



Cross-collection latent Beta-Liouville allocation model training with privacy protection and applications

Zhiwen Luo¹ · Manar Amayri^{1,2} · Wentao Fan³ · Nizar Bouguila¹

Accepted: 29 November 2022 / Published online: 13 January 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Cross-collection topic models extend previous single-collection topic models, such as Latent Dirichlet Allocation (LDA), to multiple collections. The purpose of cross-collection topic modeling is to model document-topic representations and reveal similarities between each topic and differences among groups. However, the restriction of Dirichlet prior and the significant privacy risk have hampered those models' performance and utility. Training those cross-collection topic models may, in particular, leak sensitive information from the training dataset. To address the two issues mentioned above, we propose a novel model, cross-collection latent Beta-Liouville allocation (ccLBLA), which operates a more powerful prior, Beta-Liouville distribution with a more general covariance structure that enhances topic correlation analysis. To provide privacy protection for the ccLBLA model, we leverage the inherent differential privacy guarantee of the Collapsed Gibbs Sampling (CGS) inference scheme and then propose a hybrid privacy protection algorithm for the ccLBLA model (HPP-ccLBLA) that prevents inferring data from intermediate statistics during the CGS training process without sacrificing its utility. More crucially, our technique is the first attempt to use the cross-collection topic model in image classification applications and investigate the cross-collection topic model's capabilities beyond text analysis. The experimental results for comparative text mining and image classification will show the merits of our proposed approach.

Keywords Cross-collection model · Beta-Liouville prior · Topic correlation · Comparative text mining · Image classification · Differential privacy

1 Introduction

As social media platforms proliferate, our internet collects unprecedented information from large-scale applications, making extracting knowledge and patterns from large and complex data sets more critical. Some researchers find that “comparative thinking” is the most effective way to improve learning knowledge and some real-world applications [1, 2]. So far, more and more people have begun to pay attention to privacy using real-world applications. However, most current machine learning models with comparative thinking may expose the sensitive text information in the training data, thus causing significant privacy concerns [3]. Therefore, researching efficient machine learning techniques with comparative thinking to handle massive

data collections, such as text documents and images, and addressing privacy issues in these techniques is essential.

In unsupervised topic modeling, text documents and images are generalized as documents manipulated using count vectors according to the Bag of Words (BOW) approach. The objective is to construct meaningful topics to efficiently predict unseen documents in information retrieval and document classification tasks. In further detail, topics represent the intermediate low-dimensional representations of documents [4–7]. A well-known topic model is Latent Dirichlet Allocation (LDA) [4] incorporating the Dirichlet distribution as conjugate prior to the multinomial distribution. In the LDA model, documents appear as a combination of topics, and topics are vocabulary distributions. Moreover, LDA is frequently used as a dimensionality reduction tool to examine documents by topic and extract useful information from a large amount of unstructured data. Recently, LDA has been the subject of various extension techniques [6] to cluster text documents, and images [8–10] through their latent topics based on words (or visual words in case of images) co-occurrence. Even though there are

✉ Zhiwen Luo
zhiwen.luo@mail.concordia.ca

Extended author information available on the last page of the article.

many research efforts on topic modeling to enhance LDA model in text mining [3, 6, 7], most of the existing research has failed to compare document collections. However, more and more real-world applications need to understand the relationships among various collections such as decision-making task [11, 12], interactive learning [13], and event summarization [1, 14, 15]. Moreover, privacy preservation has also drawn great attention in many real-world applications. Nevertheless, the training process of many topic models may expose sensitive textual information of training data [3], which leads to serious privacy issues when applying the topic model to real-world applications. In this work, we would like to answer the following question: How can we facilitate a comparative analysis of document collections with privacy protection using the topic model?

Comparative analysis of document collections is highly desirable in various real-world applications. To achieve this goal, we first need to reveal common and distinctive topics. Figure 1 presents an example of the results obtained by our proposed ccLBLA model on newspapers related to the COVID-19 virus in 2020. Specifically, Fig. 1(a) shows distinctive topics of news in the USA, suggesting that the most important words are related to the national economy, such as “country”, “stock”, and “product”. On the other hand, as shown in Fig. 1(c), most words from UK’s specific topics represent the coronavirus situation, such as “report”, “case”, and “outbreak”. From Fig. 1(b), we can conclude that both share common concerns with the spread of coronavirus such as “virus”, “China”, “spread”, and “global”. Because most topic models only consider a single collection of texts, they are insufficient for comparing the similarities and differences of multiple collections. This problem also makes it impossible to discover the potential common topics in all corpora. To deal with this challenge, some scholars have offered cross-collection topic models for comparative text mining tasks as a solution [2, 16–18], which entail the extraction of useful information from several datasets.

Second, it is important to capture good correlations between common and unique topics in comparative text mining. Many existing topic models and cross-collection topic models continue to rely on the traditional LDA with Dirichlet distribution as conjugate prior to multinomial distribution. Despite its popularity in topic modeling because of its convenience in computing, the Dirichlet distribution has drawbacks. Due to its restrictive negative covariance matrix structure [19–23], the Dirichlet distribution cannot capture the correlation between topics. In contrast, the topic correlation is a critical feature for recapitulating the relation of multiple collections in the cross-collections topic model. Because of the limitation of Dirichlet prior, researchers began looking into other more flexible priors, such as Generalized Dirichlet (GD) distribution [24, 25] and Beta-Liouville (BL) distribution [10, 26] in replacement of Dirichlet distribution to tackle the problem of correlation in the traditional topic model.

Third, a model capable of comparative text analysis should have good privacy protection for the whole training process. As shown in Fig. 2, traditional topic models such as LDA may be trained on datasets including sensitive information. However, these topic models will inexorably memorize some critical knowledge about the dataset after the training process, which is characteristic of typical machine learning techniques. Nevertheless, it has been demonstrated that some attack methodologies may successfully extract sensitive information from training datasets in machine learning models. Evidence suggests that model inversion attack [27] and membership inference attack [28], according to recent findings, can both pose a privacy issue for machine learning models in different ways. Dwork et al. [29] proposed the differential privacy (DP) strategy for privacy preservation in machine learning models to address these privacy problems. Because differential privacy provides a mathematical framework for measuring the security of several machine learning techniques, there has been an increasing interest in applying differential privacy in topic



Fig. 1 Topic summaries of newspapers published related to COVID-19 in 2020

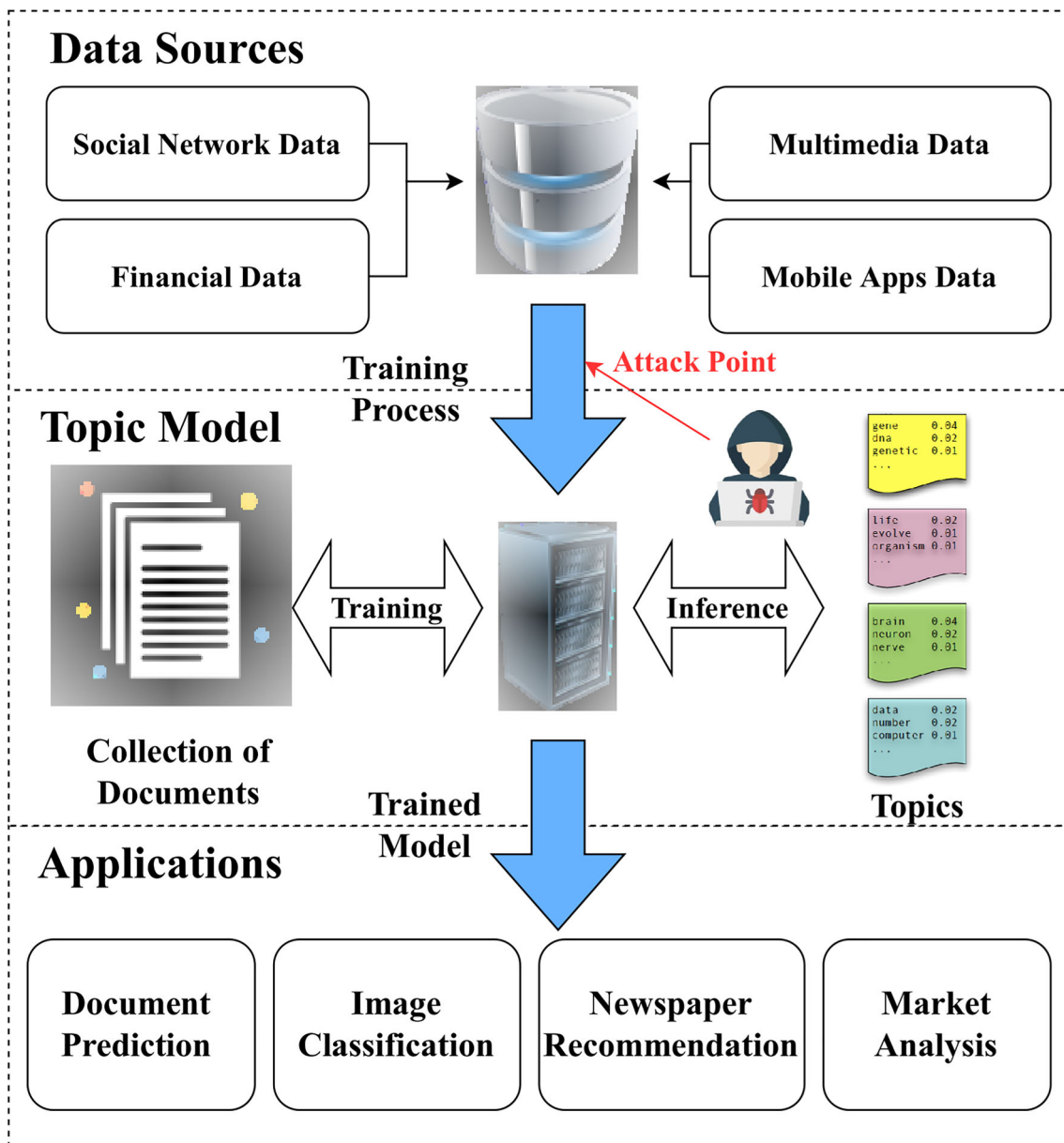


Fig. 2 Application scenarios of topic model training

models such as LDA. For example, the authors in [30] attempted to inject additional noise into the training process of LDA to create a differential privacy guarantee in Collapsed Gibbs sampling (CGS) inference scheme [31] for centralized training datasets. However, these approaches implicitly assume that the LDA model is trained on centralized datasets by a trustworthy server. But, the central server may also act as an adversary and steal training datasets (Attack Point in Fig. 2). Besides, many earlier investigations of DP in LDA training have not dealt with the inherent privacy of the CGS inference of the LDA scheme. In contrast, the CGS inference scheme provides some level of privacy guarantee because of its intrinsic unpredictability,

and uncertainty [32, 33]. Although several studies [3] recognized the intrinsic privacy of the CGS inference scheme, they only focused on the LDA model.

In this article, to answer the question and alleviate the restrictions described above, we first propose a cross-collection latent topic model (ccLBLA) with more flexibility and scalability by offering a better prior distribution, the Beta-Liouville distribution. ccLBLA is a hierarchical Bayesian model developed to learn common and distinct topics from document collections at the same time. This is also a novel enhanced cross-collection topic model that combines the state-of-the-art cross-collection topic model [17] and the completely LBLA model [10, 26].

To address privacy and utility issues, we present a hybrid privacy-preserving approach of the ccLBLA model (HPP-ccLBLA) based on a systematic analysis of the intrinsic differential privacy guarantee of topic model training on centralized datasets by taking advantage of HDP-LDA model [3]. Experimental results in text document analysis and image classification demonstrate the merits of our novel approach. This paper's overall contributions can be summarized as follows:

- The proposed ccLBLA model is a novel hierarchical Bayesian model to focus on identifying common and distinctive topics among multiple datasets emerging from a wide range of applications. The generative process of LDA [4], LBLA [10, 24, 26], and the ccLDA [17] have all been improved by the new model.
- Our proposed model replaces Dirichlet distribution with Beta-Liouville (BL) distribution as a more flexible prior to overcome its shortcomings related to document and corpus parameters. Therefore, this novel model prioritizes the topic correlation to improve the performance of comparative text mining with a more generic covariance structure that does not rely on the Dirichlet distribution's restrictive negative covariance.
- We investigate the intrinsic privacy of the CGS-based ccLBLA model by utilizing the HDP-LDA model [3]. We present the HPP-ccLBLA model, the first privacy-preserving cross-collection topic modeling technique based on the exponential mechanism of differential privacy and the consistency of the CGS inference scheme. Compared with the state-of-the-art privacy-preserving topic model (HDP-LDA), our proposed model can discover topics' similarities and differences across multiple collections. Indeed, our HPP-ccLBLA model utilizes a more flexible prior (BL) and safeguards the CGS-based training process of the ccLBLA model with centralized training datasets.
- We deliver the first study on adopting the cross-collection topic model for image classification application by processing each image as a separate document using the Bag of Visual Words methodology [8–10].
- To validate the new model's performance, our experiments include a variety of real-world datasets such as newspapers, academic articles, customer reviews, and natural scene images. Our studies indicate that our proposed model (ccLBLA) can achieve a much higher generalization performance in comparative text mining and document and image classification. Furthermore, the HPP-ccLBLA strategy can obtain a good model utility while maintaining sufficient privacy guarantees.

The paper is organized as follows. Section 2 discusses the related work regarding topic models and differential

privacy. Section 2 will also review the base information of the LDA, the ccLDA, and the LBLA models and then analyze the relationship between topic modeling and differential privacy. We present our ccLBLA model and propose the intrinsic privacy study of CGS-based ccLBLA (HPP-ccLBLA) in Section 3. Section 4 is devoted to the experimental results. Section 5 concludes this paper and gives our conclusions.

2 Related works and background

This section reviews the existing literature related to the problem studied in this article. There are three main branches of research related to this work, traditional topic modeling techniques [4, 10, 26], cross-collection topic modeling [16–18, 34], and topic modeling with privacy protection [3, 35, 36]. Although these existing models are related to the HPP-ccLBLA model, none can handle the comparison between different collections under privacy protection during the training process. The full comparison between the previous techniques and our model is provided by Table 1.

2.1 Traditional topic models

LDA model [4], as an extension of the pLSI model [37], is a complete generative probabilistic model that improves generalization capability by introducing Dirichlet prior to overcome the overfitting and the difficulty in predicting documents probability problems. In particular, the LDA model utilizes the BOW method for various applications, including text modeling and computer vision, and its generative process has been extensively documented in several articles [4, 6]. Even though the LDA model plays a fundamental role in topic modeling and many machine learning applications, numerous studies [38, 39] have shown that the constraints of Dirichlet prior hamper the LDA's performance. Bakhtiari and Bouguila [24] showed that using more flexible priors such as Generalized Dirichlet (GD) and Beta-Liouville (BL) distributions in document parameters can improve the performance of the LDA model in text modeling and computer vision applications. Moreover, Ihou and Bouguila [9, 10] proposed new models that replace the Dirichlet distribution on both the corpus and the document parameters with GD or BL priors, and their experiments show that those more flexible priors can perform well in topic correlated environments. Compared with the GD distribution, the BL distribution has fewer parameters and is also a generalization of the Dirichlet distribution [10]. Therefore, the BL distribution is superior for LDA-based topic modeling because of its computation efficiency.

Table 1 Comparison between the new HPP-LBLA model and other schemes

	Capability	Privacy protection
LDA	It is based on the Dirichlet prior, which is found to be very limited. This model conceptually focused on one single collection which is inadequate for comparative text analyses.	Without any privacy protection during the whole training process
LBLA	It replaces the Dirichlet distribution on both the corpus and the document parameter with BL prior, shown to be more flexible than the Dirichlet distribution. This model is inadequate for the comparative analysis of document collections.	Without any privacy protection during the whole training process
ccLDA	It can discover topics across multiple text collections and model their similarities and differences, but its performance suffers from limitation of Dirichlet distribution.	Without any privacy protection during the whole training process
HDP-LDA	It doesn't have ability to do comparative analysis of document collections. Also, it is a Dirichlet-based model (as a result, it is limited)	Take advantage of inherent differential privacy guarantee of CGS-based LDA training on centralized datasets. This algorithm can protect all the intermediate statistics of the whole training process.
HPP-ccLBLA	It is our proposed model to improve the ccLDA and LBLA models. It automatically combines the advantage of both BL prior and the ability of comparative text analyses.	Take advantage of inherent differential privacy guarantee of CGS-based LBLA training on centralized datasets to address the privacy issue. Our proposed model can prevent data inference from intermediate statistics during training.

2.1.1 Differences in prior information

Because the Dirichlet distribution has a very restrictive negative covariance structure, it has difficulties performing in a topic correlation analysis [21]. Even though Blei et al. [40] proposed a Correlated Topic Model (CTM) to overcome such problems in the topic model by incorporating the normal logistic distribution. However, this distribution is not a conjugate prior to the multinomial distribution [40, 41], so the CTM is very challenging to implement. Recent breakthroughs in topic modeling have highlighted the necessity for more flexible priors. Beta-Liouville distribution is becoming increasingly popular. For Beta-Liouville distribution, in dimension (K) space, $\phi = (\phi_1, \dots, \phi_K)$ and $\sum_{k=1}^K \phi_k = 1$, with hyperparameter vector $\zeta = (\alpha_1, \alpha_2, \dots, \alpha_{(K-1)}, \alpha, \beta)$ is defined by:

$$p(\phi | \zeta) = \frac{\Gamma(\sum_{k=1}^{K-1} \alpha_k) \Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \prod_{k=1}^{K-1} \frac{\phi_k^{\alpha_k - 1}}{\Gamma(\alpha_k)} \times \left(\sum_{k=1}^{K-1} \phi_k \right)^{\alpha - \sum_{k=1}^{K-1} \alpha_k} \left(1 - \sum_{k=1}^{K-1} \phi_k \right)^{\beta - 1} \quad (1)$$

ϕ is the K -dimensional multinomial parameter drawn from the BL(ϕ) distribution. When the generator has a Beta distribution with parameters $\sum_{k=1}^{K-1} \alpha_k$ and α_K , the

Beta-Liouville distribution is reduced to Dirichlet distribution [21]. Thus, Beta-Liouville includes the Dirichlet distribution as a particular case. Compared with the Dirichlet distribution, the Beta-Liouville distribution has more parameters and is more flexible for several applications [21]. The mean and the variance of the Beta-Liouville distribution are given by:

$$E[\phi_k] = \frac{\alpha}{\alpha + \beta} \frac{\alpha_k}{\sum_{k=1}^{K-1} \alpha_k} \quad (2)$$

$$var(\phi_k) = \left(\frac{\alpha}{\alpha + \beta} \right)^2 \frac{\alpha_k(\alpha_k + 1)}{\left(\sum_{k=1}^{K-1} \alpha_k \right) \left(\sum_{k=1}^{K-1} \alpha_k + 1 \right)} - E[\phi_k] \frac{\alpha_k^2}{\left(\sum_{k=1}^{K-1} \alpha_k \right)^2} \quad (3)$$

and the covariance between ϕ_i and ϕ_j is as following [26]:

$$Cov(\phi_i, \phi_j) = \frac{\alpha_i \alpha_j}{\sum_{k=1}^{K-1} \alpha_k} \left(\frac{\frac{\alpha+1}{\alpha+\beta+1} \frac{\alpha}{\alpha+\beta}}{\sum_{k=1}^{K-1} \alpha_k + 1} - \frac{\frac{\alpha}{\alpha+\beta}}{\sum_{k=1}^{K-1} \alpha_k} \right) \quad (4)$$

According to (4), the covariance matrix of the Beta-Liouville distribution is not strictly negative like the Dirichlet distribution because two variables with the same

mean value can have different variances. Therefore, the Beta-Liouville distribution has a more general covariance structure. What is more, the Beta-Liouville distribution is also a conjugate prior of the multinomial distribution. The above advantages make the BL distribution more powerful and practical in topic modeling. Hence, introducing BL distribution to replace the Dirichlet prior in the LDA model improves topic correlation and is convenient for practical applications. Consequently, the LBLA model can provide more practical capabilities than the original LDA model and includes it as a particular case [10, 26].

Because this study is an extension of the LBLA model [10], it is necessary to summarize the generative process of the original LBLA graphical model. In the LBLA scheme, there are three main generating phases:

- For each document \mathbf{d} , draw a topic mixture θ_d from $BL(\xi)$.
- Draw a corpus multinomial word distribution ϕ_k from $BL(\epsilon)$ for each topic \mathbf{z} .
- For each word w_i in \mathbf{d} :
 - Choose a topic z_i from $Mult(\theta_d)$
 - Choose a word w_i from $Mult(\phi_k)$

2.2 Cross-collection topic model

So far, natural language processing, computer vision, pattern recognition, and other disciplines are increasingly using the LDA model and its extensions, such as the LBLA. Due to different practical problems, there are more and more different new topic models inspired by LDA. For example, Zhai et al. [16] introduced a topic model, the Cross-Collection Mixture model (ccMix), based on the pLSI model [37], for handling comparative text mining problems. Due to the limitation of the ccMix model, Paul and Girju [17] presented a Cross-Cultural LDA (ccLDA) model, which is the extension of LDA and ccMix frameworks. The cross-collection topic models try to extract the common information from all collections and figure out what is unique to a specific collection in different dataset collections. As the state-of-the-art cross-collection topic model, the ccLDA model provides better generalization capabilities and less relies on user-defined parameters. Moreover, ccLDA model shares assumption with the LDA-Collection [34] and Topical N-Gram models [42]. Those models assume that each word can be generated from two different distributions. Based on ccLDA model, Julian and Ralf [18] offered an entropy-based ccLDA model which distinguishes collection-independent and collection-specific words according to information entropy. The BOW assumption is maintained in both ccLDA and entropy-based ccLDA models; thus, each word depends on the different dataset collection.

2.2.1 Differences in the generative process

The ccLDA model can detect topics among multiple data collections and differences between those data collections. Specifically, the ccLDA model first samples a collection c (observable data), then chooses a topic \mathbf{z} and flips a coin x to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of x is 1 or 0 and comes from a Beta distribution. The generative process of the ccLDA model is based on the following steps:

- Draw a collection-independent multinomial word distribution ϕ_z from $Dirichlet(\beta)$ for each topic \mathbf{z}
- Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from $Dirichlet(\delta)$ for each topic \mathbf{z} and each collection \mathbf{c}
- Draw a Bernoulli distribution $\psi_{z,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic \mathbf{z} and each collection \mathbf{c}
- For each document \mathbf{d} , choose a collection \mathbf{c} and draw a topic mixture θ_d from $Dirichlet(\alpha_c)$. Then for each word w_i in \mathbf{d} :
 - Sample a topic z_i from $Mult(\theta_d)$
 - Sample x_i from $Bernoulli(\psi_{z,c})$
 - If $x_i = 1$, sample a word w_i from $Mult(\sigma_{z,c})$ else $x_i = 0$, sample a word w_i from $Mult(\phi_z)$

Although the ccLDA model generalizes the LDA model by adding comparative analyses of different data collections, the limitations of the Dirichlet distribution to capture the correlation between topics have impeded the performance of the ccLDA model and its extensions in various text analysis or classification applications. The state-of-the-art LBLA model improves the generative data process and effectively captures the semantic relationships between topics. Integrating the BL distribution and ccLDA model can naturally improve the cross-collection topic model's performance. However, the topic models mentioned above are without any privacy protection. Specifically, those models can not defend against adversaries with full knowledge of the training process, posing severe privacy concerns.

2.3 Topic model with privacy protection

Many machine learning models [43–45] have applied differential privacy to address privacy attack vulnerabilities by perturbing the model during different training parts. Specifically, there are a lot of different ways to adopt differential privacy in ML models such as output perturbation, objective perturbation [46], intermediate perturbation [47, 48] and input perturbation. In recent years, there has been an increasing interest in input perturbation, and local

differential privacy [49], demonstrating that enormous randomized crowdsourced data may leak valuable statistics. The input perturbation can guarantee privacy by eliminating the premise of trustworthy servers.

As a classic machine learning approach, topic models also can achieve differential privacy protection by perturbing the intermediate parameters during the training process via input perturbation. For instance, by perturbing the sampling distribution in the final iteration, Zhu et al. [30] suggested a DP guarantee CGS-LDA model. While performing a variational Bayesian inference scheme, Park et al. [48] used differential privacy in LDA by perturbing the adequate statistics data in each iteration. Like the above works, Decarolis et al. [50] altered the intermediate statistics in the spectral methodology. However, those DP guarantee methods [30, 48, 50] cannot tackle the problem of untrustworthy data curators by design. Wang et al. [32] established a locally private LDA strategy for a federated environment, but this approach is not a generic solution to the standard approach for the batch-based LDA model.

2.3.1 Intrinsic privacy of CGS-LDA algorithm

Recent improvements [32, 51] in intrinsic privacy have heightened that the Bayesian sampling can generate the inherent privacy guarantee without introducing further noise to sample statistics variables. Foulds et al. [33] expanded on this work, concluding that the generic MCMC mechanism may also process inherent privacy guarantees and acquire privacy protection similar to the Laplace mechanism. Then, Zhao et al. [3] proposed a differential privacy solution for traditional batch LDA training, a hybrid privacy-preserving algorithm (HDP-LDA), which injects the noise to obfuscate the word count in each training iteration and takes advantage of the inherent randomness of Markov Chain Monte Carlo (MCMC) techniques. The inherent privacy guarantee is an essential feature of the CGS-LDA method.

Measuring the inherent privacy guarantee in a topic model such as the LDA model is still challenging. Even though HDP-LDA [3] has been demonstrated to be effective and outperforms some methods mentioned above [30, 48, 50], this scheme still suffers from the restriction of Dirichlet prior and insufficient for comparative datasets analysis. In this paper, we present a cross-collection topic model that overcomes the limitations of Dirichlet prior by adopting a more flexible prior and using differential privacy for privacy preservation, which can secure sensitive information from attackers who are aware of the training process.

3 The model

This section mainly describes our Cross-Collection Latent Beta-Liouville Allocation (ccLBLA) model and the hybrid private ccLBLA (HPP-ccLBLA) framework.

3.1 Problem statement

Our HPP-ccLBLA model is a generative probabilistic model for analyzing multiple datasets. The basic assumption is that documents are represented as random mixtures over latent topics, where each topic is a distribution over words. Specifically, the common topics are shared with all collections, while distinctive topics belong to a specific collection. This approach aims to generate common and specific topics under privacy protection during the training process.

Our approach integrates LBLA [10, 26] and ccLDA [17], and HDP-LDA [3] as a privacy preservation cross-collection topic model that takes BL distribution on both document and corpus parameters. We will start with a study of the generative process of the fundamental ccLBLA model. Then, we introduce our extension of the ccLBLA model to the hybrid privacy-preserving learning scheme applying the

Fig. 3 Graphical representation of ccLBLA

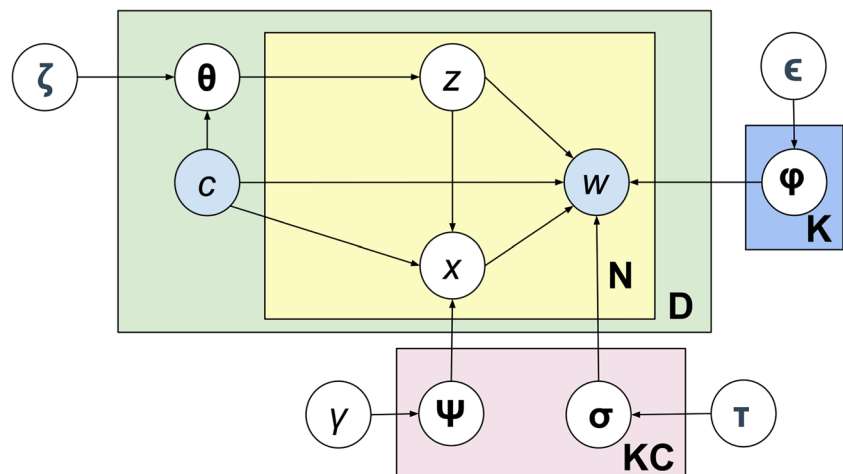


Table 2 Model variables and definitions

C - total number of collections
D - total number of documents
W - total number of words in each document
K - total number of topics
$\mathbf{w} = w_{ij}$ - observed words
$\mathbf{z} = z_{ij}$ - latent variables
θ_j - mixing proportions
ϕ_k - corpus parameters in collection-independent distribution
$\sigma_{k,c}$ - corpus parameters in collection-specific distribution
$\psi_{k,c}$ - parameter in Bernoulli distribution
$\theta_j \sim BL(\zeta_c)$ - Beta-Liouville distribution
$\phi_k \sim BL(\epsilon)$ - Beta-Liouville distribution
$\sigma_{k,c} \sim BL(\tau_c)$ - Beta-Liouville distribution
$\psi_{k,c} \sim Beta(\gamma_0, \gamma_1)$ - Beta distribution
$x \sim Bernoulli(\psi_{ck})$ - Bernoulli distribution
$z_{jk}/\theta_{jk} \sim Mult(\theta_j)$ - multinomial distribution
$x_{jk}/z_{jk}, \phi_k, x = 0 \sim Mult(\phi_k)$ - multinomial distribution
$x_{jk}/z_{jk}, \sigma_{k,c}, x = 1 \sim Mult(\sigma_{ck})$ - multinomial distribution

method on the HDP-LDA model [3], which includes cross-collection and CGS inference method with BL distribution prior. The topic graphical model (Fig. 3) is described by a list of variables. It demonstrates the conditional dependence structure between these variables. The variables in this paper are provided in Table 2 to allow readers to understand our models and follow the inference steps easily.

3.2 The cross-collection LBLA model

3.2.1 The generative process of ccLBLA model

For the complete analysis of the ccLBLA model, we will first state the generative process of the ccLBLA model, and then we will develop the inference equations when using the collapsed Gibbs sampling for learning (CGS-ccLBLA). The ccLBLA model first samples a collection \mathbf{c} (observable data), then choose a topic \mathbf{z} and flips a coin \mathbf{x} to determine whether to draw from the shared topic-word distribution or the topic’s collection-specific distribution. The probability of \mathbf{x} is 1 or 0 and is supported to be generated from a Bernoulli distribution.

- Draw a collection-independent multinomial word distribution ϕ_k from $BL(\epsilon)$ for each topic \mathbf{z}
- Draw a collection-specific multinomial word distribution $\sigma_{k,c}$ from $BL(\tau_c)$ for each topic \mathbf{z} and each collection \mathbf{c}
- Draw a Bernoulli distribution $\psi_{k,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic \mathbf{z} and each collection \mathbf{c}

- For each document \mathbf{d} , choose a collection \mathbf{c} and draw a topic mixture θ_d from $BL(\zeta_c)$. Then for each word w_i in \mathbf{d} :
 - Sample a topic z_i from $Mutl(\theta_d)$
 - Sample x_i from $Bernoulli(\psi_{k,c})$
 - If $x_i = 1$, sample a word w_i from $Mutl(\sigma_{k,c})$ else $x_i = 0$, sample a word w_i from $Mutl(\phi_k)$

3.2.2 Inference

Because the estimation of the posterior distribution in Bayesian topic models is intractable, inference methods such as VB and MCMC have become the standard choices to estimate the latent topics and the model parameters. For the inference of the ccLBLA model, we choose collapsed space representation because it contributes to the performance of batch models [31, 52]. Details about collapsed Gibbs sampling inference will be provided. Specifically, ζ_c carries the document hyperparameters α_c and β_c , ϵ includes the collection-common hyperparameters η and λ , as well as the variable τ_c holds collection-specific hyperparameters η_c and λ_c . In more detail, $(\zeta_c) = (\alpha_{c1}, \dots, \alpha_{c(K-1)}, \alpha_c, \beta_c)$ means the hyperparameter set of a document with class c , and K is the number of topics. The collection-independent hyperparameter variable ϵ can be extended as $\epsilon = (\lambda_1, \dots, \lambda_{V-1}, \lambda, \eta)$ while V is the size of the vocabulary or codebook. Similarly, the collection-specific hyperparameter variable ζ_c can be expressed as $\tau_c = (\lambda_{c1}, \dots, \lambda_{c(V-1)}, \lambda_c, \eta_c)$ while V is also the size of the vocabulary. The document, topic’s collection-common, and collection-specific distribution are sampled from Beta-Liouville distributions in our scheme. Therefore, in our implementation, ζ_c is the $K - 1$ dimensional BL hyperparameter $(\alpha_{c1}, \dots, \alpha_{c(K-1)}, \alpha_c, \beta_c)$ for the document in class c in a K dimensional space. The ϵ and τ_c are the V dimensional BL hyperparameters for the vocabulary in a V dimensional space.

In collapsed space, the parameters are marginalized, leaving only the latent variables that are conditionally independent [53], and the collapsed space of latent variables is a low dimensional space as compared with joint space. Estimation in collapsed space is faster than in joint space because the parameters ϕ , σ , and θ are marginalized. The collapsed Gibbs sampling inference approach uses a Bayesian network to estimate the posterior distributions by computing expectations through a sampling process of the latent variables. The CGS is easier to implement and computationally quicker than ordinary Gibbs sampling in the joint space. Because the CGS inference does not need the usage of digamma functions, it increases computational efficiency. As a result, when the Markov chain achieves its stationary distribution, the CGS inference

accurately approximates the actual posterior distribution. The ccLDA and its extensions [17, 18] are based on CGS inference to estimate posterior distribution because of its advantages. Furthermore, in the next section, we will describe our privacy-preserving ccLBLA method by utilizing the intrinsic privacy guarantee feature of the CGS inference scheme.

3.2.3 Hidden variables

In the CGS-ccLBLA scheme, the conditional probabilities of latent variable z_{ij} are calculated by the current state of all variables except the particular variable z_{ij} being processed in the marginal joint distribution $p(\mathbf{w}, \mathbf{z} \mid x_{ij} = 0, \zeta_c, \epsilon)$ or $p(\mathbf{w}, \mathbf{z} \mid x_{ij} = 1, \zeta_c, \tau_c)$ between collection-common and collection-specific case. This algorithm applies the collapsed Gibbs sampler for topic assignments. The conditional probability of z_{ij} is $p(z_{ij} = k \mid x_i = 0, \mathbf{z}_{-ij}, \mathbf{w}, \zeta_c, \epsilon)$ or $p(z_{ij} = k \mid x_i = 1, \mathbf{z}_{-ij}, \mathbf{w}, \zeta_c, \tau_c)$. The $-ij$ represents the counts with z_{ij} excluded [53]. This conditional probability of collection-common and collection-specific is expressed as:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon)}{p(z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon)} \tag{5}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c)}{p(z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c)} \tag{6}$$

Equations 5 and 6 can be simplified as following:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon) \tag{7}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c) \tag{8}$$

In the CGS-ccLBLA model, the parameters θ , ϕ , and σ are drawn from the BL distribution. To speed up the training process, we marginalize these parameters in the collapsed space because sampling in the collapsed space is much faster than in the joint space of latent variables and parameters [10, 53]. By integrating out the parameters, Gibbs sampler's equations are obtained as expectation expressions:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) = E_{p(z_{ij}=k|x_{ij}=0, \mathbf{w}, \zeta_c, \epsilon)}[p(z_{ij} = k \mid x_{ij} = 0, z^{-ij}, \mathbf{w}, \zeta_c, \epsilon)] \tag{9}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) = E_{p(z_{ij}=k|x_{ij}=1, \mathbf{w}, \zeta_c, \tau_c)}[p(z_{ij} = k \mid x_{ij} = 1, z^{-ij}, \mathbf{w}, \zeta_c, \tau_c)] \tag{10}$$

In the collapsed space, we can integrate out θ , ϕ , σ , and ψ to get (11)–(14) according to the conjugacy of the Beta/Binomial and BL/Multinomial distributions based on the inference equations developed for CGS-ccLDA and CGS-LBLA [10, 17]. In CGS algorithm iterations, we sample new assignment of \mathbf{z} and \mathbf{x} alternately with the following equations:

$$p(z_i = k \mid x_i = 0, \mathbf{z}_{-i}, \mathbf{w}, \zeta_c, \epsilon) \propto \frac{(\alpha_{ck} + N_{jk}^{-ij})}{(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl}^{-ij})} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl}^{-ij})}{(\alpha_c + \beta_c + \sum_{l=1}^K N_{jl}^{-ij})} \times \frac{(\lambda_v + N_{kv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl}^{-ij})} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl}^{-ij})}{(\lambda + \eta + \sum_{l=1}^V N_{kl}^{-ij})} \tag{11}$$

$$p(x_i = 0 \mid x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, s, t) \propto \frac{N_{x=0}^{k,c} + \gamma_0}{N^{k,c} + \gamma_0 + \gamma_1} \times \frac{(\lambda_v + N_{kv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl}^{-ij})} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl}^{-ij})}{(\lambda + \eta + \sum_{l=1}^V N_{kl}^{-ij})} \tag{12}$$

For (11) and (12), all counts only refer to the words for which $x_i = 0$, which are the words assigned to the topic model. Specifically, N is the total number of words for which $x_i = 0$, not the total number of words in the corpus. Same for (13) and (14), the count only includes the words for which $x_i = 1$, which means that N is the total number of words for which $x_i = 1$.

$$p(z_i = k \mid x_i = 1, \mathbf{z}_{-i}, \mathbf{w}, \zeta_c, \tau_c) \propto \frac{(\alpha_{ck} + N_{jk}^{-ij})}{(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl}^{-ij})} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl}^{-ij})}{(\alpha_c + \beta_c + \sum_{l=1}^K N_{jl}^{-ij})} \times \frac{(\lambda_{cv} + N_{kv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N_{ckl}^{-ij})} \times \frac{(\lambda_c + \sum_{l=1}^{V-1} N_{ckl}^{-ij})}{(\lambda_c + \eta_c + \sum_{l=1}^V N_{ckl}^{-ij})} \tag{13}$$

$$\begin{aligned}
 & p(x_i = 1 \mid x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, \tau_c) \\
 & \propto \frac{N_{x=1}^{k,c} + \gamma_1}{N_{\cdot}^{k,c} + \gamma_0 + \gamma_1} \times \frac{(\lambda_{cv} + N_{ckv}^{-ij})}{\left(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N_{ckl}^{-ij}\right)} \\
 & \times \frac{(\lambda_c + \sum_{l=1}^{V-1} N_{ckl}^{-ij})}{(\lambda_c + \eta_c + \sum_{l=1}^V N_{ckl}^{-ij})} \tag{14}
 \end{aligned}$$

The count N_{jk}^{ij} is the number of words w_i in the document j and topic k in class c . Besides, N_{jk}^{-ij} is the total number of words in document j and topic k in class c except for the word w_i being sampled. The count $N_{kw_{ij}}^{ij}$ is the number of times the word w_{ij} appears in topic k and document j . In addition, $N_{kw_{ij}}^{-ij}$ is the number of times the word w_{ij} appears in document j and topic k except being sampled. $N_{ckw_{ij}}^{ij}$ is the number of times the word w_{ij} appears in topic k and document j in specific collection c . In addition, $N_{ckw_{ij}}^{-ij}$ is the number of times the word w_{ij} appears in document j and topic k in specific collection c except being sampled. $N_x^{k,c}$ is the number of \mathbf{x} in topic k , and collection c . \mathbf{x} should be initialized as 0 for all tokens. We initially assume that everything comes from the shared collection word distribution.

3.2.4 Multinomial parameters

For parameters estimation, the document parameter distribution is:

$$\theta_{jk} = \frac{(\alpha_{ck} + N_{jk})}{\left(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl}\right)} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl})}{(\alpha_c + \beta_c + \sum_{l=1}^K N_{jl})} \tag{15}$$

Input : Max iteration T , Prior parameters $\zeta_c, \tau_c, \epsilon$, Topic number K , Corpus D , \mathbf{w} , Class c

Output : Parameters $\theta_j, \phi_k, \sigma_{ck}$ using (15) – (17)

//Initialization

Initialize $\mathbf{z}, \mathbf{x}, N_{jk}, N_{kw}, N_{ckw}, N_x$

//Collapsed Gibbs Sampling

while iter < T **do**

for $i \in$ document D and $j \in$ in class c **do**

if $x_{ij} = 0$ **then**

 update z_{ij} using (11)

else

 update z_{ij} using (13)

end if

 update x_{ij} using (12) and (14)

 update $N_{jk}, N_{kw}, N_{ckw}, N_x$

end for

end while

Algorithm 1 Summary of CGS-ccLBLA model.

The predictive distributions of the collection-independent and collection-specific words are:

$$\phi_{kw} = \frac{(\lambda_v + N_{kv})}{\left(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl}\right)} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl})}{(\lambda + \eta + \sum_{l=1}^V N_{kl})} \tag{16}$$

$$\begin{aligned}
 \sigma_{ckw} &= \frac{(\lambda_{cv} + N_{ckv})}{\left(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N_{ckl}\right)} \\
 &\times \frac{(\lambda_c + \sum_{l=1}^{V-1} N_{ckl})}{(\lambda_c + \eta_c + \sum_{l=1}^V N_{ckl})} \tag{17}
 \end{aligned}$$

The algorithm 1 shows the summary of the CGS-ccLBLA model.

3.3 Hybrid privacy-preserving ccLBLA scheme

This section will first introduce the differential privacy and exponential mechanism. Then, we point out the limitations of existing methods in protecting the intermediate statistics in the topic model training process. Moreover, we thoroughly analyze the inherent differential privacy guarantee of CGS-ccLBLA training on centralized datasets. Finally, based on the study above, we will present a hybrid privacy-preserving method for the cross-collection topic model (HPP-ccLBLA). In the HPP-ccLBLA scheme, all the intermediate statistics of the CGS-ccLBLA model can be protected during the training process.

3.3.1 Differential privacy and exponential mechanism

Differential privacy [29] is a de-facto standard for privacy protection framework with rigorous mathematical proof. So far, DP has been widely utilized in the past to assess the privacy issue of random algorithms by comparing the mathematical differences between neighboring datasets.

Theorem 1 (Differential Privacy [29]) *A randomized mechanism $f : \mathbf{D} \rightarrow \mathbf{Y}$ offers $(\epsilon, \delta - DP)$ if for any adjacent $D, D' \in \mathbf{D}$ and $Y \in \mathbf{Y}$, there is:*

$$Pr(f(D) \in \mathbf{Y}) \leq e^\epsilon Pr(f(D') \in \mathbf{Y}) + \delta \tag{18}$$

The $Pr()$ refers to the probability and ϵ is the privacy level of f . This definition restrains an adversary’s ability to infer whether the training or input dataset is D or D' .

According to Dework et al. [29], the exponential mechanism is a base approach to obtain $\epsilon - DP$. The main concern of the exponential mechanism is to return the result sampled from a definite distribution with a fixed output set.

Theorem 2 (Exponential Mechanism [29]) *Given a range R , a dataset D , a function u , and a privacy parameter ϵ , the mechanism $\mathcal{M}_E(x, u, \mathbf{R}) : D \rightarrow R$ satisfies $\epsilon - DP$ if $\mathcal{M}_E(x, u, \mathbf{R})$ output an element $r \in \mathbf{R}$ with probability Pr satisfies that:*

$$Pr \propto \exp\left(\frac{\epsilon}{2 \Delta u} u(x, r)\right) \quad (19)$$

where $u(x, r)$ is the utility function and Δu is sensitivity.

3.3.2 Limitations of the existing methods

The direct way to achieve DP in topic modeling is to add noise to the intermediate statistics [30, 33]. For example, Fould et al. [33] achieve DP in Gibbs sampling by adding Laplace noise to the sufficient statistics at the beginning of the Gibbs Sampling process. Zhu et al. [30] try to obtain DP by adding Laplace noise to sufficient statistics in the last iteration. Zhao et al. [3] point out that those methods cannot protect the training process against strong adversaries with full knowledge of the training and access to intermediate statistics in CGS-based topic model such as the LDA model due to two reasons: insufficient protection on word-counts and no protection on the sampled topics. Zhao et al. [3] address those issues in the CGS-LDA model, but this model is inadequate for comparative text analyses. Indeed, privacy protection of cross-collection topic models has not been previously described.

3.3.3 Model assumptions

Here we use the same assumptions as Zhao et al. [3] for adversary models and neighboring datasets.

- *Adversary Model:* We assume the data curator is trustworthy, but the adversary can observe the sampled topic assignments and the word count in each iteration during the training process.
- *Neighboring Datasets:* We construct the neighboring dataset D' by using *word replacement* [3]. Then, we assume that we can prevent the adversary from detecting the impact of *word replacement* on the training process.

3.3.4 Inherent privacy of CGS inference scheme

We will comprehensively analyze the inherent privacy of the CGS-based topic model training algorithm. Because Gibbs sampling has the same process with an exponential mechanism for differential privacy, Foulds et al. [33] highlighted that the Gibbs sampling method inherently generates some intrinsic differential privacy. The CGS

technique has the same property since it is one of the versions of Gibbs sampling. Furthermore, during each iteration of learning a topic-word distribution, the CGS inference outputs a topic from the topic set. Thus, Zhao et al. [3] began to investigate the CGS process in terms of the exponential mechanism, and they successfully concluded the inherent privacy of the CGS algorithm in the LDA model. They indeed specifically analyze the intrinsic privacy loss in each iteration before composing the privacy in total interactions of the CGS training scheme of LDA. We will employ the same concepts and then extend this idea to our proposed model so that we will use the same propositions in the HPP-ccLBLA model.

According to Zhao et al. [3], the intrinsic privacy of LDA's CGS inference technique has two significant drawbacks:

- Because privacy loss grows linearly, the privacy loss will accumulate rapidly.
- During the CGS inference process, there is no protection for word-count information since intrinsic privacy cannot secure the word-count data, leading to a privacy leakage issue.

We will address these two potential difficulties of inherent privacy after leveraging CGS's inherent privacy feature and present a privacy-preserving solution for our new model (HPP-ccLBLA).

3.3.5 Hybrid privacy-preserving ccLBLA algorithm

We first summarized the limitations of existing works in protecting sampled topics in the cross-collection topic model. Then, we studied the inherent privacy lack of protection on the word-count information. We propose a hybrid privacy protection algorithm for the ccLBLA model (HPP-ccLBLA) to address these issues.

The HPP-ccLBLA model described in this section integrates the inherent privacy of the CGS inference approach with external privacy provided by noise injection. We provide suitable noise in each iteration of the CGS technique to secure the word-count statistical information to overcome the possible privacy concern of intrinsic privacy.

We introduce the noise to obfuscate the difference between N_{dk} or N_{cdk} in each iteration. Besides, we minimize the rapid accumulation of privacy loss by setting the upper bound of the topic-word count. We choose the same method for HDP-LDA [3], which resorts to a clipping method to restrict the inherent privacy in each iteration. Specifically, the clipping only impacts a copy of N_{dk} or N_{cdk} in the computation of sampling but not the updating of

CGS inference. Algorithm 2 meets $(\epsilon_L + \epsilon_I) - DP$ in each iteration. ϵ_I is the inherent privacy loss:

$$\epsilon_I = \begin{cases} 2 \log \left(\frac{C}{\lambda_v} + 1 \right), & \text{if } x_{ij} = 0 \\ 2 \log \left(\frac{C}{\lambda_{cw}} + 1 \right), & \text{if } x_{ij} = 1 \end{cases} \quad (20)$$

The ϵ_L denotes the privacy loss incurred by the Laplace noise, and the CC is the clipping bound for N_{dk} or N_{cdk} .

```

Input : Max iteration  $T$ , Prior parameters  $\zeta_c, \tau_c, \epsilon,$ 
        Topic number  $K$ , Corpus  $D$ ,  $\mathbf{w}$ , Class  $c$ 
Output : Parameters  $\theta_j, \phi_k, \sigma_{ck}$  using (15) – (17)
        Privacy loss  $\epsilon = (\epsilon_L + \epsilon_I)$  using (20)
//Initialization
Initialize  $\mathbf{z}, \mathbf{x}, N_{jk}, N_{kw}, N_{ckw}, N_x$ 
//Collapsed Gibbs Sampling
while iter <  $T$  do
  for word  $i \in$  document  $D$  and document  $j \in$  class  $c$ 
  do
     $\eta \sim \text{Lap} \left( \frac{2}{\epsilon_L} \right)$  // Laplace noise
    if  $x_{ij} = 0$  then
       $N_{kw} = N_{kw} + \eta$  // Add noise to each  $N_{kw}$ 
      Clip:  $(N_{kw})^{temp} = \min(N_{kw}, C)$  // Restrict
the inherent privacy
      Compute:  $\epsilon_I = 2 \log \left( \frac{C}{\lambda_v} + 1 \right)$  //Compute
inherent privacy loss
      update  $z_{ij}$  using (11)
    else
       $N_{ckw} = N_{ckw} + \eta$  // Add noise to each  $N_{ckw}$ 
      Clip:  $(N_{ckw})^{temp} = \min(N_{ckw}, C)$  // Restrict
the inherent privacy
      Compute:  $\epsilon_I = 2 \log \left( \frac{C}{\lambda_{cv}} + 1 \right)$  //Compute
inherent privacy loss
      update  $z_{ij}$  using (13)
    end if
    update  $x_{ij}$  using (12) and (14)
    update  $N_{jk}, N_{kw}, N_{ckw}, N_x$ 
  end for
end while

```

Algorithm 2 Summary of HPP-ccLBLA algorithm.

In Algorithm 2, the privacy loss in the HPP-ccLBLA model includes privacy loss ϵ_L incurred by Laplace noise and the inherent privacy loss ϵ_I of CGS inference. According to (20), we can conclude that the rapid increase of inherent privacy loss has been limited, and the word-count statistical information also gets privacy protection.

4 Experimental results

The cross-collection topic model was evaluated via perplexity, classification accuracy, and topic coherence using several applications such as comparative text mining and image classification. We also compare topic examples across multiple text datasets to demonstrate the strengths of our technique. The experiments utilize four text datasets with different collection numbers, document lengths, domains, and one well-known image dataset. In this section, we use the Scale Invariant Feature Transform (SIFT), and K-means approaches to successfully apply our cross-collection topic model (ccLBLA) to an image classification assignment using the Bag of Visual Words (BOVW) approach. Finally, we validate the HPP-ccLBLA algorithm’s performance in model utility, such as perplexity, to show our approach’s merits.

4.1 The datasets

Table 3 displays an overview of each dataset size for the text datasets. The COVID-19 newspapers dataset contains online newspapers from the United States of America, which is collected from COVID-NEWS-US-NNKDATASET¹. Whereas, the second collection of this dataset is from several different British newspaper websites.² Indeed, we can use this novel dataset for comparative text mining tasks in aggregation and summarization to extract common and different effects and knowledge about the virus in two countries and demonstrate our proposed model’s merits. Besides, the second text dataset mainly focuses on computer science academic papers, including the abstracts of NeurIPS³ and CVPR⁴ papers published in 2019. We apply our model to comparative text analysis to automatically spot different topics and trends in these two conferences. The third text dataset consists of a subset of the New York Times (NYT) comments,⁵ which contains more than two million comments from 2017 to 2018. We decided to use all comments posted on NYT articles in the period Jan - April 2017 and Jan - April 2018 to compare the performance of the ccLBLA model with ccLDA [17], and LDA [31] models. The dataset of NYT comments forms the largest dataset in our evaluation. We also reuse the dataset⁶ reported in ccLDA [17] so that we can make

¹<https://github.com/nnk−dataset/usa−nnk>

²<https://www.kaggle.com/jwallib/coronavirus−newspaper−classification/data>

³<https://www.kaggle.com/rowhitwami/nips−papers−1987−2019−updated>

⁴<https://www.kaggle.com/paultimothymooney/cvpr−2019−papers>

⁵<https://www.kaggle.com/aashita/nyt−comments>

⁶<http://www.michaeljpaul.com/downloads/ccdata.php>

Table 3 Datasets - number of documents D and average number of words per document W/D (without stop words)

Text Datasets			
Dataset	Collection	D	W/D
COVID-19 Newspapers	USA UK	2731	433
Academic Papers	NIPS CVPR	2787	91
Traveler Forum	India Singapore UK	4174	247
NYT Comments	2017	44465	214
	2018	48903	237

a fair comparison. The last text dataset crawled from an online travel platform, including three different countries' discussion forums of India, Singapore, and the UK, with thousands of threads in each collection [17]. Therefore, our experiment utilized four domains of datasets: newspapers, academic papers, customer comments, and travel blogs, to prove that our approach can handle different types of documents.

For the image-based application, we used the famous grayscale natural scenes dataset [55]. As shown in Table 4 and Fig. 4, this image dataset includes the following categories: kitchen, office, bedroom, suburb, highway, living room, street, downtown, industry, store, forest, skyscraper, coast, mountain, and rural area.

4.2 Experiments for text mining

For comparative text mining application, we preprocess the text datasets by first tokenizing words with the Natural

Table 4 Size of each image category

Natural scenes images dataset	
Categories	Size
Kitchen	210
Office	215
Bedroom	216
Suburb	241
Highway	260
Living Room	289
Street	292
Downtown	308
Industry	311
Store	315
Forest	328
Skyscraper	356
Coast	360
Mountain	374
Rural Area	410

Language ToolKit (NLTK) [56], removing punctuation, stop-words and then lemmatizing tokens to derive their common base form. We choose BL priors hyperparameters following the same setting of the asymmetric BL priors in [10]. For Dirichlet-based model, the topic distribution priors are fixed and $\alpha = 0.1$. Then, we set β and δ to 0.01; for γ_0 and γ_1 , we use the same value, 1.0. The LDA and ccLDA (LDA and ccLDA)⁷ are based on a widely used open-source package GibbsLDA++. For the text experiment validation, we use ten-fold cross-validation, which separates each dataset with a 90% training set and 10% test set. In the Gibbs sampling, the burn-in period is five hundred, and then we collect ten samples separated by lags of ten iterations. The average of ten samples is the final result of the model. After, we calculate the document-topic parameter θ , the collection-independent word distribution parameter ϕ , the collection-specific word distribution parameters σ , and ψ . Moreover, we assessed model perplexity, document classification accuracy, and mixed topic coherence based on these parameters and results.

4.2.1 Perplexity

Perplexity evaluates how well a trained topic model predicts the co-occurrence of words on the unseen test data. Perplexity focuses on the topic model's ability to generate word probabilities for the unseen dataset, so a lower perplexity score indicates better generalization performance. Based on Hofmann [37], we use the "fold-in" approach for this experiment. This method evaluates the model by only learning the test dataset's document-topic probabilities θ . All other topic model probabilities parameters are kept the same from the training dataset—the validation Gibbs sampling measure only the document-topic distributions on the test documents.

In the cross-collection topic model, for a test dataset of M documents, the perplexity is:

$$Perplexity(D_{test}) = 2^{-\frac{1}{M} \sum_w \text{likelihood}(w|\theta_{d_{new},c})} \quad (21)$$

In this formula, after getting the topic probabilities θ_d and the collection c of a test document d , the likelihood of a word w in test document d is:

$$\begin{aligned} \text{likelihood}(w | \theta_{d_{new}, c}) &= \sum_z P(z | \theta_{d_{new}}) \\ &\times [P(w | z, x = 0)P(x = 0) \\ &+ P(w | z, c, x = 1)P(x = 1)] \end{aligned} \quad (22)$$

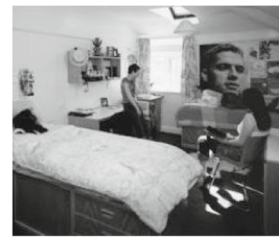
⁷<http://www.michaeljpaul.com/downloads/mftm.php>



(a) Kitchen



(b) Office



(c) Bedroom



(d) Suburb



(e) Highway



(f) Living Room



(g) Street



(h) Downtown



(i) Industry



(j) Store



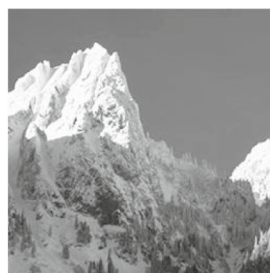
(k) Forest



(l) skyscraper



(m) Coast



(n) Mountain



(o) Rural Area

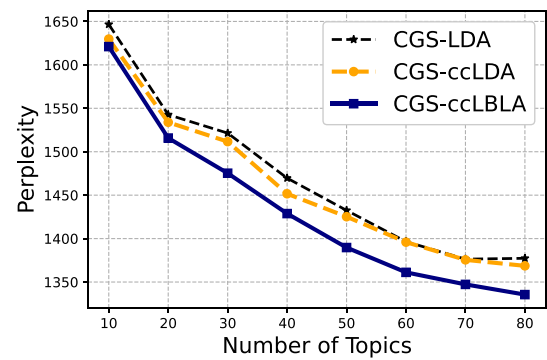
Fig. 4 Examples from the natural scenes images dataset (Total Fifteen Categories)

$P(x = 0)$ is the probability that word w is collection-independent, and $x = 1$ means the likelihood of word w being collection-specific. $P(w | z, x)$ denotes the possibility of word w sampled from collection-common or collection-specific when topic z is sampled.

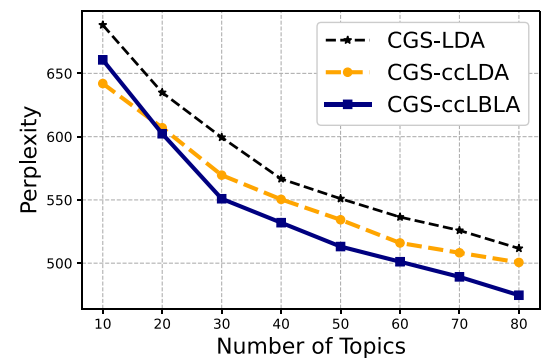
The perplexity for each model on both corpora for different values of topics is shown in Fig. 5. As expected, cross-collection topic models (ccLDA and ccLDA) generally achieve a lower perplexity than single-collection topic models such as the LDA model because these models utilize extra information to assign a greater probability to words more likely to exist in a document. According to Fig. 5, The ccLBLA and ccLDA models have comparable performance when the number of topics is negligible since the topic number is not ideal for specific datasets. The ccLBLA models achieve lower perplexity than the ccLDA models as the number of topics increases. The ccLBLA and ccLDA models produce similar results on the traveler forum dataset, although the difference between the two models is not significant. After examining the traveler forum dataset, we notice that each collection contains many duplicate documents, implying that this dataset cannot accurately demonstrate the capabilities of a cross-collection topic model to predict unseen documents. In the other three text datasets, ccLBLA has a lower perplexity than the ccLDA model. This result also demonstrates the flexibility of the BL prior (general covariance structure in (4)) compared to the Dirichlet distribution, which is very limited for its inability to perform in the case of positively correlated datasets. Therefore, we can conclude that the main reason for our proposed model's (ccLBLA) improvement is that the BL distribution prior has better topic correlation, flexibility, generalization, and modeling capabilities [10, 26].

4.2.2 Document classification

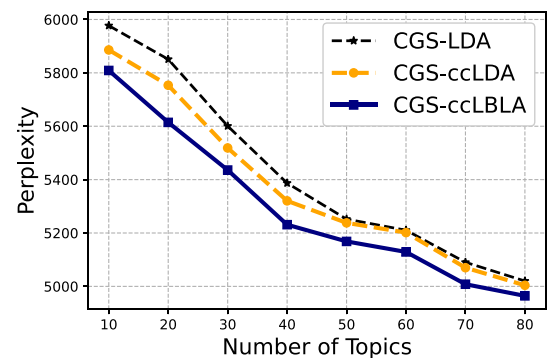
Cross-collection topic models like ccLBLA and ccLDA can produce collection predictions for unseen documents since they can generate a document likelihood that relies on the document's collection [17]. Each model predicts the collection of test documents based on the words in this task. Furthermore, the document classification accuracy may be used to assess the model's separation of collection-common and collection-specific words [17, 18]. The cross-collection topic model provides a probability for each collection and assigns the most likely collection for the test document. This probabilistic classification enables a more precise assessment of each topic model's degree of certainty. Therefore, we can objectively measure the performance of these models in document classification. The cross-collection topic model calculates the category of an unlabeled document d for choosing collection



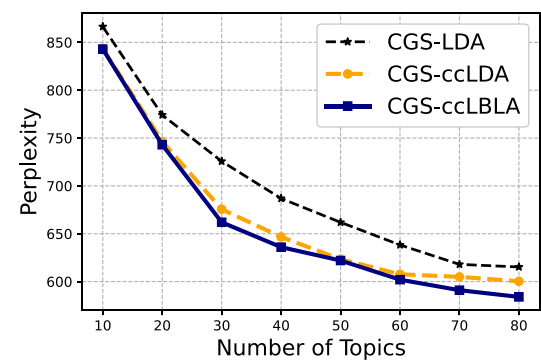
(a) COVID-19 Newspapers



(b) Academic Papers



(c) NYT Comments



(d) Traveler Forum

Fig. 5 Perplexity results on four different datasets for LDA, ccLDA and ccLBLA

Table 5 Document classification accuracy results on four different datasets for ccLDA and ccLBLA

Document Classification Accuracy		
Dataset	ccLDA	ccLBLA
COVID-19 Newspapers	0.40	0.59
Academic Papers	0.76	0.91
NYT Comments	0.67	0.81
Traveler Forum	0.45	0.63

c as:

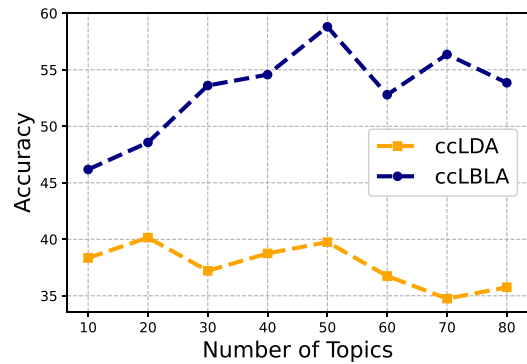
$$\begin{aligned}
 \text{label} = \arg \max_c P(c) \prod_w \sum_z P(z | \theta_{d_{new}}, c) \\
 \times [P(w | z, x = 0)P(x = 0) \\
 + P(w | z, c, x = 1)P(x = 1)] \quad (23)
 \end{aligned}$$

We can get the predicted collection c by using (23). Expect for $P(z | \theta_d, c)$ and $P(c)$; other probabilities are generated from the training document because $P(z | \theta_d, c)$ and $P(c)$ depend on the new test document. Following Paul’s approach [17], we assign a collection c for the unlabeled document, and then we use another Gibbs sampling procedure to learn these probabilities. The classification accuracy for the new test datasets is $\frac{D_{correct}}{D_{testset}}$.

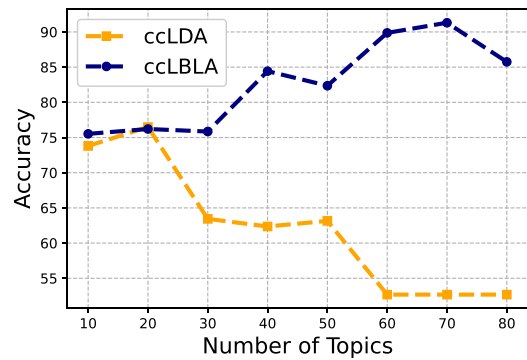
Table 5 and Fig. 6 demonstrates all document classification accuracy results for four different datasets among ccLDA and ccLBLA models. As shown in Table 5, the performance of the ccLBLA model is much better than the ccLDA model in the document classification task on the whole. On the COVID-19 newspapers dataset, the document classification accuracy of the ccLBLA model is almost 45% higher than the ccLDA model. Also, the ccLBLA model achieves about 40% greater than ccLDA’s accuracy. The ccLBLA model gets about 20% higher accuracy in academic papers and NYT comments datasets than the other two datasets. We can find that the ccLBLA model’s accuracy does not drop like the ccLDA model with increasing the number of topics from Fig. 6. Based on those results, compared with the ccLDA model, we can conclude that the ccLBLA model obtains a better ability to separate collection-common and collection-specific words by introducing BL distribution.

4.2.3 Topic coherence

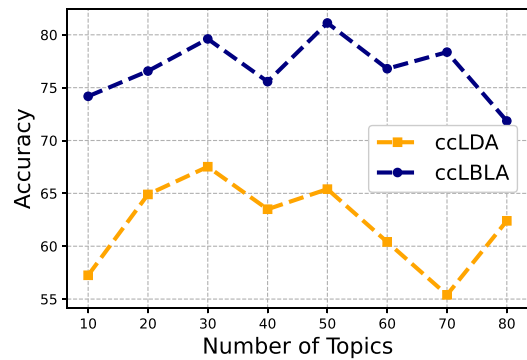
The topic coherence evaluation compares the ccLBLA and ccLDA models for clustering words inside the collection-independent topic and between multiple collection-specific topics through semantic similarity. In particular, the model’s capacity to align topics from distinct collections among different collection-specific topic-word distributions was



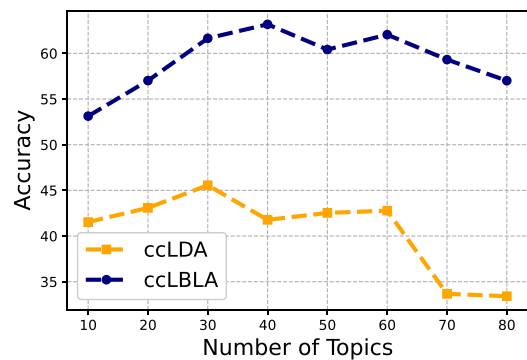
(a) COVID-19 Newspapers



(b) Academic Papers



(c) NYT Comments



(d) Traveler Forum

Fig. 6 Document classification results on four different datasets for ccLDA and ccLBLA

tested. On the other hand, the current topic coherence metric only examines a single word distribution per topic, not several word distributions inside a single topic. As a result, we use the mix topic coherence [18], which mixes the topic representation of the collection-independent word distribution with the collection-specific word distribution. As a result, we employ the union of these representations as a unified topic representation, which is distributed by particular topic terms and is independent of the individual collections. This union's coherence can be evaluated to determine the current topic coherence score.

This mixed topic coherence may also be used to evaluate the topical alignment of different collection word distributions according to Risch, and Krestel [18]. The C_V technique [57] is chosen as the topic coherence evaluation method. This coherence measurement is based on a sliding window, segmentation of a set of top words, indirect confirmation measures using normalized pointwise mutual information (NPMI), and cosine similarity. This coherence metric retrieves the co-occurrence count for a given word using a sliding window and a constant window size. The NPMI is calculated using these counts. When a collection of top-level words is segmented, the cosine similarity between each top word vector and the sum of all complete word vectors is calculated. The arithmetic mean of these similarities is thus C_V Coherence. Even though C_V coherence measurement considers human judgments, this topic coherence has limits since C_V coherence implies that words that never appear together in the reference dataset are inconsistent. This assumption is not suitable for some datasets with strong language contrast.

In this experiment, we use the Palmetto library⁸ to evaluate the topic coherence automatically. Table 6 shows the C_V -based topic coherence of four datasets, which averages all topics' coherence scores. The number of topics in the mixed topic coherence evaluation is based on the result from perplexity and document classification. From Table 6, we can conclude that the ccLBLA model obtains slightly higher topic coherence values than the ccLDA model. Especially for the academic papers dataset, our proposed model gets around 8.3% improvement. Indeed, the ccLBLA model obtains almost 4.5% advancement compared with the ccLDA model in the COVID-19 newspapers and travel forum dataset.

4.2.4 Topics analysis and discussion

We modeled this dataset with 30 topics based on perplexity and topic coherence findings in the COVID-19 newspapers datasets. The top 10 words for collection-independent and each collection local word distribution from the ccLBLA

Table 6 Topic coherence comparison with ccLDA and ccLBLA models

Topic Coherence		
Dataset	ccLDA	ccLBLA
COVID-19 Newspapers	0.3832	0.4008
Academic Papers	0.3886	0.4211
NYT Comments	0.4173	0.4291
Traveler Forum	0.3833	0.4013

model are shown in Table 7. Topic 15, which is about maintaining public health during the Covid-19 pandemic, may be deduced from the collection-independent topic terms. Indeed, when comparing the methods used in the United Kingdom and the United States, it is evident that the United States government advises individuals to work from home and stay at a safe distance from public places to prevent the spread of Covid-19 in the USA collection. The UK government recommends that people wear masks and wash their hands to protect themselves.

Moreover, Topic 19 presents the symptom of COVID-19. Topic 23 is a Coronavirus study report. The newspapers in the United States and the United Kingdom have distinct concerns. The US newspaper emphasized the virus's instances and patients in China. In contrast, the COVID-19 virus's data across the world and vaccine manufacture were the focus of the UK media.

Furthermore, Table 8 compares ccLDA and ccLBLA models to world economic issues from the New York Times Comments dataset. Our method, the ccLBLA model, also results in superior separation of collection-specific terms and theme coherence in this dataset. The 2017 collection is assigned the terms "bank" and "estate" by the ccLDA model, whereas the world economy themes are assigned the words "job", "work", and "worker" by the ccLBLA model. Labor costs have a considerably more significant impact on the global economy than real estate and banks because real estate and banks can affect the local economy. Moreover, both models provide the same outcome in the 2018 collection regarding China's impact on global commerce. The ccLDA model, on the other hand, is limited to the Sino-Canadian economic connection. ccLBLA, on the other hand, assigns "China" and "global" to 2018 collections, which is more relevant to the collection's specific topic: the global economy.

4.3 Image classification

This section successfully applies the cross-collection topic model in an image classification following the bag of visual words framework [8, 10]. Figure 7 illustrates an overview of the feature extraction, clustering, and ccLBLA pipeline.

⁸<https://github.com/dice-group/Palmetto>

Table 7 ccLBLA model with three topics for COVID-19 newspapers dataset

Topic 15		Topic 19		Topic 23	
Coronaviru, health, work, week, continue, virue, time emerg, country, clear		Symptom, infect, viru day, ill, coronavirus, sever people, cough, fever		viru, disease, conronavirue animal, vaccine, spread, human research, study, scientist	
UK Collection	USA Collection	UK Collection	USA Collection	UK Collection	USA Collection
Mask	peopl	health	hand	vaccin	infect
Worker	govern	case	breath	world	China
Suppli	stay	peopl	test	data	patient
Protect	home	covid19	cough	use	Wuhan
Wear	social	test	covid19	develop	outbreak
Face	test	viru	lung	medium	ill
Product	distanc	infect	suffer	research	test
Hand	offic	diseas	bodi	work	pandem
Equip	public	spread	throat	inform	cent
Hospit	rule	death	clean	report	public

Specifically, we use the Scale Invariant Feature Transform (SIFT) algorithm to extract the local features from local patches through the whole corpus collection, the vectors of counts in each image. The K-means algorithm clusters the set of training image descriptors to find the unique local feature representation. After that, we can obtain the codeword from the cluster center and the codebook or the dictionary of image vocabulary. The codebook contains a vector of counts for each image. Using this bag of visual words approach, we can consider each image as a document and train them into our proposed ccLBLA model. Besides, in this well-known grayscale fifteen-categories

natural scenes dataset, the data is separated into training and testing sets in each category: the testing set has a hundred random images while the remaining constitute the training set. In the model section, we set the range of topic numbers from 10 to 80. Then, we can use the bags of visual word representation for each image to evaluate the performance of the ccLBLA model in the image classification task based on Eq. 23.

Table 8 Example of topics from the NYT Comments dataset as discovered by the ccLDA and ccLBLA models

ccLDA		ccLBLA	
busi, market, product, money, trade, compani, economi, econom, good, price		econom, economi, job, polici, worker, increas, corpor, employ, product, cost	
2017 Collection	2018 Collection	2017 Collection	2018 Collection
Regul	trade	job	trade
Bank	china	work	china
Estat	tariff	worker	market
Reduc	steel	labor	global
Econom	manufactur	class	industri
Growth	chine	busi	good
2008	aluminum	incom	compani
Doddfrank	canada	rate	rate
Mortaga	industri	rich	cost
Banker	impos	growth	product

$$\begin{aligned}
 class = arg \max_c & \prod_w \sum_z P(z | \theta_{d_{new}}, c) \\
 & \times [P(w | z, x = 0)P(x = 0) \\
 & + P(w | z, c, x = 1)P(x = 1)]
 \end{aligned}
 \tag{24}$$

Because the cross-collection topic model can generate an image (document) likelihood which depends on the image’s collection [17], cross-collection models like ccLBLA and ccLDA are capable of making collection predictions for unseen documents. Therefore, the cross-collection topic model naturally suits the classification task, and each model can predict the collection of test documents based on the visual words. Specifically, The predictive model is created by estimating the topic parameters using (15). The predictive topic distributions and the empirical likelihood framework lead to the estimation of the class likelihood. Based on (24), we can obtain the class conditionals to predict the class label of unseen images. Therefore, the collection of the unseen image is chosen with the highest class posterior distribution.

Our experiment uses the same training and testing dataset to implement the LDA, LBLA, ccLDA, and ccLBLA models by estimating the class likelihood to predict the

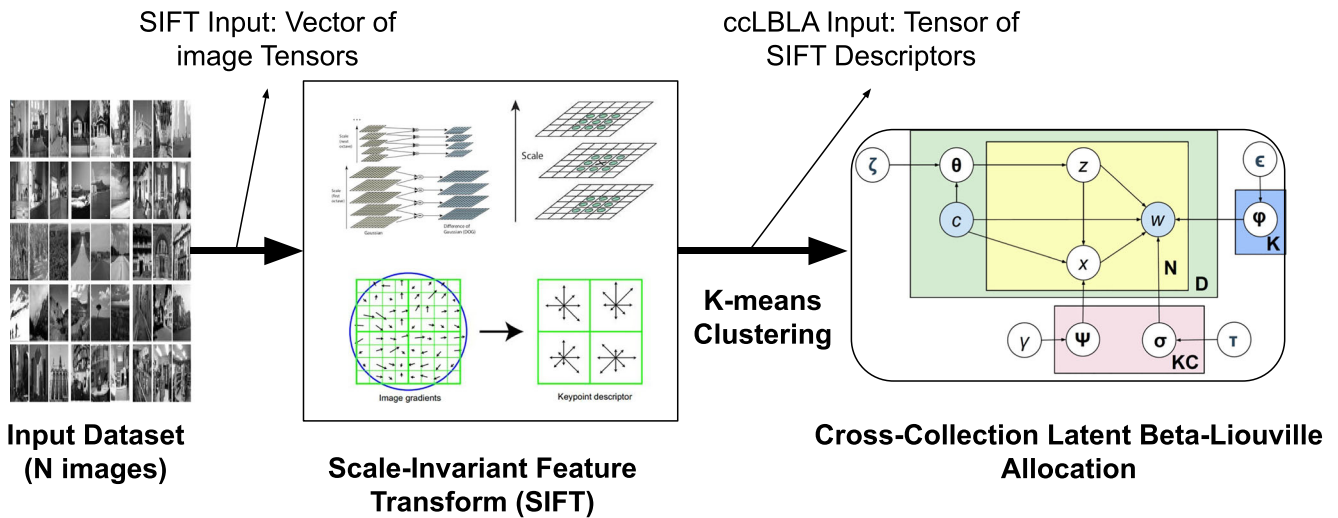


Fig. 7 An overview of the feature extraction, clustering, and ccLBLA pipeline

class label of unseen images. The highest class posterior distribution will assign the class for the unseen image.

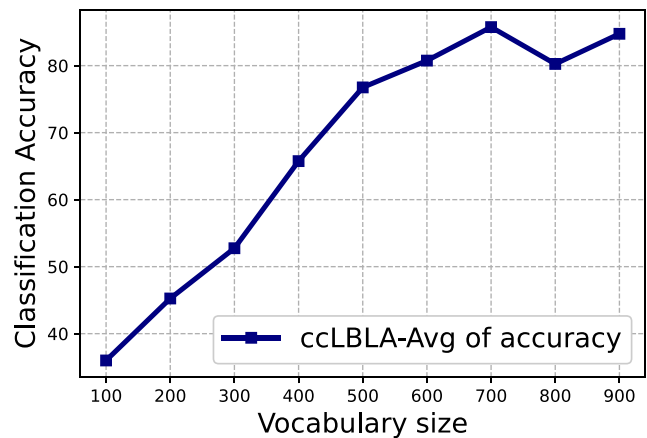
From Table 9, we can conclude that the ccLBLA model provides better accuracy than the other topic models. Precisely, our proposed model achieves 57% (CGS-LDA), 25% (CGS-LBLA), and 12% (CGS-ccLDA) higher accuracy. Figure 8(a) and (b) show that the optimal vocabulary size is $V=700$, and we find that the optimal number of topics is $K=50$ in model selection. The high average accuracy is 85.75 and high accuracy rate is 90.97, shown in the confusion matrix (Fig. 9), which outperforms its competitors (see Table 9). These results demonstrate that the generative schemes with more flexible priors (BL distribution) can enhance the cross-collection topic model’s performance and reinforce the ccLDA model’s generalization by overcoming the negative covariance structure of the Dirichlet distribution.

4.4 Performance of HPP-ccLBLA

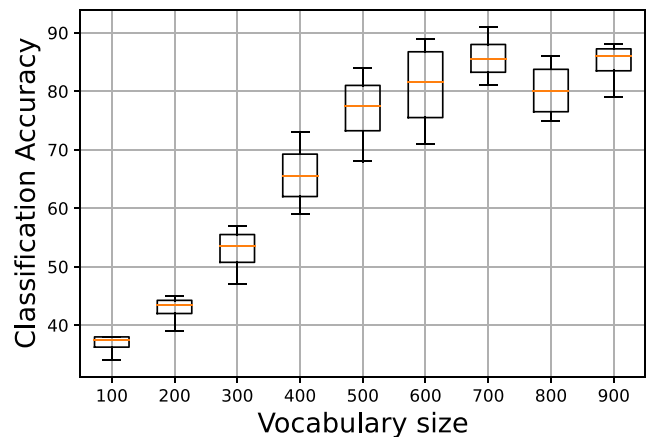
This section details our assessment of the HPP-ccLBLA model, emphasizing its utility. We implement our method on three real-world text datasets: Covid19 newspapers, academic publications, and comments from the New York Times. The statistics for these datasets are presented in Table 6. Because the traveler forum dataset contains many duplicate documents which cannot accurately demonstrate

Table 9 The accuracies of different tested models applied to the natural scene dataset

LDA	LBLA	ccLDA	ccLBLA
57.93%	72.67%	81.37%	90.97%



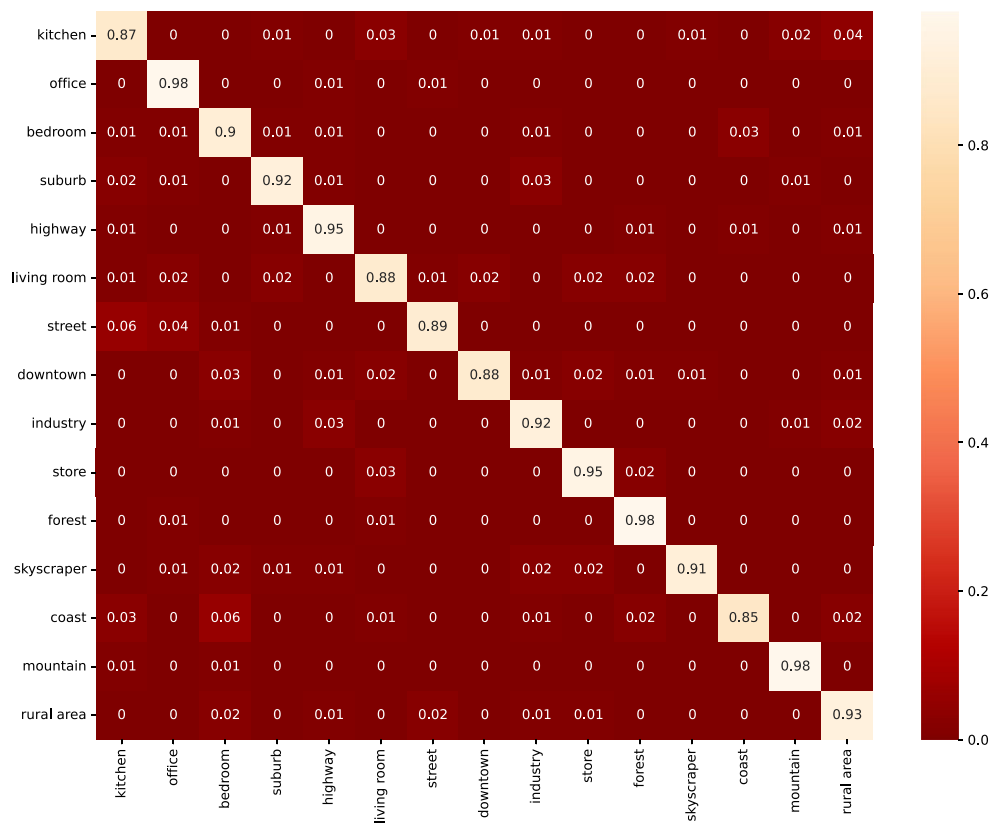
(a) Average accuracy



(b) Accuracy

Fig. 8 The accuracy as a function of the vocabulary size for image classification

Fig. 9 Confusion matrix for the natural scenes classification



the capabilities of a cross-collection topic model, we will not use this dataset in this experiment.

In this experiment, we select perplexity as the evaluation standard for topic model utility, similar to Zhao et al. [3], because perplexity emphasizes the generative aspect of topic models to predict word probabilities for unseen documents in the test dataset [4, 18]. A lower perplexity indicates a higher likelihood and better model utility. To compute the perplexity and likelihood of a cross-collection topic model, we apply (21) and (22). To evaluate our strategy, we will compare it to CDP-ccLBLE+, which protects the training process by introducing Laplace noise into N_{dk} , N_{kw} , and N_{ckw} in each iteration. In addition, we will compare the differences in topic samples between HPP-ccLBLE and Non-privacy protection ccLBLE to validate the utility of our approach.

4.4.1 Utility

The perplexity of HPP-ccLBLE and CDP-ccLBLE+ with different Laplace privacy ϵ settings is shown in Fig. 10. In Fig. 10, we also compare the plain CGS algorithm (Non-Privacy), which lacks privacy protection. Furthermore, we employ several BL parameter configurations in ccLBLE experiments in this utility experiment. To limit the inherent privacy, we explicitly set a larger λ_w , and λ_{cw} , as well as a proper clipping bound C , during the training process.

Then, we set the intrinsic privacy level of HPP-ccLBLE to 10 in each iteration. Because we utilize a more significant parameter in BL distribution, the prior information can improve the model utility ability to the noise. The limited Inherent means that the HPP-ccLBLE has the same setting for inherent privacy level but no Laplace noise for N_{kw} and N_{ckw} . From Fig. 10, we can infer that Limited Inherent has a utility degradation compared with the plain CGS algorithm (Non-Privacy) because Limited Inherent integrates a stronger inherent privacy guarantee. Even though CDP-ccLBLE+ introduces more Laplace noise and privacy loss than the HPP-ccLBLE scheme, including the intrinsic privacy loss, the utility of HPP-ccLBLE outperforms the CDP-ccLBLE+ method in that three real-world datasets based on the BL prior information.

We tested our ccLBLE model and compared it with LDA and ccLDA models using four evaluation methods in text application and image classification. The model in all the evaluation methods shows similar or improved results compared to LDA and ccLDA models. Compared with these models, the ccLBLE model not only replaces the Dirichlet prior for the document parameter but also does it for the corpus parameter. Therefore, our model provides a stronger generalization than those Dirichlet-based topic models. Specifically, in dimension D , the Dirichlet has $D + 1$ parameters while the Beta-Liouville has $D + 2$. Thus, compared to the Dirichlet, Beta-Liouville has one

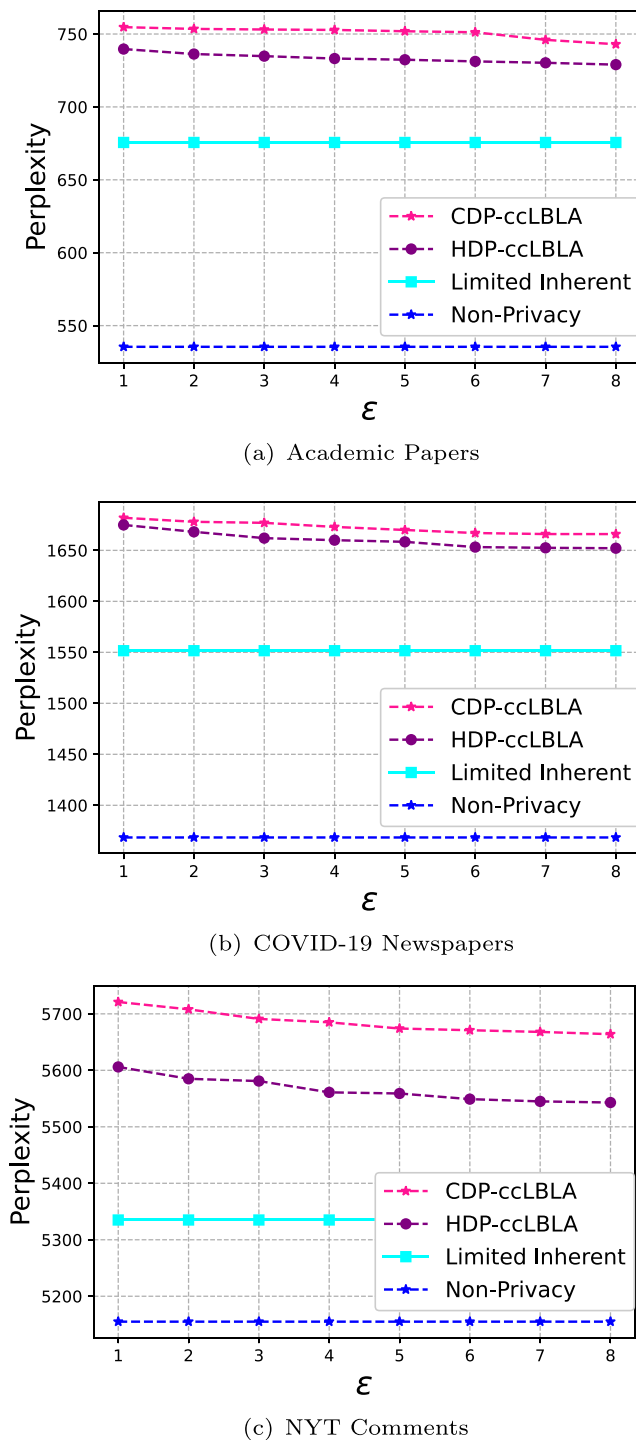


Fig. 10 Perplexity results on three different datasets vs. Privacy level of HPP-ccLBLA

extra parameter. Indeed, the general covariance structure in the BL is also suitable for any data modeling within the BoW framework. This is not the case for the Dirichlet for its limitation in the case of positively correlated datasets. Then, our proposed model (HPP-ccLBLA) naturally outperforms the HDP-LDA model in utility with a similar assumption

and privacy protection algorithm. Moreover, compared with the HDP-LDA model, our HPP-ccLBLA model can extract more useful information with a better topic correlation structure by modeling multiple document collections.

5 Conclusion

This paper first presents and implements a novel cross-collection topic model (ccLBLA model) that utilizes the BL distribution instead of Dirichlet for various domain text collections to improve previous cross-collection topic models because the BL distribution can provide a better topic correlation representation. Furthermore, we investigate the privacy protection of topic models with differential privacy and propose a centralized privacy-preserving algorithm for the ccLBLA model (HPP-ccLBLA), which takes advantage of the Collapsed Gibbs Sampling inference approach's inherent differential privacy guarantee to address the privacy issue.

The ccLBLA model extends the ccLDA and LBLA models. These previous models suffer from the limitation of Dirichlet prior or focusing only on one individual data collection. All of these issues are addressed by the ccLBLA model. In particular, our new model replaced the Dirichlet distribution with the BL prior in the generating process, making our model more flexible. We compare our experimental results to the ccLDA and LDA models to demonstrate the merit of our new technique. The perplexity of the topic model, document classification accuracy, topic coherence, and topic samples are all examined. Experimental findings show that our ccLBLA beats ccLDA and LDA models on all four quality metrics across four real-world text datasets with varying domains and numbers of collections. Moreover, we present the first study on applying the cross-collection topic model to image classification applications. Because of the general covariance structure in the BL distribution, the performance of the ccLBLA model in image classification demonstrates a higher classification accuracy than the ccLDA, LBLA, and LDA models. Extensive experiments reveal that our HPP-ccLBLA model can prevent data inference from intermediate statistics during training, and this algorithm can achieve a good model utility under differential privacy.

For our future work, we plan to improve the efficiency of the HPP-ccLBLA model so that it can be used for real-time streaming data such as investigating an online-based ccLBLA model under privacy protection. In addition, it is possible to extend our model from a centralized privacy-preserving algorithm to a local privacy algorithm. One can also naturally extend the proposed model by introducing a more flexible prior to improve the model's performance. For reproducibility and future improvement by other

researchers, the complete source code is provided in the following repository: <https://github.com/zuol149/ccLBLA>.

Acknowledgements The completion of this work was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Natural Science Foundation of China (62276106, 61876068). The authors would like to thank the associate editor and the reviewers for their helpful comments.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

- Hua T, Lu C, Choo J, Reddy CK (2020) Probabilistic topic modeling for comparative analysis of document collections. *ACM Trans Knowl Discov Data* 14(2):24–12427
- Le TMV, Akoglu L (2019) Contravis: contrastive and visual topic modeling for comparing document collections. In: Liu L, White RW, Mantrach A, Silvestri F, McAuley JJ, Baeza-Yates R, Zia L (eds) *The world wide web conference, WWW 2019*, 13–17 May 2019. ACM, pp 928–938
- Zhao F, Ren X, Yang S, Han Q, Zhao P, Yang X (2021) Latent dirichlet allocation model training with differential privacy. *IEEE Trans Inf Forensics Secur* 16:1290–1305
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Blei DM, Lafferty JD (2009) Topic models. In: *Text mining*. Chapman and hall/CRC, pp 101–124
- Blei DM (2012) Probabilistic topic models. *Commun ACM* 55(4):77–84
- Blei DM, Carin L, Dunson DB (2010) Probabilistic topic models. *IEEE Signal Process Mag* 27(6):55–65
- Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer society conference on computer vision and pattern recognition (CVPR 2005), 20–26 June 2005. IEEE computer society, pp 524–531
- Ihou KE, Bouguila N (2019) Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing* 332:372–395
- Ihou KE, Bouguila N (2020) Stochastic topic models for large scale and nonstationary data. *Eng Appl Artif Intell*, vol 88
- Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Inf Sci* 275:1–12
- Fan W, Yang L, Bouguila N (2021) Unsupervised grouped axial data modeling via hierarchical bayesian nonparametric models with watson distributions. *IEEE Trans Pattern Anal Mach Intell*: 1–1
- Yuan M, Durme BV, Ying JL (2018) Multilingual anchoring: interactive topic modeling and alignment across languages. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) *Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, neurIPS 2018*, 3–8 Dec 2018. Montréal, Canada, pp 8667–8677
- Setty V, Anand A, Mishra A, Anand A (2017) Modeling event importance for ranking daily news events. In: De Rijke M, Shokouhi M, Tomkins A, Zhang M (eds) *Proceedings of the tenth ACM international conference on web search and data mining, WSDM 2017*. ACM, 6–10 Feb 2017, pp 231–240
- Rudrapal D, Das A, Bhattacharya B (2018) A survey on automatic twitter event summarization. *J Inf Process Syst* 14(1):79–100
- Zhai C, Velivelli A, Yu B (2004) A cross-collection mixture model for comparative text mining. In: Kim W, Kohavi R, Gehrke J, DuMouchel W (eds) *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, Seattle, Washington, USA, 22–25 Aug 2004, pp 743–748
- Paul MJ, Girju R (2009) Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: *Proceedings of the 2009 conference on empirical methods in natural language processing, EMNLP 2009*, 6–7 Aug 2009, Singapore, a meeting of SIGDAT, a special interest group of the ACL. ACL, pp 1408–1417
- Risch J, Krestel R (2018) My approach = your apparatus? In: Chen J, Gonçalves MA, Allen JM, Fox EA, Kan M, Petras V (eds) *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*. ACM, JCDL 2018, fort worth, TX, USA, 03–07 June 2018, pp 283–292
- Bouguila N (2008) Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Trans Knowl Data Eng* 20(4):462–474
- Bouguila N, Ziou D, Hammoud RI (2009) On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling. *Pattern Anal Appl* 12(2):151–166
- Bouguila N (2011) Count data modeling and classification using finite mixtures of distributions. *IEEE Trans Neural Netw* 22(2):186–198
- Fan W, Bouguila N (2013) Variational learning of a dirichlet process of generalized dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognit* 46(10):2754–2769
- Ihou KE, Bouguila N (2017) A new latent generalized dirichlet allocation model for image classification. In: *Seventh international conference on image processing theory, tools and applications, IPTA 2017*. IEEE, 28 Nov – 1 Dec 2017, pp 1–6
- Bakhtiari AS, Bouguila N (2014) Online learning for two novel latent topic models. In: Linawati, Mahendra MS, Neuhold EJ, Tjoa AM, You I (eds) *Information and communication technology - second IFIP TC5/8 international conference, ICT-eurasia 2014*, Bali, Indonesia, 14–17 Apr 2014. *Proceedings. Lecture notes in computer science*. Springer, vol 8407, pp 286–295
- Ihou KE, Bouguila N (2018) A smoothed latent generalized dirichlet allocation model in the collapsed space. In: *IEEE 61st international midwest symposium on circuits and systems, MWSCAS 2018*, windsor. IEEE, ON, Canada, 5–8 Aug 2018, pp 877–880
- Bakhtiari AS, Bouguila N (2016) A latent beta-liouville allocation model. *Expert Syst Appl* 45:260–272
- Fredrikson M, Lantz E, Jha S, Lin SM, Page D, Ristenpart T (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Fu K, Jung J (eds) *Proceedings of the 23rd USENIX security symposium*. USENIX association, San Diego, CA, USA, 20–22 Aug 2014, pp 17–32
- Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: *2017 IEEE symposium on security and privacy, SP 2017*. IEEE computer society, San Jose, CA, USA, 22–26 May 2017, pp 3–18
- Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006) Our data, ourselves: privacy via distributed noise generation. In: Vaudenay S (ed) *Advances in cryptology - EUROCRYPT 2006*, 25th annual international conference on the theory and applications of cryptographic techniques, st. petersburg, Russia, 28 May – 1 June 2006, *Proceedings. Lecture notes in computer science*. Springer, vol 4004, pp 486–503

30. Zhu T, Li G, Zhou W, Xiong P, Yuan C (2016) Privacy-preserving topic model for tagging recommender systems. *Knowl Inf Syst* 46(1):33–58
31. Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235
32. Wang Y, Fienberg SE, Smola AJ (2015) Privacy for free: posterior sampling and stochastic gradient monte carlo. In: Bach FR, Blei DM (eds) *Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, 6–11 July 2015*. *JMLR Workshop and Conference Proceedings*. JMLR.org, vol 37, pp 2493–2502
33. Foulds JR, Geumlek J, Welling M, Chaudhuri K (2016) On the theory and practice of privacy-preserving bayesian data analysis. In: Ihler AT, Janzing D (eds) *Proceedings of the thirty-second conference on uncertainty in artificial intelligence, UAI 2016*. *AUAI Press*, 25–29 June 2016
34. Griffiths TL, Steyvers M, Tenenbaum JB (2007) Topics in semantic representation. *Psychol Rev* 114(2):211
35. Dwork C, McSherry F, Nissim K, Smith AD (2016) Calibrating noise to sensitivity in private data analysis. *J Priv Confidentiality* 7(3):17–51
36. Park M, Foulds JR, Chaudhuri K, Welling M (2020) Variational bayes in private settings (VIPS). *J Artif Intell Res* 68:109–157
37. Hofmann T (1999) Probabilistic latent semantic indexing. In: Gey FC, Hearst MA, Tong RM (eds) *SIGIR '99: proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. *ACM*, 15–19 Aug 1999, Berkeley, CA, USA, pp 50–57
38. Espinosa KLC, Barajas J, Akella R (2012) The generalized dirichlet distribution in enhanced topic detection. In: Chen X, Lebanon G, Wang H, Zaki MJ (eds) *21st ACM international conference on information and knowledge management, CIKM'12*. *ACM*, maui, HI, USA, 29 Oct - 02 Nov 2012, pp 773–782
39. Bakhtiari AS, Bouguila N (2014) A variational bayes model for count data learning and classification. *Eng Appl Artif Intell* 35:176–186
40. Blei DM, Lafferty JD (2005) Correlated topic models. In: *Advances in neural information processing systems 18 [neural information processing systems, NIPS 2005, 5–8 Dec 2005, Vancouver, British Columbia, Canada]*, pp 147–154
41. Putthividhya D, Attias HT, Nagarajan SS (2009) Independent factor topic models. In: Danyluk AP, Bottou L, Littman ML (eds) *Proceedings of the 26th annual international conference on machine learning, ICML 2009, Montreal, Quebec, Canada, 14–18 June 2009*. *ACM international conference proceeding series*. *ACM*, vol 382, pp 833–840
42. Wang X, McCallum A, Wei X (2007) Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: *Proceedings of the 7th IEEE international conference on data mining (ICDM 2007)*. *IEEE computer society*, 28–31 Oct 2007, Omaha, Nebraska, USA, pp 697–702
43. Chaudhuri K, Sarwate AD, Sinha K (2012) Near-optimal differentially private principal components. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012*. *Proceedings of a meeting held 3–6 December 2012, Lake Tahoe, Nevada, United States*, pp 998–1006
44. Xu C, Ren J, Zhang D, Zhang Y, Qin Z, Ren K (2019) Ganobfuscator: mitigating information leakage under GAN via differential privacy. *IEEE Trans Inf Forensics Secur* 14(9):2358–2371
45. Huang Z, Hu R, Guo Y, Chan-Tin E, Gong Y (2020) DP-ADMM: Admm-based distributed learning with differential privacy. *IEEE Trans Inf Forensics Secur* 15:1002–1012
46. Bassily R, Smith AD, Thakurta A (2014) Private empirical risk minimization: efficient algorithms and tight error bounds. In: *55th IEEE annual symposium on foundations of computer science, FOCS 2014*. *IEEE computer society*, philadelphia, PA, USA, 18–21 Oct 2014, pp 464–473
47. Shokri R, Shmatikov V (2015) Privacy-preserving deep learning. In: *53rd Annual allerton conference on communication, control, and computing, allerton 2015, allerton park & retreat center*. *IEEE*, monticello, IL, USA, 29 Sept – 2 Oct 2015, pp 909–910
48. Park M, Foulds JR, Chaudhuri K, Welling M (2020) Variational bayes in private settings (VIPS). *J Artif Intell Res* 68:109–157
49. Sun M, Tay WP (2020) On the relationship between inference and data privacy in decentralized iot networks. *IEEE Trans Inf Forensics Secur* 15:852–866
50. Decarolis C, Ram M, Esmaeili S, Wang Y, Huang F (2020) An end-to-end differentially private latent dirichlet allocation using a spectral algorithm. In: *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 Jul 2020, virtual event*. *Proceedings of machine learning research*. *PMLR*, vol 119, pp 2421–2431
51. Dimitrakakis C, Nelson B, Mitrokotsa A, Rubinstein BIP (2014) Robust and private bayesian inference. In: Auer P, Clark A, Zeugmann T, Zilles S (eds) *Algorithmic learning theory - 25th international conference, ALT 2014, bled, Slovenia, 8–10 Oct 2014*. *Proceedings. Lecture notes in computer science*. *Springer*, vol 8776, pp 291–305
52. Porteous I, Newman D, Ihler AT, Asuncion AU, Smyth P, Welling M (2008) Fast collapsed gibbs sampling for latent dirichlet allocation. In: Li Y, Liu B, Sarawagi S (eds) *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*. *ACM*, Las Vegas, Nevada, USA, 24–27 Aug 2008, pp 569–577
53. Teh YW, Newman D, Welling M (2006) A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In: Schölkopf B, Platt JC, Hofmann T (eds) *Advances in neural information processing systems 19*, *Proceedings of the twentieth annual conference on neural information processing systems*. *MIT Press*, Vancouver, British Columbia, Canada, 4–7 Dec 2006, pp 1353–1360
54. Sadman N, Anjum N, Gupta KD (2020) Introduction of covid-news-us-nnk and covid-news-bd-nnk dataset
55. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR 2006)*. *IEEE computer society*, 17–22 June 2006, New York, pp 2169–2178
56. Wagner W, Bird S, Klein E, Loper E (2010) *Natural language processing with python, analyzing text with the natural language toolkit - o'reilly media*, Beijing, 2009. *Lang Resour Eval* 44(4):421–424
57. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Cheng X, Li H, Gabrilovich E, Tang J (eds) *Proceedings of the eighth ACM international conference on web search and data mining, WSDM 2015*. *ACM*, Shanghai, China, 2–6 Feb 2015, pp 399–408

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Zhiwen Luo received the B.S. degree in computer science from George Mason University, VA, USA, in 2019, and the M.Sc. degree in information systems security from Concordia University, Montreal, QC, Canada, in 2022. He is currently a Ph.D. student at the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada. His research interests include machine learning, deep learning and pattern recognition.



Nizar Bouguila received the engineer degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently a Professor with the Concordia Institute for Information Systems Engineering (CIISE) at Concordia University, Montreal, Quebec, Canada. His research interests include image processing,

machine learning, data mining, computer vision, pattern recognition, smart buildings, and energy. Prof. Bouguila received the best Ph.D Thesis Award in Engineering and Natural Sciences from Sherbrooke University in 2007. He was awarded the prestigious Prix d'excellence de l'association des doyens des études supérieures au Québec (best Ph.D Thesis Award in Engineering and Natural Sciences in Quebec), and was a runner-up for the prestigious NSERC doctoral prize. He was the holder of a Concordia University research Chair Tier 2 from 2014 to 2019 and was named Concordia University research Fellow in 2020. He is the author or co-author of more than 500 publications in several prestigious journals and conferences. He is a regular reviewer for many international journals and serving as associate editor for several journals such as Pattern Recognition journal. Dr. Bouguila is a licensed Professional Engineer registered in Ontario, and a Senior Member of the IEEE.



Manar Amayri received a bachelor's degree in power engineering from Damascus University Syria, a master's degree in electrical power systems from Power Department, Damascus University, a master's degree in smart grids and buildings from ENES3, INP-Grenoble (Institute National Polytechnique de Grenoble), in 2014, and the Ph.D. degree in energy smart-buildings from the Grenoble Institute of Technology, in 2017. She was a postdoctoral


researcher at INP-Grenoble and then at Concordia University from 2017 to 2020. She is currently an assistant professor in ENES3, INP-Grenoble, G-SCOP Lab (Sciences pour la conception, l'Optimisation et la Production). Her research interests include data mining, machine learning, explainable AI, energy, and smart buildings.



Wentao Fan received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Concordia University, Montreal, QC, Canada, in 2009 and 2014, respectively. He is currently an Associate Professor with the Department of Computer Science, Beijing Normal University-Hong Kong Baptist University (BNU-HKBU) United International College, Zhuhai, Guangdong, China. He has published more than 100 publications in several

prestigious peer-reviewed journals and international conferences. His research interests include machine learning, computer vision, and pattern recognition.

Affiliations

Zhiwen Luo¹  · Manar Amayri^{1,2} · Wentao Fan³ · Nizar Bouguila¹

Manar Amayri
manar.amayri@grenoble-inp.fr

Wentao Fan
wentaofan@uic.edu.cn

Nizar Bouguila
nizar.bouguila@concordia.ca

¹ The Concordia Institute for Information Systems
Engineering (CIISE), Concordia University,
Montréal, H3H 1M8, Québec, Canada

² G-SCOP Lab, Grenoble Institute of Technology,
Grenoble, 38031, France

³ Department of Computer Science,
Beijing Normal University-Hong Kong Baptist University
United International College (UIC),
Zhuhai, Guangdong, 519088, China