w|z_k)$ will use the statistics of all the documents. Thus, it should be done by some communications of the parties and the server.

Specifically, a party t should upload the following value:

$$M_t(w, z_k) = \sum_{i=1}^{D}\sum_{j=1}^{S_i} \hat{N}_{ijw}\hat{r}_{ijk}. \tag{10}$$

For the server, after getting the statistics from all the parties, it can calculate the distribution of topic to word as follows:

$$\hat{p}(w|z_k) = \frac{\sum_{t=1}^{n} M_t(w, z_k)}{\sum_{t'=1}^{n}\sum_{w'=1}^{W} M_{t'}(w', z_k)} \tag{11}$$

where n is the number of participants.

Federated Framework

The whole framework is summarized in Algorithm 1. As the initialization, each party randomly generates $p(z_k|d_i)$ and $p(w|z_k)$ (line 1). After that, for each iteration i, each party will first get $\hat{r}_{ijk}$ in lines 4-7 (E-step). In lines 8-10, they will calculate $M_t(w|z_k)$ and push it to server. The M-step will be finished on server by collecting $M_t(w|z_k)$ and calculating $\hat{p}(w|z_k)$ (line 11). Finally, the server will push $\hat{p}(w|z_k)$ to every party in lines 12–13. Note that our algorithms will run for a prespecified number of iterations T, and with a prespecified σ; this ensures a certain level of (ϵ, δ) guarantee in the released expected sufficient statistics from Algorithm 1.

Algorithm 1. Federated Framework of PC-TD

foreach party t **do**
  Initialize $p(z_k|d_i), p(w|z_k)$ randomly;
for $i = 1, 2, \cdots, T$ **do**
  foreach party 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} ;
    Get $\hat{r}_{ijk}$ according to (5);
    Get $p(z_k|d_i)$ according to (7);
    Get $M_t(w|z_k)$ according to (10);
    push $M_t(w|z_k)$ to server;
  Merge $M_t(w|z_k)$ from every party and get $\hat{p}(w|z_k)$ according to (11);
  foreach party t **do**
    push $\hat{p}(w|z_k)$ to party t;

Global Semantic Consistency

The previous sections illustrate the local semantic consistency and federated framework of PC-TD. In this section, we discuss how to ensure global semantic consistency in PC-TD and present an approach to adapt the federated inference framework 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 and automatically built knowledge base. In this article, 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 where words or phrases from the vocabulary are mapped to vectors of real numbers. We can use a popular method, Word2vec,[14] 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

$$R_{ab} = \frac{v_a \cdot v_b}{||v_a||_2 ||v_b||_2}. \quad (12)$$

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 tradeoff factor τ. Formally, for a given R, we adjust the topic-word distribution $P(w|z_k)$ as follows:

$$p'(w|z_k) \leftarrow p(w|z_k) + \tau \frac{p(w|z_k) \sum_{i=1}^{W} R_{iw} p(i|z_k)}{P(\cdot|z_k)^T R P(\cdot|z_k)}. \quad (13)$$

In our federated framework, we can do this optimization on server and push the $p'(w|z_k)$ to each party. 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, (13) 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.

Privacy Analysis

In this section, 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[15] to account the privacy loss incurred by successive iterations of our EM algorithm.

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

$$L_{\mathcal{M}}(\mathcal{D}, \mathcal{D}', w) = \log \frac{\Pr[\mathcal{M}(\mathcal{D}, w) = o]}{\Pr[\mathcal{M}(\mathcal{D}', w) = o]}. \quad (14)$$

In PC-TD, 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

$$\alpha_{\mathcal{M}_t} = \sup_{\mathcal{D}, \mathcal{D}', w} \log \mathbb{E}[\exp(\lambda L_{\mathcal{M}_t}(\mathcal{D}, \mathcal{D}', w))]. \quad (15)$$

Since each iteration of PC-TD $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T$ adds noise independently, the log moment generating function has the following property according to[15]

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=1}^{T} \alpha_{\mathcal{M}_t}(\lambda). \quad (16)$$

Additionally, given a log moment function $\alpha_{\mathcal{M}}$,[15] shows that the corresponding mechanism $\mathcal{M}$ satisfies a range of privacy parameters (ϵ, δ) with the following equation:

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon). \quad (17)$$

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

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

Proof. From the lemma 3 in M. Abadi *et al.*,[15] 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 $\sigma = \Theta(\frac{\sqrt{T \log(1/\delta)}}{\epsilon})$. □

EXPERIMENTS

In this section, we evaluate the performance of PC-TD. In the "Experimental Setup" section, we describe the experimental setup. In the "Privacy Evaluation" section, we demonstrate the impact of privacy. Finally, we demonstrate the effectiveness of PC-TD with quantitative evaluation in the "Performance Evaluation" section.

Experimental Setup

Dataset

We evaluate our method on a corpus collected from New York Times. (https://www.kaggle.com/nzalake52/new-york-times-articles) We sample 500 documents from the news of June 26–30, 2016 as our training dataset. After removing the stopwords, we get 18,286 unique words.

Metric

We use the perplexity of documents and average topic coherence 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. The perplexity of a topic model on a set of documents is defined as

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

Topic coherence scores a single topic by measuring the semantic similarity between high scoring words in the topic. We use UMass metric to evaluate the topic coherence, which is defined by

$$\text{coherence}(Z) = \sum_{(w_i,w_j)\in Z} \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

where $D(w_i, w_j)$ counts the number of documents containing words w_i and w_j, and $D(w_i)$ counts the number of documents containing w_i.

Implementation

To simulate the federated scenario, we assume there are three participants t_1, t_2, and t_3 and split the dataset into three parts according to their release time. Specifically, t_1, t_2, and t_3 store 85, 165, and 252 documents, respectively.

For PC-TD, we 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 pretrained Word2vec by Google.[14]

We tune the parameter τ which serves as the weight parameter for global semantic consistency. 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 PC-TD with the typical general-purpose topic model LDA to verify its effectiveness. We use Markov chain Monte Carlo sampling method to train an LDA model, with parameter $\alpha = Z/50, \beta = 0.01$.

Privacy Evaluation

We first demonstrate the impact of privacy. Figure 2 shows the tradeoff between ϵ and per-word perplexity on our dataset for the different methods under a variety of conditions. As expected, the perplexity gradually decreases as the number of iterations increases in all cases. Besides, 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 $\sigma = 0.25$, the convergence of PC-TD and LDA are slightly

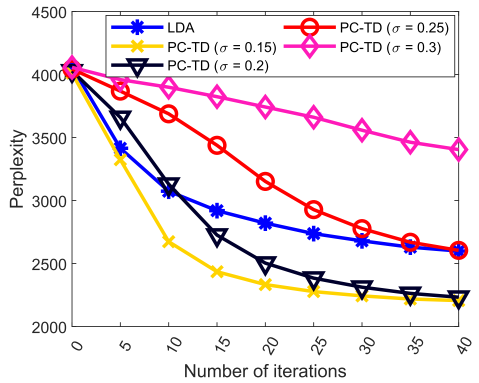

FIGURE 2. Convergence curves of varying σ.

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.

Performance Evaluation

In this section, we evaluate the performance of PC-TD by perplexity and topic coherence.

To verify the effectiveness of federated topic discovery, we compare PC-TD trained by three participants with those trained by a single participant relying on its own data. The experimental result of perplexity is shown in Figure 3. We observe that PC-TD achieves the lowest perplexity by utilizing the documents from all participants. This observation demonstrates it is meaningful to alleviate data scarcity with federated topic discovery and PC-TD is an effective method.

As the amount of topics ranges from 10 to 50, the experimental result of topic coherence is shown in Figure 4. We can observe that the average topic coherence of all the three topic models gradually increases. This phenomenon indicates that a fairly large number

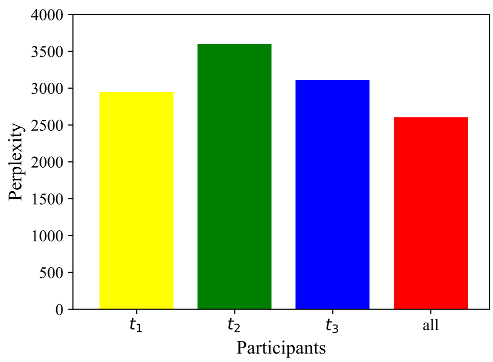

FIGURE 3. Perplexity of different participants.

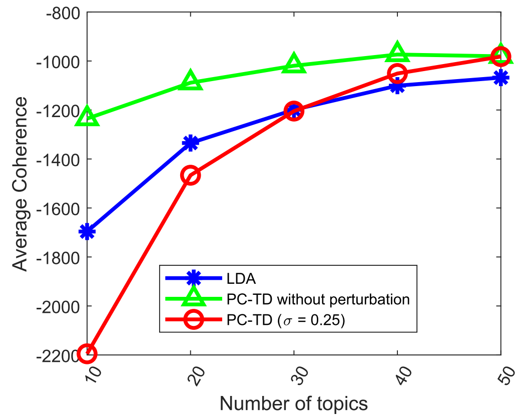

FIGURE 4. Average coherence of varying number of topics.

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. It further illustrates that the PC-TD can achieve similar or even better performance than LDA with privacy protection.

CONCLUSION

In this article, we propose a federated topic modeling approach named PC-TD to discover latent topics with semantic consistency and privacy guarantee. PC-TD utilizes a federated inference algorithm with differential privacy to ensure the privacy of sensitive documents for each party. We implement the global semantic consistency by the prior knowledge about word relations. Meanwhile, in light of the existence of semantic units such as sentences, PC-TD seamlessly integrates such local semantic consistency during its generation process. Experimental results on real datasets show that our approach outperforms the conventional LDA in terms of both privacy and performance.

ACKNOWLEDGMENTS

The authors are grateful to anonymous reviewers for their constructive comments. This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0101100, and in part by the National Science Foundation of China (NSFC) under Grant No. 61822201 and U1811463.

REFERENCES

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

2. A. Das, M. Datar, A. Garg, and S. Rajaram, “Google news personalization: Scalable online collaborative filtering,” in *Proc. Int. World Wide Web Conf. Committee*, 2007, pp. 271–280.
3. Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 12:1–12:19, 2019.
4. Y. Wang, Y. Tong, and D. Shi, “Federated latent Dirichlet allocation: A local differential privacy based framework,” in *Proc. Assoc. Adv. Artif. Intell.*, 2020, pp. 6283–6290.
5. D. Jiang *et al.*, “Federated topic modeling,” in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1071–1080.
6. C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3/4, pp. 211–407, 2014.
7. J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” 2016, *arXiv:abs/1610.05492*.
8. Y. Tong, C. C. Cao, and L. Chen, “TCS: Efficient topic discovery over crowd-oriented service data,” in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 861–870.
9. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
10. T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.
11. Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis,” in *Proc. 4th Int. Conf. Web Search Web Data Mining*, 2011, pp. 815–824.
12. D. Ramage, D. L. W. Hall, R. Nallapati, and C. D. Manning, “Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 248–256.
13. D. Jiang, Y. Tong, and Y. Song, “Cross-lingual topic discovery from multilingual search engine query log,” *ACM Trans. Inf. Syst.*, vol. 35, no. 2, pp. 9:1–9:28, 2016.
14. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2013, *arXiv:1301.3781*.
15. M. Abadi *et al.*, “Deep learning with differential privacy,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

YEXUAN SHI is currently working toward the Ph.D. degree from Beihang University, Beijing, China. His major research interests are federated data analysis and crowdsourcing. Contact him at skyxuan@buaa.edu.cn.

YONGXIN TONG received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include big spatio-temporal data analytics, crowdsourcing, crowd intelligence, federated learning, privacy preserving data analytics, and uncertain data management. He is the corresponding author of this article. He is a member of the IEEE. Contact him at yxtong@buaa.edu.cn.

ZHIYANG SU received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, Hong Kong. His research interests include natural language processing, machine learning, software-defined networking, and cloud computing. Contact him at zsuab@ust.hk.

DI JIANG received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a Senior Research Scientist with WeBank AI. His research interests include information retrieval, natural language processing, and massive data management. Contact him at dijiang@webank.com.

ZIMU ZHOU received the B.E. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2015. He is currently an Assistant Professor with the School of Information Systems, Singapore Management University, Singapore. His research focuses on mobile and ubiquitous computing. Contact him at zimuzhou@smu.edu.sg.

WENBIN ZHANG is currently working toward a master's degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. His major research interests are crowdsourcing and spatio-temporal data management. Contact him at zhangwenbin@buaa.edu.cn.