A Model-Agnostic Approach to Differentially Private Topic Mining

Han Wang*
Jayashree Sharma*
hwang185@hawk.iit.edu
jsharma6@hawk.iit.edu
Illinois Institute of Technology
USA

Shuya Feng
Kai Shu
sfeng16@hawk.iit.edu
kshu@iit.edu
Illinois Institute of Technology
USA

Yuan Hong
yuan.hong@iit.edu
yuan.hong@uconn.edu
Illinois Institute of Technology
University of Connecticut
USA

ABSTRACT

Topic mining extracts patterns and insights from text data (e.g., documents, emails and product reviews), which can be used in various applications such as intent detection. However, topic mining can result in severe privacy threats to the users who have contributed to the text corpus since they can be re-identified from the text data with certain background knowledge. To our best knowledge, we propose the first differentially private topic mining technique (namely TopicDP) which injects well-calibrated Gaussian noise into the matrix output of any topic mining algorithm to ensure differential privacy and good utility. Specifically, we smoothen the sensitivity for the Gaussian mechanism via sensitivity sampling, which addresses the major challenges resulted from the high sensitivity in topic mining for differential privacy. Furthermore, we theoretically prove the differential privacy guarantee under the Rényi differential privacy mechanism and the utility error bounds of TopicDP. Finally, we conduct extensive experiments on two real-word text datasets (Enron email and Amazon Reviews), and the experimental results demonstrate that TopicDP is a model-agnostic framework that can generate better privacy preserving performance for topic mining as compared against other differential privacy mechanisms.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**; • **Information systems → Data mining**.

KEYWORDS

Differential Privacy; Topic Mining; Model-Agnostic; Utility

ACM Reference Format:
Han Wang, Jayashree Sharma, Shuya Feng, Kai Shu, and Yuan Hong. 2022. A Model-Agnostic Approach to Differentially Private Topic Mining. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539417

1 INTRODUCTION

Billions of documents are generated everyday via personal computers, email servers, IoT devices, cloud, among others. These documents include considerable amounts of information, and they are frequently collected and shared for text analysis in various applications. One important type of applications is to extract topics from those documents, which can facilitate sentiment analysis [22], opinion summarization [28], recommender systems [31], and anomalous texts detection [23]. Topic mining statistically analyzes a corpus of documents to identify the discussion topics in them. The text data in each document is analyzed using probabilistic models and statistical analysis to discover patterns for the underlying topics. Thus, topic mining is widely used in the above real-world applications.

However, when topics of those documents are extracted and shared to untrusted third parties for further analysis, it raises severe privacy risks since the untrusted recipients may re-identify the owners of those documents with a diverse set of possible background knowledge related to the user and his/her documents (e.g., some keywords in the documents, and linguistic patterns). Thus, privacy preserving solutions for topic mining should be explored.

A simple privacy-enhancing technique is to replace the real user IDs of these documents with pseudonyms. This has been proven to be vulnerable to re-identification attacks [18, 27] (e.g., the AOL data leak incident [21]). As a rigorous privacy model against arbitrary background knowledge known to the adversaries, *differential privacy* (DP) has been extensively studied to address the privacy concerns in the text data [17, 21, 32]. However, such techniques only consider the privacy of term frequencies or related quantities in the documents. In practice, topic mining with differential privacy for documents should be a more complicated function rather than calculating the frequency of terms.

To our best knowledge, we propose the first differentially private topic mining technique (namely TopicDP) that protects the privacy of individuals involved in the documents used for any topic mining model (model-agnostic). It ensures indistinguishable analysis result derived from the input data with and without any user's all the documents. Specifically, we attempt to inject well-calibrated Gaussian noise into the result of topic mining (usually as a matrix with probability entries for different keywords in different topics), which would work regardless of the topic mining models. Thus, the untrusted recipient cannot distinguish whether any user is included in the documents or not.

Recall that topic mining generates a matrix output in which each row represents a topic and each entry in the row is a keyword and its probability of occurrence in the topic. Therefore, different

*Both authors contributed equally to this research.

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539417

from generic differential privacy mechanisms on statistical queries [10, 19, 25], TopicDP should address three major challenges: (1) topic mining generates a matrix output (many entries) rather than an aggregated value such as count, max, average, and sum, (2) each user's documents may include a unique word, which is not found in other users' documents, and (3) each user may include a large number of documents (high sensitivity).

To this end, we define a novel privacy notion for protecting the individuals in the documents for topic mining: "$(\epsilon, \delta, \gamma)$-random differential privacy", which is a relaxed notion extended from ϵ-differential privacy. First, δ (close to 0, e.g., 0.0001) is used to ensure that the probability of generating any unique word in the output matrix is bounded by δ. Second, γ (close to 0, e.g., 0.02) is used to smoothen the sensitivity [29] such that at least (1-γ) portion of users can be protected with (ϵ, δ)-differential privacy, which ensures indistinguishability (bounded by e^{ϵ}) for the topic mining results regardless of the presence or absence of each user's all documents. Moreover, we apply the Rényi differential privacy (RényiDP) mechanism to approximate (ϵ, δ)-differential privacy, and design algorithms that evaluate the Rényi divergence for the output of neighboring matrices and add noise for (ϵ, δ)-differential privacy.

Thus, the contributions of this paper are summarized as below:

- We propose the first model-agnostic differentially private technique TopicDP to protect the privacy of users in the documents used for topic mining. The well-calibrated noise is generated for the matrix output of topic mining, and thus works for any topic mining algorithm (model-agnostic). It can also be readily extended for other machine learning models which generate matrix outputs.
- We design a novel differential privacy mechanism for sensitivity smoothing of topic mining. This new relaxed privacy notion could significantly improve the utility of topic mining while approximating (ϵ, δ)-differential privacy.
- We learn the relationship between (ϵ, δ)-differential privacy and Rényi differential privacy in topic mining, and provide an approach to approximate (ϵ, δ)-DP for $1 - \gamma$ portion of the dataset with the RényiDP mechanism.
- We formally prove the privacy, utility error bound and the smaller noise with the RényiDP mechanism to ensure (ϵ, δ)-DP for $1 - \gamma$ portion of the dataset.
- We have conducted extensive experiments to validate the performance of TopicDP on two real text datasets under different topic mining models.

2 PRELIMINARIES

In this section, we present some preliminaries, including the topic mining and the differential privacy models.

2.1 Topic Mining Models

Recall that topic mining aims to identify the foci of discussions from a set of text documents (e.g., emails). Specifically, topic mining discovers the patterns of words to represent the topics, each of which is defined as a set of words that frequently occur together in the text corpus. For instance, while mining topics from a bunch of emails, the email bodies are analyzed to identify the topics to discover patterns of word usage. There are many existing algorithms used for topic mining, e.g., Latent Dirichlet Allocation (LDA) [1], Latent Semantic Indexing (LSI) [6], Probabilistic Latent Semantic Analysis (PLSA) [14], and Hierarchical Dirichlet Processing Model (HDP) [34]. Our TopicDP algorithm can guarantee differential privacy for any of the topic mining models.

Some generative probabilistic models (e.g., LDA) define the topic as a group of words that have a high likelihood of co-occurrence in the document corpus (see details in Appendix B). These words can be noun, verb, adjective and adverb. Therefore, topic mining generates a set of keywords and their corresponding probabilities in the topic. Thus, the output of topic mining can be denoted as a matrix of keywords and their probabilities. Each row in the matrix represents a topic discussed in the text documents, and the entries in the row refer to the probabilities of different keywords in the given topic (sum of the entries in each row is 1).

Formally, we denote the output of topic mining as a matrix W (rows: top m topics, columns: n words) with $m \times n$ probabilities $W \in [0, 1]^{m \times n}$. Since different topics may involve different sets of keywords, n words are the union of the keywords in all m top topics, and the probability in the matrix would be 0 if any topic does not include such word.

2.2 Differential Privacy (DP)

To protect the matrix outputs of topic mining, we first consider two input document datasets D and D' that differ in any user as two neighboring datasets. It is worth noting that any user may have multiple documents (e.g., all his/her emails) in the dataset. The DP model in case of topic mining would be interpreted as: adding or removing any user's all documents should not cause significant changes to the output of the topic mining. Thus, the privacy risks resulted from each user's documents can be bounded, even if the adversary possesses arbitrary background knowledge on all the users. Therefore, the DP model can be defined as below:

Definition 1 (ϵ-Differential Privacy). *A randomized mechanism $\mathcal{A}$ satisfies ϵ-differential privacy if for any two input datasets D and D' that differ in any user u (including at least one document), and for any output $S \in range(\mathcal{A})$, we have $e^{-\epsilon} \leq \frac{Pr[\mathcal{A}(D) \in S]}{Pr[\mathcal{A}(D') \in S]} \leq e^{\epsilon}$.*

Before designing a DP mechanism for topic mining, we need to explore its sensitivity. We start from the global sensitivity which refers to the maximum output difference of the function applied to the neighboring datasets. For example, the L_2-norm sensitivity of the query function f for Gaussian mechanism [7] is:

Definition 2 (Global Sensitivity). *Given a function f, its L_2-sensitivity is defined as:*

$$\Delta_{gs}(f) = \max_{D,D'} \left\| f(D) - f(D') \right\|_2 \tag{1}$$

As discussed in Section 1, different from generic DP mechanisms (e.g., Laplace, and Gaussian), we should address three new challenges in the mechanism design for TopicDP. First, topic mining will be considered as a complex function that generates a matrix output with probability entries. Thus, we first define the sensitivity for such function that returns matrix entries [3]:

Definition 3 (Sensitivity for Matrix-Output Function). *Given a matrix-output function $f(D) \in \mathbb{R}^{m \times n}$, the L_2-sensitivity is,*

$$\Delta(f) = \max_{D,D'} \|f(D) - f(D')\|_F \tag{2}$$

where $||\cdot||_F$ is the Frobenius norm.

The global sensitivity of topic mining might be very high since each user may include a large number of documents. Thus, the noise may be very large for the output.

Second, every user may include some unique words. Given any output P that includes any unique keyword from any user, the probabilities of applying topic mining to D and D' to generate such output P cannot be bounded with $\frac{Pr[\mathcal{A}(D)=S]}{Pr[\mathcal{A}(D')=S]} \leq e^{\epsilon}$ and $\frac{Pr[\mathcal{A}(D')=S]}{Pr[\mathcal{A}(D)=S]} \leq e^{\epsilon}$ since one of $Pr[\mathcal{A}(D) = S]$ and $Pr[\mathcal{A}(D') = S$ is equal to 0. Thus, the probabilities of such extreme cases should be bounded by a small number δ to ensure (ϵ, δ)-differential privacy.

To address these two challenges, we relax the protection to $(\epsilon, \delta, \gamma)$-random differential privacy (RDP) [29] where the confidence parameter of satisfying (ϵ, δ)-DP is denoted as $\gamma \in [0, 1)$. Then, $1 - \gamma$ portion of the dataset satisfies (ϵ, δ)-DP while the probability of any unique keyword in the output is bounded by δ.

Definition 4 ($(\epsilon, \delta, \gamma)$-Random Differential Privacy). *A randomized mechanism $\mathcal{A} : D^N \longrightarrow \mathbb{R}$ responding with values in arbitrary response set $\mathbb{R}$ preserves $(\epsilon, \delta, \gamma)$-RDP, at privacy level $\epsilon > 0, \delta \in [0, 1)$, and confidence parameter $\gamma \in [0, 1)$, if $Pr[\forall S \subset \mathbb{R}, Pr[(\mathcal{A}(D) \in S) \leq e^{\epsilon} \cdot Pr(\mathcal{A}(D') \in S) + \delta] \geq Pr[|f(D) - f(D')| \leq \Delta] \geq 1 - \gamma$, with the inner probabilities over the mechanism's randomization, and the outer probability over neighboring datasets D, D'.*

Intuitively, given the sensitivity $\Delta > 0$, when neighboring datasets D and D' satisfy $|f(D) - f(D')| \leq \Delta$, the randomized mechanism $\mathcal{A}(D, \epsilon, \delta, \gamma)$ enjoys (ϵ, δ)-DP. Then, the probability of holding ϵ-DP is at least $(1 - \delta)(1 - \gamma)$ since the maximum leakage occurs if two leakages are disjoint [16]. Thus, TopicDP satisfies such DP notion with minor relaxations since very small δ and γ make the probability $(1 - \delta)(1 - \gamma)$ very close to 1.

Third, we have to inject the same well-calibrated Gaussian noise to all the matrix entries. In practice, some real-world topics might be disjoint (e.g., from completely different users), then the matrix entries may follow parallel composition [25]. However, to protect the worst case that the involved user records of each topic in the matrix W are all correlated, all entries of matrix satisfy sequential composition and the privacy budgets have to be allocated to each entry. Thus, the budget for each entry might be too small to retain good utility in the randomized matrix output.

To address this challenge, we apply a relaxation of differential privacy based on the Rényi divergence [26], which evaluates the divergence of overall data. For two probability distributions P and Q over $\mathbb{R}$, the Rényi divergence of order α is $\mathcal{D}_{\alpha}(P||Q) = \frac{1}{\alpha-1} \log \mathcal{E}_{x \sim Q} [\frac{P(x)}{Q(x)}]^{\alpha}$. Thus, the privacy notion is defined as below:

Definition 5 ((α, ϵ)-Rényi Differential Privacy). *A randomized mechanism $\mathcal{A} : \mathcal{D} \mapsto \mathbb{R}$ is said to satisfy ϵ-Rényi differential privacy (RényiDP) of order α, if for any D, D' it holds that $D_{\alpha}[f(D)||f(D')] \leq \epsilon$.*

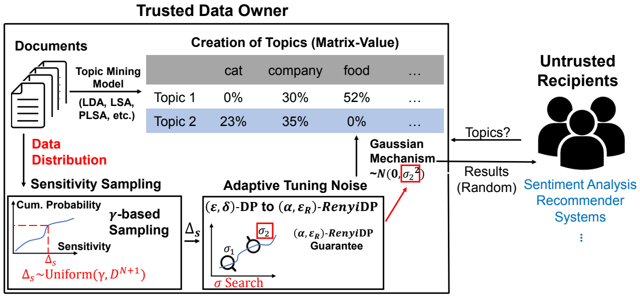
The definition of ϵ-DP coincides with (∞, ϵ)-RényiDP. By monotonicity of the Rényi divergence, (∞, ϵ)-RényiDP implies (α, ϵ)-RDP for all finite α[26]. In turn, an (α, ϵ)-RDP implies (ϵ, δ)-differential privacy for any given probability $\delta > 0$. The α can be any number other than 1. In our formulation, we treat α as a parameter in optimization of error bound. The outcomes of $f(D)$ and $f(D')$ should be all feasible probability distribution output of topic mining.

3 TOPIC MINING WITH DIFFERENTIAL PRIVACY

In this section, we illustrate the details of TopicDP.

3.1 Threat Model

We adopt the standard threat model setting of differential privacy in this paper. A trusted data owner collects a large number of users' documents and perturbs the topic mining algorithm with DP guarantee. Thus, other untrusted data recipients (adversaries) would request the topic mining analysis on the documents, but cannot infer if any user's document (e.g., any user's email) is included in the topic mining even if the adversaries possess arbitrary background knowledge (e.g., knowing the contents of all the users' emails). We assume that all the parties are honest-but-curious to follow the procedures without maliciously manipulating the data.



	cat	company	food	…
Topic 1	0%	30%	52%	…
Topic 2	23%	35%	0%	…

Figure 1: Overview of the TopicDP Framework.

3.2 The TopicDP Framework

Since TopicDP requires (ϵ, δ)-DP, we adapt the Gaussian mechanism [7] in TopicDP which is widely used to ensure (ϵ, δ)-DP by injecting a Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the query (or analysis function) where $\sigma^2 = 2\Delta^2 \log(1.25/\delta)/\epsilon^2$ and Δ refers to the sensitivity of the query/function. However, as discussed earlier, the sensitivity might be too large and the privacy budget might be too small to generate noise. Thus, we propose a framework that injects well-calibrated Gaussian noise into the matrix. Such noise can satisfy the privacy requirement and provide better utility.

Figure 1 illustrates the framework of TopicDP. We carefully inject Gaussian noise into the topic matrix (*output perturbation*) with the sampled sensitivity and variance σ^2 of Gaussian mechanism. Thus, DP can be ensured for any topic mining algorithm (*model-agnostic*). TopicDP includes the following major steps.

Step 1 : The trusted data owner collects all documents D from N users, and specifies the privacy parameters (ϵ, δ).

Step 2 : The untrusted recipient requests to identify topics from all documents D.

- **Step 3** : The trusted data owner applies any topic mining algorithm (e.g., LDA) to extract the topic matrix W with probability entries for a set of keywords.
- **Step 4** : The trusted data owner first uniformly samples the smooth sensitivity Δ (denoted as Δ_s) of dataset D with parameter γ. Then, the trusted data owner converts the given (ϵ, δ)-DP guarantee to (α, ϵ_R)-RényiDP with privacy parameters ϵ and δ. Finally, the parameter σ_R of Gaussian noise is searched to satisfy the (α, ϵ_R)-RényiDP with sampled sensitivity Δ_s. The noise with variance σ_R from Gaussian mechanism is injected to the topic matrix W.
- **Step 5** : The noisy topic matrix W' (normalized) is returned to the untrusted recipient for further analyses.

Note that the sensitivity sampling in Step 4 is extended from the Pain-Free algorithm [29] (which only considers a single aggregated result) to the matrix-output query. We will discuss the details for smoothing the sensitivity using the distribution in the dataset D and the value of σ_R to satisfy the privacy requirement as follows.

```
Input   : Database size $N$, topic mining function $f$, the
          distribution $P$, the confidence parameter $\gamma$
Output  : The sampling sensitivity $\Delta_s$
Compute the sample size $h = \left\lceil \frac{\log(1/\rho)}{2(\gamma-\rho)^2} \right\rceil$ where
  $\rho = exp(W_{-1}(-\frac{\gamma}{2\sqrt{e}}) + 0.5)$
Compute the order statistic index
  $k = h(1 - \gamma + \rho + \sqrt{\log(1/\rho)/(2h)})$
$P \leftarrow Uniform()$
foreach $i = 1$ to $h$ do
    Sample $D_{1,..,N-1} \sim P^N$, $D_N \sim P^N$ and $D_{N+1} \sim P^N$
    $D \leftarrow D_{1,..,N-1} \cup D_n$
    $D' \leftarrow D_{1,..,N-1} \cup D_{N+1}$
    $\Delta_i = \|f(D) - f(D')\|_F$
end
Sort $\Delta_1, ...., \Delta_h$ with the ascending order
Return $\Delta_s = \Delta_k$
```

Algorithm 1: Sensitivity Sampling

3.3 Sensitivity Derivation

The sensitivity sampler in Pain-Free algorithm [29] obviates the challenge of unbounded sensitivity in DP. With it, we can approximate the global sensitivity with very high probability, assuming only oracle access to the target query function f evaluations.

Algorithm 1 presents the details of sampling sensitivity for topic mining. With the given confidence parameter γ, we compute the value of sampling size h and order index k, which can guarantee $(\epsilon, \delta, \gamma)$-RDP (see detailed privacy analysis in Section 3.6). These two parameters are involved in the the Lambert-W function [30]. The distribution P is chosen to match the desired distribution of dataset D. There are a number of natural choices for the dataset distribution P [29]. We use the uniform distribution for our documents since the Pain-Free algorithm and its privacy guarantee are derived by assuming a uniform distribution defined over the domain D. We empirically demonstrate the improvement of utility with the sensitivity sampler for matrix-output query rather than the single aggregated result (see details in Section 4). However, the accuracy should be significantly boosted if we can approximate the distribution of dataset with some background knowledge (e.g., users' linguistic patterns).

With the distribution P, in each iteration, we independently sample the $N+1$ records from the domain to construct the database D and D' which differ in one user. Then, the sensitivity can be computed for these pairs of neighboring datasets. After h iterations, there are h sensitivities and sort them in an ascending order. The final smooth sensitivity should be the kth sensitivity.

3.4 Parameter σ_R Derivation

Proposition 1 (From RényiDP to (ϵ, δ)-DP [26]). *If f satisfies (α, ϵ)-RényiDP, it also satisfies $(\epsilon + \frac{log\frac{1}{\delta}}{\alpha-1}, \delta)$-differential privacy.*

Proposition 2 (From (ϵ, δ)-DP to (α, ϵ_R)-RényiDP [2]). *For any $\alpha > 1$, $\epsilon \geq 0$, $\delta \in (0, 1)$, and $0 < \alpha\delta < 1$, we have:*

$$\epsilon_R \geq \max\{\epsilon - \frac{1}{\alpha - 1}\log\frac{\zeta_\alpha}{\delta}, \epsilon + \frac{1}{\alpha - 1}\log((e^\epsilon - \alpha\delta)(\frac{\delta - 1}{\delta - e^\epsilon})^\alpha + \alpha\delta)\}$$

where $\zeta_\alpha = \frac{1}{\alpha}(1 - \frac{1}{\alpha})^{\alpha-1}$.

With the Proposition 1 and 2, the relationship between DP and RényiDP enables us to derive the optimal RényiDP parameters of a mechanism that satisfies a given level of strict DP. Thus, with the given (ϵ, δ)-DP requirement, we apply the mechanisms that satisfy the corresponding RényiDP. It is worth noting that we only consider α as an parameter to provide results when calculating the Rényi divergence. We set the value $\alpha = 2$ and use the exact map of privacy values from ϵ to ϵ_R [2]. We illustrate the relationship between DP and RényiDP with $\alpha = 2$, $\delta = 0.0001$, $\gamma = 0.1$ as an example in Figure 2. The parameter σ_R then has to satisfy $\sigma_R^2 \geq \frac{\alpha\Delta_s^2}{2\epsilon_R}$ for satisfying the (α, ϵ_R)-RényiDP with Gaussian mechanism. We will prove that it satisfies (α, ϵ_R)-RényiDP with the condition that $\sigma_R^2 \geq \frac{\alpha\Delta_s^2}{2\epsilon_R}$ and the noise amount to guarantee (ϵ, δ)-DP under a composition of RényiDP mechanisms is less than the noise amount under direct DP mechanism in Section 3.6.

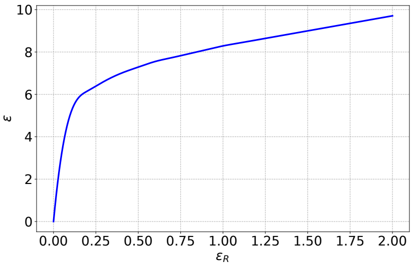


Figure 2: Privacy Bound: ϵ in DP equals ϵ_R in RényiDP ($\alpha = 2$, $\delta = 0.0001$, $\gamma = 0.1$).

3.5 Algorithm for Differential Privacy

After computing the parameter σ_R, the DP algorithm can be designed. Algorithm 2 presents the details for topic mining (matrix outputs) with RényiDP mechanism to satisfy (ϵ, δ)-DP for $1 - \gamma$ portion of dataset. We then add the noise matrix to the topic mining output (the probability matrix) and release the noisy result. Note that the estimation of sampled sensitivity can always be the same for a specific domain. Thus, such sensitivity sampling could be performed entirely in an offline stage and executed once.

```
Input  : Database $D$, the query size $m \times n$, the sampled sensitivity
         $\Delta_s$, the parameter $\sigma_R$ and the privacy parameter $\epsilon_R$
Output : The noisy matrix result $W'$
Apply the topic mining algorithm to extract the $m$ topics and
  keywords probabilities $W$ of dataset $D$
foreach entry $e_i$ in the matrix $W$ do
    $\tilde{e}_i \leftarrow e_i + \mathcal{N}(0, \sigma_R^2)$ and $\sigma_R^2 \geq \frac{\alpha\Delta_s^2}{2\epsilon_R}$
    Update the $e_i$ with $\tilde{e}_i$
end
$W' \leftarrow W$
Return the noisy matrix $W'$
```

Algorithm 2: Generating Noise with Gaussian Mechanism

3.6 Privacy and Utility Analysis

We first analyze the privacy bound of TopicDP.

Theorem 1. *[29]: Consider any non-private function $f\colon D^N \to \mathcal{B}$, any sensitivity-induced (ϵ, γ)-differentially private mechanism mapping $\mathcal{B}$ to (randomized) responses in $\mathbb{R}$, any database D of N records, privacy parameters $\epsilon > 0, \delta \in [0, 1], \gamma \in (0, 1)$, and sampling parameters size $h \in \mathbb{N}$, order statistic index $h \geq k \in \mathbb{N}$, approximation confidence $0 < \rho < min\{\gamma, 1/2\}$, distribution P on D. If*

$$h \geq \left\lceil \frac{\log(1/\rho)}{2(\gamma - \rho)^2} \right\rceil \tag{3}$$

$$k \geq h\big(1 - \gamma + \rho + \sqrt{\log(1/\rho)/(2h)}\big) \tag{4}$$

then Algorithm 1 running with D, Δ_s, f, h, k, P, preserves $(\epsilon, \delta, \gamma)$-random differential privacy.

Proof. See Appendix A.1. □

Theorem 1 proves that the probability of RDP of $\mathcal{A}_{\Delta_s}$ running on fixed Δ_s is at least $Pr(G < \Delta_s)$, where G is the sampled sensitivity group $\Delta_1, ..., \Delta_h$ drawn from the Algorithm 1. Then, by the Dvoretzky-Kiefer-Wolfowitz inequality [24], we can prove the probability of RDP of $\mathcal{A}_{\Delta_s}$ is at least $1 - \gamma$.

Theorem 2. *Let $\epsilon > 0, \delta > 0, \alpha > 1$, for $\sigma_R^2 \geq \frac{\alpha\Delta_s^2}{2\epsilon_R}$, the Gaussian mechanism $\mathrm{G}_{\mathrm{f}}(\sigma_R^2)$ satisfies (α, ϵ_R)-RényiDP.*

Proof. With the definition of RényiDP, we can calculate the error bound of Rényi divergence between distribution P and Q of Gaussian distribution $\mathcal{N}(0, \sigma_R^2)$.

$$\mathcal{D}_\alpha(P(x)||Q(x)) = \frac{1}{\alpha - 1}\log\int_{-\infty}^{\infty} p(x)^\alpha q(x)^{1-\alpha}dx$$
$$= \frac{1}{\alpha - 1}\log\int_{-\infty}^{\infty} \frac{1}{\sigma_R\sqrt{2\pi}} exp(\frac{-\alpha x^2}{2\sigma_R^2}) exp(\frac{(\alpha - 1)(x - \Delta_s)^2}{2\sigma_R^2})dx$$
$$= \frac{1}{\alpha - 1}\log\{exp\frac{\alpha(\alpha - 1)\Delta_s^2}{2\sigma_R^2}\} = \frac{\alpha\Delta_s^2}{2\sigma_R^2}$$

Since $\mathcal{D}_\alpha(P||Q) \leq \epsilon_R$, we have $\sigma_R^2 \geq \frac{\alpha\Delta_s^2}{2\epsilon_R}$. Thus, the Gaussian mechanism satisfies (α, ϵ_R)-RényiDP. □

Theorem 3. *When $\alpha < [\frac{\xi - \sqrt{\xi^2 - 4\epsilon\log(1/\delta)}}{2\epsilon} + 1]^2$ in which $\xi = (2\sqrt{\log(1.25/\delta)}\epsilon - \epsilon)$, the noise amount to guarantee the given level of differential privacy under a composition of RényiDP mechanisms is less than the noise amount using DP mechanism directly.*

Proof. First, assuming that the goal is to achieve $(\epsilon + \frac{\log\frac{1}{\delta}}{\alpha - 1}, \delta)$-DP, we then can apply $\sigma = \frac{\sqrt{2\log(1.25/\delta)}\Delta}{\epsilon + \frac{\log\frac{1}{\delta}}{\alpha-1}}$ to guarantee $(\epsilon + \frac{\log\frac{1}{\delta}}{\alpha-1}, \delta)$-DP directly. Given the Proposition 1, we know that it needs to guarantee (α, ϵ)-RényiDP with RényiDP mechanism, which also satisfies $(\epsilon + \frac{\log\frac{1}{\delta}}{\alpha-1}, \delta)$-DP. To achieve (α, ϵ)-RényiDP, we can apply Gaussian noise with $\sigma_R = \sqrt{\frac{\alpha}{2}} \times \frac{\Delta}{\epsilon}$. Thus, we have:

$$\frac{\sigma}{\sigma_R} = \frac{\frac{\sqrt{2\log(1.25/\delta)}\Delta}{\epsilon + \frac{\log\frac{1}{\delta}}{\alpha-1}}}{\sqrt{\frac{\alpha}{2}} \times \frac{\Delta}{\epsilon}} \tag{5}$$

Note that $\frac{\sigma}{\sigma_R} > 1$ means the noise with DP mechanism to guarantee (ϵ, δ)-DP is greater than the noise with RényiDP mechanism, when $\frac{\sigma}{\sigma_R} > 1$, $0 < \delta < 1$, and $\alpha > 1$. Thus, we have $\alpha < [\frac{\xi - \sqrt{\xi^2 - 4\epsilon\log(1/\delta)}}{2\epsilon} + 1]^2$ in which $\xi = (2\sqrt{\log(1.25/\delta)}\epsilon - \epsilon)$. □

Theorem 4. *The noisy matrix output in Algorithm 2 satisfies (ϵ, δ)-differential privacy for $1 - \gamma$ portion of dataset.*

Proof. It is straightforward to prove that Algorithm 2 ensures (ϵ, δ)-DP. Per Theorem 1, the sampled sensitivity Δ_s ensures (ϵ, δ)-DP for $1 - \gamma$ portion of dataset D. With Theorem 2, adding the Gaussian noise with Δ_s to each entry of matrix W can make the divergence of two matrices of neighboring datasets D and D' bounded by ϵ_R, which also satisfies (ϵ, δ)-DP for $1 - \gamma$ portion of dataset. □

Finally, we analyze the utility error bound of TopicDP.

Theorem 5. *The expectation of the amplitude of noise in TopicDP is $\frac{2\sigma}{\sqrt{2\pi}}$ where $\sigma_R = \sqrt{\frac{\alpha\Delta_s^2}{2\epsilon_R}}$.*

Proof. See Appendix A.2. □

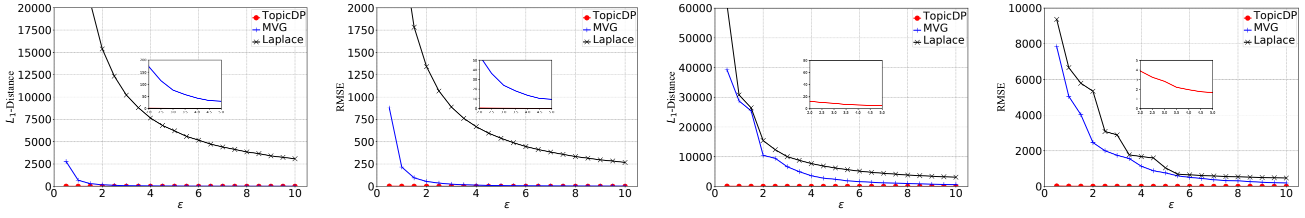
4 EXPERIMENTS

In this section, we present the experimental evaluations on TopicDP. Without loss of generality, we first use LDA as the example topic mining algorithm to examine the performance of TopicDP. We then also investigate the performance of proposed TopicDP under different topic mining models.

4.1 Experimental setting

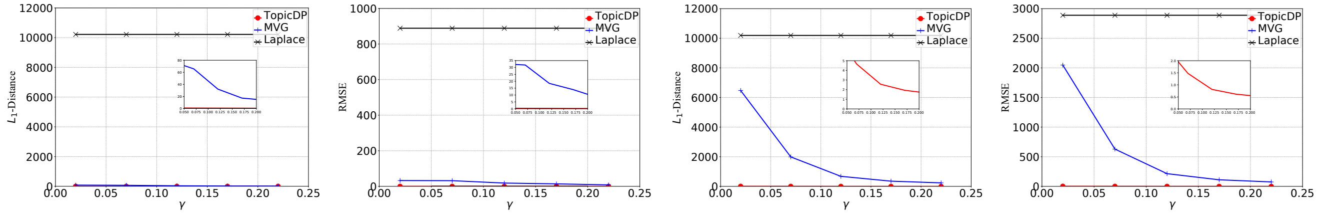
We conduct experiments on two real text datasets as below.

(1) **Enron Email Dataset** [20] is collected by the CALO Project. It contains data from about 158 users, organized into folders and contains thousands of mails exchanged among the employees. The dataset has been pre-processed to remove the email headers and duplicate emails. Figure 11(a) in Appendix C shows the three attributes of the dataset ("body", "to", "from"). The attribute "to" is used to identify the email recipients (users). All the emails associated to each "to" email address will be considered as a specific user's emails.



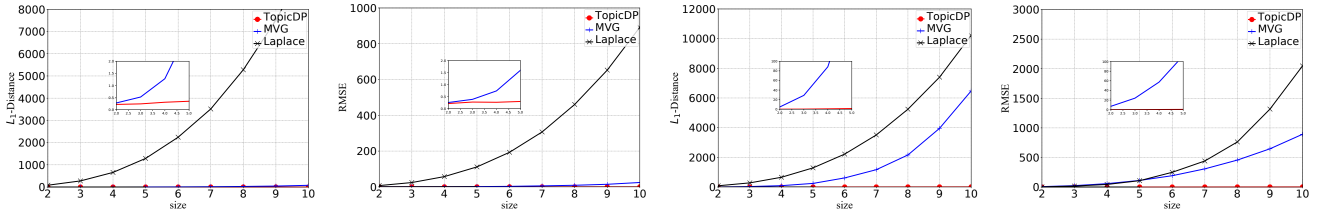
(a) L_1-Distance (b) RMSE (c) L_1-Distance (d) RMSE

Figure 3: L_1-Distance and RMSE vs Privacy Bound ϵ on the Enron Dataset (a, b) and Amazon Dataset (c, d).



(a) L_1-Distance (b) RMSE (c) L_1-Distance (d) RMSE

Figure 4: L_1-Distance and RMSE vs γ on the Enron Email Dataset (a, b) and Amazon Product Review Dataset (c, d).



(a) L_1-Distance (b) RMSE (c) L_1-Distance (d) RMSE

Figure 5: L_1-Distance and RMSE vs Matrix Size with Fixed $\epsilon = 3$ on the Enron Dataset (a, b) and Amazon Dataset (c, d).

(2) **Amazon Product Review Dataset** [12] is a collection of product reviews created by users on the product pages. It includes reviews (ratings, text, helpfulness votes), product metadata (e.g., descriptions, category information), and links (also viewed/bought graphs). We pre-processed the dataset to retain two attributes of this dataset. Figure 11(b) in Appendix C shows the retained attributes: user name and text reviews.

Table 1: Characteristics of Experimental Datasets

Dataset	User #	Docs #	Avg Word #	Avg Docs #/User
Enron	158	24,151	154	152
Amazon	3,815	50,000	31	13

Table 1 presents the characteristics of the datasets. To evaluate the utility of TopicDP, we perform three groups of experiments. First, we adopt the L_1*-distance* and the *Root-Mean-Square Error (RMSE)* metric to quantify the noise. Specifically, we compare the original output probability matrix and noisy output probability matrix with these two metrics. Second, we use the *Kendall's Tau* distance to evaluate the misalignment between the keyword ranking before and after adding noise. Finally, we visualize the keyword distributions in some example topics before and after noise.

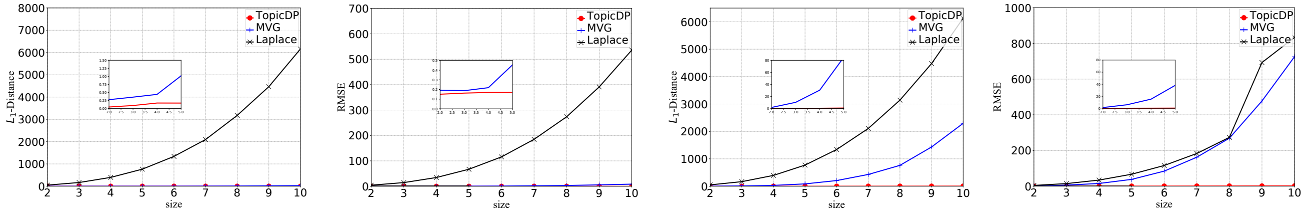
Moreover, we evaluate TopicDP by varying the privacy parameters ϵ and δ, confidence γ, and output matrix size $m \times n$ (top m topics and n keywords in each topic). We also set δ as a very small bound 10^{-4} and $m = n$ since one benchmark Multivariate Gaussian (MVG) mechanism [3] only works for square matrix outputs.

4.2 Evaluating TopicDP

We compare the TopicDP with the well-known Laplace mechanism (adapted for matrix outputs) and MVG mechanism [3] which also aims to protect privacy for matrix outputs.[1] Specifically, in the Laplace mechanism, we use the L_1 sensitivity and a global sensitivity $2m$ based on the extreme case that the topics and keywords are totally different (the probability difference of each topic should be 2). The setting of the MVG mechanism is the same as TopicDP in which the neighboring datasets differ in a single user and the sensitivity may be unbounded. However, in the MVG mechanism, there is a threshold for the unbounded sensitivity. For a fair comparison, we also use the smooth sensitivity in the MVG mechanism.

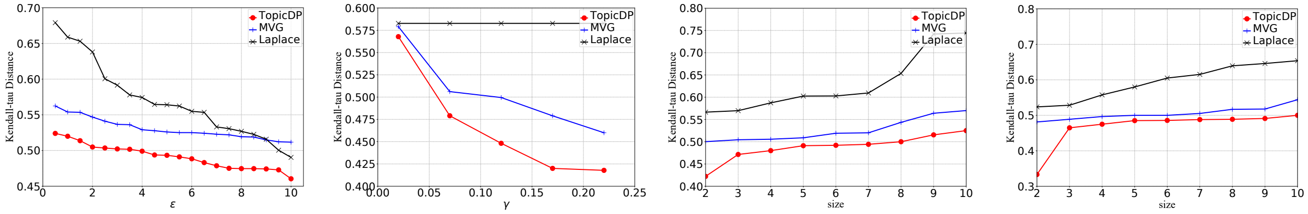
L_1 and RMSE. First, Figure 3 demonstrates the L_1-distance and RMSE by varying the privacy bound ϵ from 1 to 10 with a step

[1] [34] uses Laplace mechanism for the LDA algorithm, which cannot be model-agnostic for multiple topic mining algorithms (due to input/query perturbation). Thus, we adapt it to output perturbation for fair comparisons with TopicDP and the MVG mechanism.



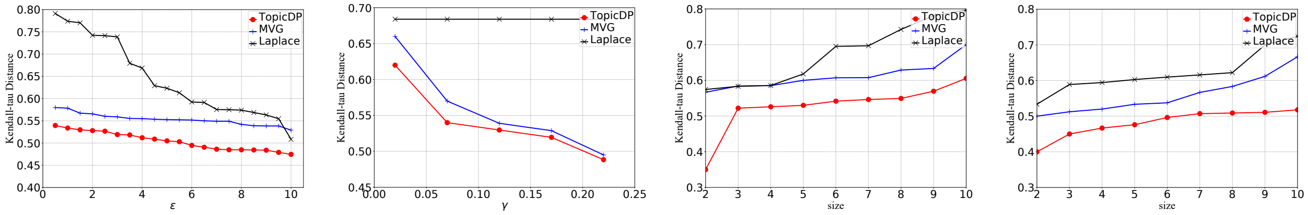
(a) L_1-Distance (b) RMSE (c) L_1-Distance (d) RMSE

Figure 6: L_1-Distance and RMSE vs Matrix Size with Fixed $\epsilon = 5$ on the Enron Dataset (a, b) and Amazon Dataset (c, d).



(a) ϵ (b) γ (c) size (ϵ=3) (d) size (ϵ=5)

Figure 7: Kendall's Tau Distance on the Enron Dataset.



(a) ϵ (b) γ (c) size (ϵ=3) (d) size (ϵ=5)

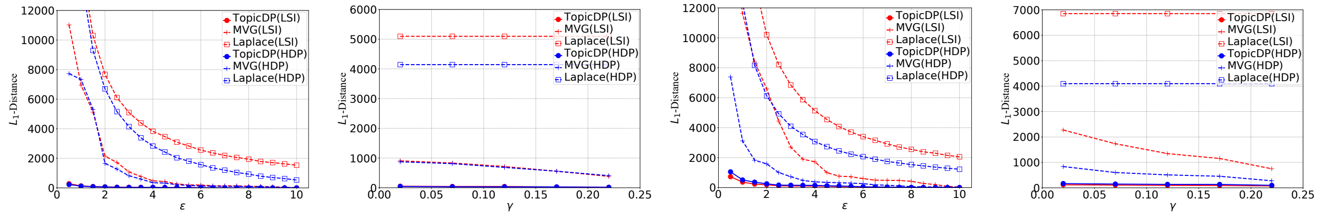
Figure 8: Kendall's Tau Distance on the Amazon Dataset.

of 0.5 while fixing the confidence $\gamma = 0.1$ and output matrix size as 10 × 10. Figure 3(a) and 3(b) show the L_1-distance and RMSE results on the Enron Email Dataset, respectively. Similarly, Figure 3(c) and 3(d) demonstrate the results on the Amazon Product Review Dataset. As ϵ increases, the L_1-distance and RMSE decrease (noise gets smaller for all mechanisms while increasing ϵ). Second, the noise generated by Laplace mechanism (output perturbation) is far larger than other two mechanisms due to the very large global sensitivity. Third, in Figure 3(a), given a small ϵ (strong privacy), the L_1-distance between the actual output and the noisy output is greater than 2500 in MVG mechanism but less than 6 in our TopicDP. For large ϵ (weak privacy, e.g., $\epsilon = 10$), the L_1-distance for the MVG mechanism is still much higher than TopicDP. Similarly, we can also observe such trend from Figure 3(b), 3(c) and 3(d).

Second, Figure 4 shows the L_1-distance and RMSE by varying the confidence parameter γ from 0.02 to 0.22 with a step of 0.05 and fixing $\epsilon = 3$ and output matrix size 10×10. Figure 4(a), 4(b), 4(c), and 4(d) show the results for L_1-distance and RMSE, respectively. Since the global sensitivity is not related to γ, the sensitivity and the utility of Laplace mechanism should not be changed. We can observe that, as the γ increases, the L_1-distance and RMSE of other two mechanisms generate smaller noise. The main reason is that smaller γ gives stronger privacy by ensuring ϵ-differential privacy for a higher percent of records (thus the sampled sensitivity should be larger). Second, although the MVG mechanism decreases drastically as γ increases, the lowest result of MVG is still higher than the result of TopicDP (e.g., the range of L_1 distance for TopicDP is from 0.1389 to 8.4426 whereas the range for MVG is from 4.8034 to 6475.0496, as shown in Figure 4(a)).

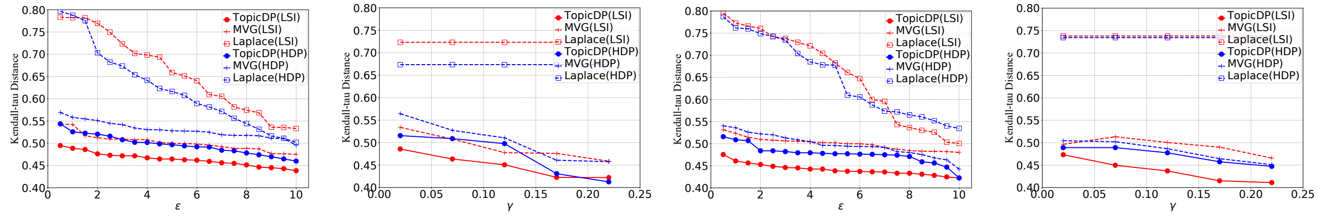
Third, Figure 5 shows the L_1-distance and RMSE by varying the matrix size $m \times n$ and fixing $\epsilon = 3$ and $\gamma = 0.1$. Both m and n vary from 2 to 10 with a step of 1. For both datasets, as m, n get larger, the L_1-distance exponentially increases for both Laplace and MVG, whereas TopicDP increases much slower. The increasing trends on growing $m \times n$ are consistent with the sequential composition for adding noise to the matrix entries. Larger m and n mean less privacy budget allocated for each entry and the utility should be worse. Clearly, the Laplace mechanism generates a much larger noise than other two mechanisms.

Thus, we will compare the utility of MVG and TopicDP. In Figure 5(a), the range of L_1-distance for TopicDP is from 0.0179 to 0.8954, whereas the range for MVG is from 0.0794 to 77.1487. In Figure 5(b), the range of RMSE for TopicDP is from 0.0122 to 0.2838, whereas it is from 0.0562 to 24.3965 for MVG. When the matrix size is small (e.g., 2 × 2 and 3 × 3), the noise generated by these two mechanisms can be similar in terms of the L_1-distance and RMSE metric. However,



(a) ϵ (b) γ (c) ϵ (d) γ

Figure 9: L_1 Distance under Different Models on the Enron Dataset (a, b) and Amazon Dataset (c,d).



(a) ϵ (b) γ (c) ϵ (d) γ

Figure 10: Kendall's Tau Distance under Different Models on the Enron Dataset (a, b) and Amazon Dataset (c,d).

when the matrix output size is large (e.g., greater than 5), TopicDP significantly outperforms MVG. We can draw the same conclusions from Figure 5(b), 5(c) and 5(d) as well. Furthermore, in Figure 6, we make the privacy bound $\epsilon = 5$, and compare the results with Figure 5. Then, all metric values get smaller, and this is consistent with the trends in Figure 3. Thus, we can conclude that TopicDP greatly outperforms the MVG mechanism and Laplace mechanism.

Kendall's Tau. We next evaluate the keywords rank of each topic for top popular topic before and after adding noise, by using the Kendall's Tau to measure the misalignment between every two sets of ranked keywords. The larger distance means higher dissimilarity.

Figure 7 shows the Kendall's Tau distance on the Enron Email dataset. With a larger ϵ, the Kendall's Tau distance is slightly smaller (see Figure 7(a)). This is consistent with what we observe in the results of L_1-distance and RMSE. We also observe that TopicDP outperforms MVG and Laplace mechnaisms. Furthermore, we can see the results get smaller as γ goes larger in Figure 7(b). Finally, in Figure 7(c) and 7(d), the distance increases as the output matrix size gets larger, but TopicDP always has the smallest distance for the best utility. In Figure 8, four sub-figures illustrate the similar observations and trends on the Amazon review dataset.

Keyword Distribution. Finally, we visualize specific topics extracted from two datasets and show the keyword distributions of them before and after adding noise. In Figure 12 and 13 (Appendix C), we randomly pick four topics from each dataset (topic 2, 4, 5 and 9 from the Enron dataset and topic 2, 3, 6 and 7 from the Amazon dataset). From these figures, we observe that the keyword distributions after injecting the noise using TopicDP are still close to the original distributions. This proves the practicality of the TopicDP.

4.3 Model-Agnostic Evaluation

In this section, we investigate the performance of TopicDP under other two topic mining models: the Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Processing (HDP). These two models both generate the probability matrix for key words. The setting is the same as the setting in the Section 4.2.

We first compare the noise amount with L_1-distance for these two models. Figure 9(a) and 9(c) present the L_1-distance with varying ϵ for the Enron and Amazon datasets. Blue lines denote the LSI model while red lines denote the HDP model. We can observe that, for any model, the noise is smaller as ϵ increases. Moreover, the noise generated by TopicDP is the smallest that is very close to the output probability matrix. Figure 9(b) and 9(d) present the L_1-distance results with varying confidence parameter γ. They also show similar observations and trends on these two datasets. Next, we investigate the performance of TopicDP on keyword rank, as shown in Figure 10. Figure 10(a) and 10(c) present the Kendall's Tau distance with varying privacy ϵ. Figure 10(b) and 10(d) present the Kendall's Tau distance with varying confidence γ. For any model and dataset, the distance generated by TopicDP is the smallest in which the rank is more similar before and after adding noise.

In summary, such experimental results validate that TopicDP retains good utility for different topic mining models (better than the two benchmarks) with rigorous privacy guarantees, which make TopicDP more practical in the real-world applications.

5 DISCUSSION

Model-Agnostic Differential Privacy. The proposed TopicDP only needs to perturb the output matrix, regardless of the topic mining algorithm used by the trusted data collector. Thus, TopicDP is a model-agnostic differentially private technique that can be deployed with any topic mining model.

Other Applications with Matrix Outputs. Besides the application of topic mining, many other applications (e.g., collaborative filtering recommender systems [13], graph queries [11], and histogram releasing [10]) also return the matrix outputs. TopicDP can also be readily adapted to preserve privacy in those applications by providing provable guarantees and good utility. Thus, TopicDP can universally work for any analysis that returns a matrix output even with a high sensitivity. It can provide better utility while guaranteeing strong privacy.

6 RELATED WORK

Privacy Preserving Text Analysis. Preserving user privacy in text analysis has been extensively studied in the past decades [4, 8, 15, 17, 32–34]. For instance, [4] proposes a privacy-preserving classification technique for personal text messages based on the secure multiparty computation, which encompasses both private feature extraction from texts, and private classification with logistic regression and tree ensembles. It proves that when using the secure text classification method, the application cannot learn anything about the texts, and the author of the text cannot learn anything about the text classification model either. [8] proposes a privacy preserving keyword search scheme for searching over encrypted data. To avoid the high computational cost of asymmetric encryption, this scheme employs symmetric encryption and Bloom filter.

Differentially Private Text Analysis. Several other works focus on the text analysis with differential privacy. Specifically, [32] proposes an automated text anonymization approach that produces synthetic term frequency vectors for the input documents with differential privacy. [9] presents a formal approach to preserve privacy in text perturbation using the notion of d_χ-privacy which is also extended from differential privacy. It considers the input distance between any two inputs of the domain to achieve indistinguishability. Some other works address privacy concerns in the Latent Dirichlet Allocation (LDA) training process [5, 34, 35]. For instance, [34] mainly proposes a HDP-LDA algorithm to protect the entire training process on centralized datasets. However, in their privacy model, the neighboring datasets only differ in one record and the HDP-LDA algorithm is based on the collapsed Gibbs sampling which adds noise to the word count statistics. [21] and [17] release search query logs with differential privacy while ensuring differential privacy for the query keywords and URLs.

However, these models are not model-agnostic and cannot be applied to topic mining (due to high sensitivity and matrix outputs).

7 CONCLUSION

There is a high risk on re-identifying individuals from the topic mining on documents with certain background knowledge. To our best knowledge, we propose the first model-agnostic differentially private topic mining technique that injects well-calibrated Gaussian noise (with smooth sensitivity) to the output of any topic mining algorithm. It can ensure high utility and guarantee differential privacy for at least $1 - \gamma$ portion of records (γ is close to 0). The noisy result can be privately shared to any untrusted recipient for further downstream analyses on the extracted topics.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation (NSF) under the Grants No. CNS-2046335 and CNS-2034870, as well as the Cisco Research Award. Also, the authors would like to thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] LDA. 2021.. NLP with LDA: Analyzing Topics in the Enron Email dataset.

[2] Shahab Asoodeh, Jiachun Liao, Flavio P Calmon, Oliver Kosut, and Lalitha Sankar. 2021. Three variants of differential privacy: Lossless conversion and applications. *IEEE Journal on Selected Areas in Information Theory* 2, 1 (2021), 208–222.

[3] Thee Chanyaswad, Alex Dytso, H Vincent Poor, and Prateek Mittal. 2018. MVG mechanism: Differential privacy under matrix-valued query. In *CCS*, 230–246.

[4] Martine De Cock, Anderson C Nascimento, Devin Reich, Rafael Dowsley, and Ariel Todoki. 2019. Privacy-Preserving Classification of Personal Text Messages with Secure Multi-Party Computation. In *NeurIPS*, 3752.

[5] Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. 2020. An end-to-end differentially private latent Dirichlet allocation using a spectral algorithm. In *ICML*, 2421–2431.

[6] Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology* 38, 1 (2004), 188–230.

[7] Cynthia Dwork. 2007. Ask a better question, get a better answer a new approach to private data analysis. In *International Conference on Database Theory*, 18–27.

[8] Ibrahim Elhenawy, Salwa H Mahmoud, Ahmed Moustafa, et al. 2021. A Lightweight Privacy Preserving Keyword Search Over Encrypted Data in Cloud Computing. *Journal of Cybersecurity and Information Management* 3, 2 (2021), 29–9.

[9] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the Web Search and Data Mining*, 178–186.

[10] Soheila Ghane, Lars Kulik, and Kotagiri Ramamohanarao. 2018. Publishing spatial histograms under differential privacy. In *SSDBM*, 1–12.

[11] Chris Godsil and Gordon F Royle. 2001. *Algebraic graph theory*, Vol. 207.

[12] Ruining He, Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, 507–517.

[13] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *TOIS* (2004), 5–53.

[14] Thomas Hofmann. 2013. Probabilistic latent semantic analysis. *arXiv:1301.6705.*

[15] Yuan Hong, Xiaoyun He, Jaideep Vaidya, Nabil Adam, and Vijay Atluri. 2009. Effecrtive anonymization of query logs. In *CIKM*, 1465–1468.

[16] Yuan Hong, Wen Ming Liu, and Lingyu Wang. 2017. Privacy Preserving Smart Meter Streaming Against Information Leakage of Appliance Status. *TIFS* 12, 9.

[17] Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. 2015. Collaborative Search Log Sanitization: Toward Differential Privacy and Boosted Utility. *TDSC* 12, 5 (2015), 504–518.

[18] Marek Jawurek, Martin Johns, and Konrad Rieck. 2011. Smart metering de-pseudonymization. In *ACSAC*, 227–236.

[19] Noah Johnson, Joseph P Near, and Dawn Song. 2018. Towards practical differential privacy for SQL queries. *VLDB Endowment* 11, 5 (2018), 526–539.

[20] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML*, Vol. 3201, 217–226.

[21] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. 2009. Releasing search queries and clicks privately. In *WWW*, 171–180.

[22] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[23] Andrei Manolache, Florin Brad, and Elena Burceanu. 2021. DATE: Detecting Anomalies in Text via Self-Supervision of Transformers. (2021). arXiv:2104.05591

[24] Pascal Massart. 1990. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability* (1990), 1269–1283.

[25] Frank D McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, 19–30.

[26] Ilya Mironov. 2017. Rényi differential privacy. In *CSF*, IEEE, 263–275.

[27] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 111–125.

[28] Vijay B Raut and DD Londhe. 2014. Opinion mining and summarization of hotel reviews. In *Computational Intelligence and Communication Networks*, 556–559.

[29] Benjamin I. P. Rubinstein and Francesco Alda. 2017. Pain-free random differential privacy with sensitivity sampling. In *ICML*, 2950–2959.

[30] Sree Ram Valluri, David J Jeffrey, and Robert M Corless. 2000. Some applications of the Lambert W function to physics. *Can. J. Phys.* 78, (9) (2000), 823–831.

[31] Chenyang Wang, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2019. Modeling Item-Specific Temporal Dynamics of Repeat Consumption for Recommender Systems. In *WWW*, 1977–1987.

[32] Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *SIGIR*, 305–314.

[33] Shangyu Xie and Yuan Hong. 2021. Reconstruction attack on instance encoding for language understanding. In *EMNLP*, 2038–2044.

[34] Fangyuan Zhao, Xuebin Ren, Shusen Yang, Qing Han, Peng Zhao, and Xinyu Yang. 2020. Latent Dirichlet Allocation Model Training With Differential Privacy. *TIFS* 16 (2020), 1290–1305.

[35] Tianqing Zhu, Gang Li, Wanlei Zhou, Ping Xiong, and Cao Yuan. 2016. Privacy-preserving topic model for tagging recommender systems. *KAIS*, 33–58.

SUPPLEMENTARY APPENDIX

A PROOFS

A.1 Proof of Theorem 1

PROOF. Assume that the sampled sensitivities $\Delta_1, \cdots, \Delta_h$ are sorted as $\Delta_1 \leq \cdots, \leq \Delta_h$, given any $\rho' \in (0, 1)$ satisfying the following:

$$1 - \gamma + \rho + \rho' \leq 1 \Leftrightarrow \rho' \leq \gamma - \rho \tag{6}$$

Then, the random sensitivity $\Delta_s = \Delta_k$, where $h(1 - \gamma + \rho + \rho')$, is the smallest $\Delta \geq 0$ such that $\Phi_h(\Delta) \geq 1 - \gamma + \rho + \rho'$. Thus, we have:

$$\Phi_h(\Delta_s) = \frac{1}{h} \sum_{i=1}^{h} \mathbb{I}(\Delta_i \leq \Delta_s) \geq 1 - \gamma + \rho + \rho' \tag{7}$$

Define the events as

$$\begin{aligned} A_{\Delta_s} = \{\forall S \subset \mathbb{R}, Pr(\mathcal{A}_{\Delta_s}(D) \in S) \leq \\ e^{\epsilon} \cdot Pr(\mathcal{A}_{\Delta_s}(D') \in S) + \delta\} \\ B_{\rho'} = \left\{ \sup_{\Delta_s}(\Phi_h(\Delta_s) - \Phi(\Delta_s)) \leq \rho' \right\} \end{aligned}$$

where $\Phi(\Delta_s)$ is the unknown CDF and $\Phi_h(\Delta_s)$ is the corresponding random empirical CDF. The former is the event that DP holds for a specific DB pair, when the mechanism is executed with (possibly random) sensitivity parameter Δ_s; the latter records the empirical CDF uniformly one-sided approximating the CDF to level ρ'. Moreover, per the definition of differential privacy, we have

$$\forall \Delta_s > 0,\ Pr_{D,D' \sim P^{N+1}}(A_{\Delta_s}) \geq \Phi(\Delta_s). \tag{8}$$

The random D, D' on the left-hand side induce the distribution on Δ_s on the right-hand side under which $\Phi(\Delta_s) = Pr(\Delta \leq \Delta_s)$. The probability on the left-hand side is the level of random differential privacy of $\mathcal{A}_{\Delta_s}$ while running on the fixed Δ_s. By the Dvoretzky-Kiefer-Wolfowitz inequality [24], we have: for all $\rho' \geq \sqrt{(\log 2)/(2h)}$,

$$Pr_{\Delta_1, \cdots, \Delta_h}(B_{\rho'}) \geq 1 - e^{-2h\rho'^2} \tag{9}$$

Thus, we have

$$\begin{aligned} & Pr_{D,D',\Delta_1,\cdots,\Delta_h}(A_{\Delta_s}) \\ =& \mathbb{E}(\mathbb{I}[A_{\Delta_s}]|B_{\rho'})Pr(B_{\rho'}) + \mathbb{E}(\mathbb{I}[A_{\Delta_s}]|\bar{B}_{\rho'})Pr(\bar{B}_{\rho'}) \\ \geq& \mathbb{E}[\Phi_h(\Delta_s)|B_{\rho'}]Pr(B_{\rho'}) \\ \geq& \mathbb{E}[\Phi_h(\Delta_s) - \rho'|B_{\rho'}](1 - e^{-2h\rho'^2}) \\ \geq& (1 - \gamma + \rho + \rho' - \rho')(1 - e^{-2h\rho'^2}) \\ \geq& (1 - \gamma + \rho)(1 - \rho) \\ \geq& 1 - \gamma + \rho - \rho \\ =& 1 - \gamma \end{aligned} \tag{10}$$

The last inequality subjects to $\rho < \gamma$; the penultimate inequality subjects to the setting

$$\rho' \geq \sqrt{(\log \frac{1}{\rho})/(2h)} \tag{11}$$

and then the DKW condition, $\rho' \geq \sqrt{(\log 2)/(2h)}$, is satisfied given $\rho \leq 1/2$. □

A.2 Proof of Theorem 5

PROOF. In this paper, we use the Gaussian mechanism with sampled sensitivity. The Gaussian distribution has a probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{12}$$

where $\sigma_R = \sqrt{\frac{\alpha\Delta_s^2}{2\epsilon_R}}$. Then, the expectation of the amplitude of noise by Gaussian distribution is

$$\begin{aligned} \mathbb{E}(V) &= \int_{-\infty}^{\infty} |x| f(x) dx = \int_{0}^{\infty} 2x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{-2\sigma}{\sqrt{2\pi}} (\int_{0}^{\infty} de^{-\frac{x^2}{2\sigma^2}}) = \frac{2\sigma}{\sqrt{2\pi}} \end{aligned}$$

We observe that the noise is related to the sampled sensitivity that is dependent on the confidence γ. □

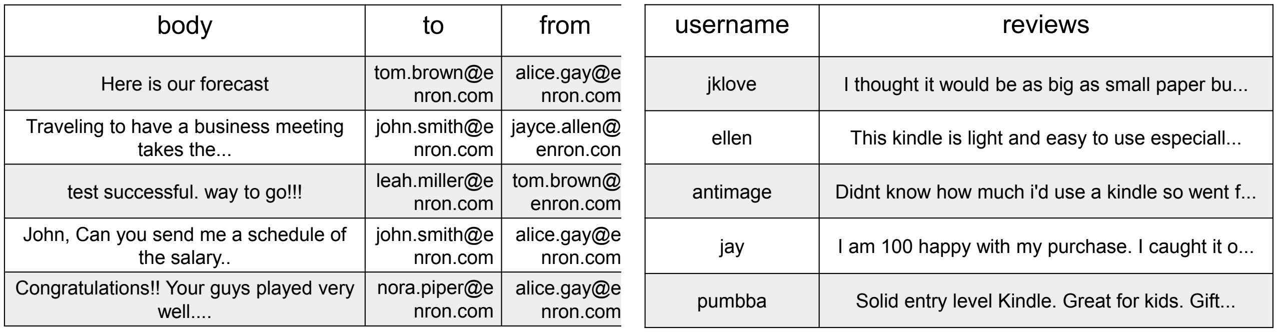
B LATENT DIRICHLET ALLOCATION

We illustrated the example topic mining model used in the experiments: Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model for collections of discrete data such as text corpora. The goal is to find short descriptions of the members in a collection that enable efficient processing of large collections while preserving the essential statistical relationships useful for the basic tasks such as classification, novelty detection, etc.

In generative probabilistic modeling, data are generated in a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. In LDA, the observed variables are the words of the documents; the hidden variables are the topic structure. We calculate the hidden topic structure from the documents by computing the posterior distribution (the conditional distribution of the hidden variables given the documents).

We can describe LDA with formal notations. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary. The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in the document d. Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary. Thus, the LDA algorithm corresponds to the joint distribution of the words:

$$\begin{aligned} & Pr(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ =& \prod_{i=1}^{k} Pr(\beta_i) \prod_{d=1}^{D} Pr(\theta_d) [\prod_{n=1}^{N} Pr(z_{d,n}|\theta_d) Pr(w_{d,n}|\beta_{1:K}, z_{d,n})] \end{aligned}$$



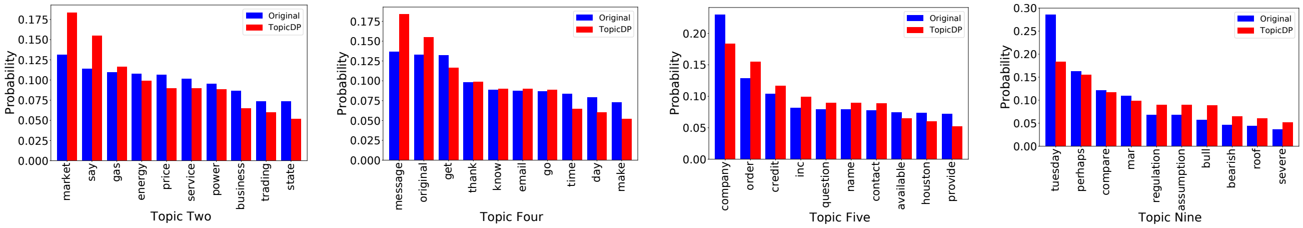
body	to	from
Here is our forecast	tom.brown@e nron.com	alice.gay@e nron.com
Traveling to have a business meeting takes the...	john.smith@e nron.com	jayce.allen@ enron.con
test successful. way to go!!!	leah.miller@e nron.com	tom.brown@ enron.com
John, Can you send me a schedule of the salary..	john.smith@e nron.com	alice.gay@e nron.com
Congratulations!! Your guys played very well....	nora.piper@e nron.com	alice.gay@e nron.com

(a) Enron Email Dataset

username	reviews
jklove	I thought it would be as big as small paper bu...
ellen	This kindle is light and easy to use especiall...
antimage	Didnt know how much i'd use a kindle so went f...
jay	I am 100 happy with my purchase. I caught it o...
pumbba	Solid entry level Kindle. Great for kids. Gift...

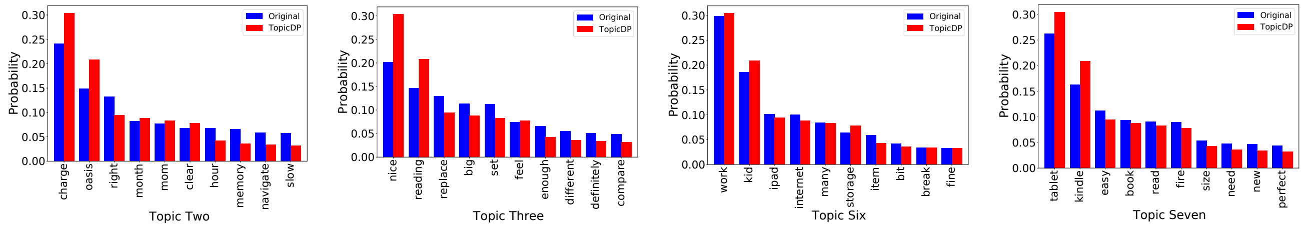
(b) Amazon Product Review Dataset

Figure 11: The Attributes of Datasets (emails and usernames are replaced with pseudonyms).



(a) Topic 2 (b) Topic 4 (c) Topic 5 (d) Topic 9

Figure 12: Keyword Distribution of Four Randomly Selected Topics in Enron Dataset.



(a) Topic 2 (b) Topic 3 (c) Topic 6 (d) Topic 7

Figure 13: Keyword Distribution of Four Randomly Selected Topics in Amazon Dataset.

The given distribution specifies various dependencies such as the topic assignment $z_{d,n}$ which in turn depends on the topic proportions per document θ_d, the observed word $w_{d,n}$ which in turn depends on the topic assignment $z_{d,n}$ and all of the topics.

C ADDITIONAL FIGURES

In this section, we present some additional figures of the experimental datasets and results. Figure 11(a) and 11(b) demonstrate the attributes of two datasets (Enron Email and Amazon Product Review datasets) used for experiments. In the Enron dataset, topic mining will be performed on the body of all the emails, each of which is considered as a separate document. In the Amazon dataset, topic mining will be performed on the specific reviews, each of which is considered as a separate document. Finally, Figure 12 and 13 show the keyword distribution of four randomly selected topics in the two datasets, respectively. The probability distributions of all the keywords are quite close in the outputs of both TopicDP and original topic mining.

D THE NOTATION TABLE

Table 2: Frequently Used Notations

Notations	Comments
W	output matrix of topic mining
h	sampling size
γ	confidence parameter
Δ_s	the smooth sensitivity
σ^2	scale parameter of Gaussian distribution
ϵ, δ	privacy parameter