Membership Inference Attacks and Privacy in Topic Modeling

A THESIS PRESENTED
BY
NICO A. MANZONELLI
TO
THE JOHN A. PAULSON SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
IN THE SUBJECT OF
DATA SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2023

Thesis advisor: Professor Salil Vadhan Nico A. Manzonelli

Membership Inference Attacks and Privacy in Topic Modeling

Abstract

From a privacy perspective, machine learning models should not leak sensitive information about their training data. Research shows that large deep learning models are especially susceptible to privacy attacks that exploit the learned model to infer aspects of the training data. However, the extent to which simpler probabilistic models, like topic models, demonstrate similar vulnerabilities is unclear. To shed light on the privacy concerns associated with topic modeling, we demonstrate membership inference attacks' effectiveness against the popular topic model Latent Dirichlet Allocation (LDA). Our results indicate that LDA shows symptoms of training data memorization despite being much simpler compared to large deep learning models. Like in other fields of machine learning, we argue that strong membership inference attacks should be used to empirically measure privacy in topic modeling.

To address their vulnerabilities, we explore differential privacy's (DP) application to topic modeling. While several DP topic modeling algorithms attempt to protect individual privacy, they largely neglect the privacy associated with the vocabulary set. We present a framework for DP topic modeling that incorporates DP vocabulary selection as a pre-processing step. Overall, our results show that DP vocabulary selection improves the privacy guarantees offered by DP and protects against membership inference attacks while having limited effects on practical utility.

Acknowledgments

First, I would like to thank my thesis advisor **Salil Vadhan**. I'm extremely grateful that Salil introduced me to the field of privacy preserving data analysis, allowed me to pursue my research interests, and offered invaluable guidance while encouraging me to consider nuances in my research.

To my committee members, **Swaroop Vattam** and **Weiwei Pan**, thank you for serving on my committee and providing valuable feedback.

A special thanks to **Wanrong Zhang** who helped me work though technical details, provided feedback, and allowed me to think out loud about my research.

This endeavor would not have been possible without funding and support from **MIT Lincoln Laboratory**. Specifically, the leadership in **Group 52** who allowed me to pursue this degree through the MIT Lincoln Laboratory Military Fellowship. I also would like to acknowledge the **MIT Lincoln Laboratory Supercomputing Center** for providing high performance computing and consultation resources that have contributed to the research and results within this thesis.

Finally, to my **family** and **friends**, specifically Joey, Jeff, Charlie, Iggy, Stone, and Maxime, thank you.

Contents

1

Introduction

In recent years, the scale of data collection by both public and private organizations has increased exponentially. While harnessing this data can provide benefits to various sectors of our society, the privacy concerns over sharing and processing our information are mounting, especially in the field of machine learning (ML). The rapid development and increasing integration of ML into everyday tasks creates more opportunities for ML models to be trained on sensitive data. This thesis explores the privacy risks associated with training and releasing ML models with a focus on topic modeling.

The privacy concerns regarding deep neural network based ML models have garnered considerable attention from the research community. Researchers actively study deep learning models' propensity to memorize training data and the consequent effect on privacy. Large language models are especially vulnerable to privacy attacks that exploit memorization. In fact, researchers at Google extracted verbatim pieces of text from GPT-2 which included personal emails, phone numbers, and other personally identifiable information (Carlini et al., 2021b). These large language models' have complex architectures with millions, if not billions, of parameters. However, it is unclear whether simpler probabilistic models, such as topic models, present similar privacy concerns.

Topic modeling is an unsupervised ML technique that aims to discover the underlying themes in a collection of text data. There are many types of topic models, but we are primarily interested in simple probabilistic topic models, specifically the foundational model Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Although deep learning models seem to dominate current ML research, probabilistic topic models remain relevant today. LDA's ease of interpretability and implementation often make it common solution for summarizing large collections of text. As evidence, companies that solicit ML as a service, like AWS, offer solutions for topic modeling based on LDA.*

Additionally, topic models are widely used in various government or national defense applications that analyze social networks and information campaigns. For instance, researchers recently applied LDA to better understand twitter activity aimed to discredit NATO during the Trident Juncture Exercises in 2018 (Uyheng et al., 2020). Therefore, considering the privacy concerns associated with LDA, particularly when used to analyze real-world scenarios of war or political unrest, is essential for promoting the ethical and responsible use of topic modeling.

From a ML perspective, researching topic models' privacy allows us to consider notions of memorization in simple probabilistic generative models. As opposed to large language models, topic models have a limited number of parameters and a Bag-of-Words (BoW) text representation. The

*https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html

BoW assumption requires that documents are represented as multisets of words where word ordering or punctuation does not matter. Furthermore, topic models require the generative process for documents to be explicitly defined *a priori* while large-language models learn the generative process based on observed patterns in the training data. Studying topic models' vulnerabilities to privacy attacks informs our understanding of memorization in ML models in general.

This research applies membership inference attacks (MIAs) to infer whether or not a specific document was used to train the learned topic model. Particularly, we extend the Likelihood Ratio Attack (LiRA) introduced by Carlini et al. to attack LDA (2021a). Our results demonstrate that the LiRA can confidently infer the membership of documents included the training data of LDA. LDA's vulnerabilities indicate that memorization may not be sequestered to large language models; topic models also exhibit some degree of memorization. Furthermore, we argue that strong attacks, like the LiRA, should be used to empirically evaluate the privacy of topic modeling like other ML models.

To address topic models' vulnerabilities we explore defenses to privacy attacks. Particularly, we focus on applying differential privacy (DP) to topic modeling. DP is a notion of individual level privacy that guarantees the output of some data analysis is indistinguishable regardless of any one user's inclusion in the analysis. While several DP topic modeling algorithms in the literature attempt to protect individual privacy, many fail to recognize the intricacies associated with DP releases of topic models.

Topic modeling primarily returns distributions over words indexed from a vocabulary set. It follows that their output is only interpretable with access to said vocabulary set. However, the vocabulary set accompanied by a set of documents inherently holds private information on the documents. Previous work on DP topic models largely disregards the privacy associated with the vocabulary set. To address this gap in the literature, we propose a framework for DP topic modeling that includes DP vocabulary selection as a pre-processing step. We show that introducing DP vocabulary selec-

tion into the topic modeling workflow strengthens the privacy grantees offered by DP and improves defenses against the LiRA while maintaining limited effects on practical utility.

1.1 Contributions

The organization and contributions of this thesis are as follows:

- Chapter 2 provides a background on topic modeling, establishes notation, and describes the technical details associated with LDA.
- Chapter 3 focuses on applying MIAs to topic models. We extend the LiRA, originally introduced by Carlini et al., to topic models, and show that the LiRA can confidently identify members of the training data in LDA (2021a).
- Chapter 4 discusses privacy preserving ML and proposes methods for defending against MIAs using DP. We introduce Fully DP Topic Modeling as a modular framework for releasing topic models, and discuss it's theoretical and empirical benefits over existing DP topic modeling procedures.

2

Background on Topic Modeling

Topic Modeling is an unsupervised machine learning technique that seeks to discover the underlying themes in unstructured data. Primarily designed for natural language processing and information retrieval, topic models attempt to summarize large collections of text data by grouping together similar words. Researchers and practitioners apply topic models in various domains to summarize documents, cluster documents, or perform down-stream tasks such as classification.

Probabilistic topic models originated from Latent Semantic Analysis (LSA), a technique that

employs Singular Value Decomposition (SVD) to identify the relationships between words in documents (Deerwester et al., 1990). LSA was the primary technique used in text mining and information retrieval until 1999 when Hofmann introduced a probabilistic variant of Latent Semantic Analysis (pLSA) (1999). Unlike LSA, Hofmann's model provided a formal probabilistic interpretation for learned topics. However, pLSA did not define a full generative model for documents. To address the flaws in LSA and pLSA, Blei et al. introduced Latent Dirichlet Allocation (LDA) as a hierarchical Bayesian model to describe the joint probability distribution over words, topics, and documents (2003). LDA set the standard for topic modeling and has remained the most widely used topic model in research and industry since its introduction in 2003. Contemporary research in topic modeling focuses on improving or extending LDA. For example, the Author-Topic Model is an extension of LDA used to characterize the relationship between topics, authors, and their documents (Rosen-Zvi et al., 2004). Additionally, there is an active research area focused on designing better topic models for short text documents (Murshed et al., 2022).

In recent years, language models have improved significantly due to complex deep neural network architectures and better text representation with embeddings. Following these themes, topic models are evolving as researchers work to incorporate word embeddings and develop "neural topic models" (Xun et al., 2017, Zhao et al., 2021c). While these developments are exciting, this thesis focuses the popular and foundational topic model LDA. We chose to focus on the privacy concerns associated with topic modeling using LDA because of its simplicity compared to large language models and other more complex topic models. First, LDA assumes a Bag-of-Words document structure which dictates that each document is simply a collection of words where multiplicity matters but word ordering, grammar, and punctuation do not. Additionally, the generative process associated with LDA relies on simple yet powerful assumptions which allow the learned model to be informative with limited parameters.

For the duration of this thesis we consider LDA, and topic models in general, under their origi-

nal use-case as a tool for text-mining. However, it is important to note that topic modeling can be applied in many other domains where the data follows a similar assumed structure. For example, researchers working with music apply topic models to understand common themes amongst songs (Shalit et al., 2013). Furthermore, researchers in bioinformatics apply topic models to a variety medical data. In one study, researchers represented genomic data from lung cancer patients as words in documents, and applied LDA to cluster different types of cancerous cells (Zhao et al., 2014).

Applications of LDA in the medical domain make privacy concerns surrounding topic modeling more immediate. For instance, privacy violations could occur if an adversary leverages results from LDA to learn sensitive information about the patients included in the data. However, before we consider the intricacies associated with privacy violations and adversarial attacks against machine learning, we first provide definitions and notation for understanding LDA as a topic model.

2.1 Definitions and Notation

To formally describe the generative model associated with LDA, we must define the definitions and notation for describing probabilistic topic models in general. Topic models that assume a Bag-of-Words (BoW) document structure, like LDA, have simple text representations built on words.

Definition 2.1.1 (Word) A *word* is a piece of discrete data indexed from a fixed vocabulary set that contains V entries.

We can represent a specific word as a V-dimensional array where the component indices correspond to term indices in the vocabulary set, and where a single component is equal to one and all others are equal to zero. To be precise, using superscripts to denote a component's index in word w, the i^{th} term in the vocabulary is a V-dimensional array w such that $w^i = 1$ and $w^j = 0 \; \forall \, i \neq j$. In a BoW document structure, words are the basic building blocks for documents.

Definition 2.1.2 (Document) A *document* is a multi-set of N words denoted as $d = \{w_1, w_2, ..., w_N\}$.

Because we assume a BoW document structure, we can represent a single document as a V-dimensional array where the i^{th} component represents the number of times the i^{th} term in the vocabulary appears in the document. This vector representation of documents highlights the assumptions made by a BoW document structure. First, multiplicity matters because each array component represents the i^{th} term's occurrence in d. Next, word ordering is not captured by this representation because each array component i corresponds to a unique word in the vocabulary set as opposed to some position relative to other words in the document. Finally, punctuation is irrelevant unless included in the vocabulary set which is not standard practice. Just like words are the building blocks for documents, documents are the base of a corpus.

Definition 2.1.3 (Corpus) A *corpus* is a set of M documents denoted by $D = \{d_1, d_2, ..., d_M\}$.

If we represent each document as a V-dimensional array, then a corpus can assume a document-term matrix structure with shape $M \times V$ where each row corresponds to a document in the corpus. In general, the objective of topic modeling is to summarize a corpus by its topics.

Definition 2.1.4 (Topic) A *topic* is a discrete probability distribution over words from a fixed vocabulary set denoted by φ such that $\sum_{i=1}^{V} \varphi_i = 1$.

We can define a corpus' *topic-word distribution* as $\Phi = \{\varphi_1, ..., \varphi_k\}$ where each φ_z is a topic distribution. The topic-word distribution for a corpus is often represented as a $k \times V$ shaped matrix where each row corresponds to a topic's distribution over words φ_z. In general, each entry of the topic-word distribution $\Phi_{z,w}$ corresponds to the probability of word w in topic z.

Thus far, we've described topic models as a general unsupervised machine learning technique used learn topics from a corpus. In this thesis, we focus on probabilistic generative models as topic models and can define them using standard machine learning notation.

Definition 2.1.5 (Topic Model) A probabilistic topic model $\mathcal{M}$ is an assumed latent structure and generative process for documents with an associated function $f_{\mathcal{M}} : [0, \infty]^{M \times V} \rightarrow [0, 1]^{k \times V}$ that maps a corpus $D \in [0, \infty]^{M \times V}$ to a topic-word distribution $\Phi \in [0, 1]^{k \times V}$.

As per machine learning convention, our definition assumes that the model's parameters, a set of latent variables defined in $\mathcal{M}$, are learned from the training data. We write $f_{\mathcal{M}}(D)$ to denote that $\mathcal{M}$ is learned on some corpus D from an underlying data distribution $\mathbb{D}$. The function $f_{\mathcal{M}}$ incorporates the learning algorithm to estimate $\mathcal{M}$'s latent variables and returns the topic-word distribution Φ.

Note that our definition relies on the assumption that topic modeling's primary goal is to estimate Φ. While $\mathcal{M}$ defines the structure of other latent variables or distributions, we'd argue that Φ is generally the most important. Intuitively, Φ best achieves the goals associated with topic modeling by summarizing the relevant themes in the corpus. Therefore, our assumption is valid for most topic modeling implementations like LDA, the primary topic model under review in thesis.

2.2 Latent Dirichlet Allocation

In 2003, Blei et al. introduced Latent Dirichlet Allocation (LDA) as a generative model for a corpus where each document contains a mixture of topics. In general, LDA assumes that each document is generated word-by-word by first selecting a word's topic assignment from the document's distribution over topics and then selecting a word from the assigned topic's distribution over words. Using the notation established in the definitions above, LDA assumes that each document d in a corpus D is generated by the following process:

1. The number of words in the document is drawn from a Poisson: $N \sim \text{Poisson}(\varsigma)$
2. The document-topic distribution is drawn from a Dirichlet: $\theta_d \sim \text{Dirichlet}(\alpha)$

3. For each word $w_n \in \{w_1, w_2, ..., w_N\}$:

   (a) Sample a topic assignment from the document's topic-distribution: $z_n \sim \text{Multinomial}(\theta_d)$

   (b) Sample a term from the given topic's topic-word distribution: $w_n \sim \text{Multinomial}(\varphi_{z_n})$

This generative process relies on a few fundamental assumptions. First, sampling the number of words in a document is independent of the latent or hidden word generating variables θ_d, z_n, and φ_z. Therefore, its randomness is generally ignored when defining probability distributions over the hidden and observed variables. Second, the dimensionality of the Dirichlet prior α controls the number of topic components in the model and is assumed to be fixed *a priori*. The parameter α is a symmetric k-dimensional vector that typically is assigned such that $\alpha < 1$ to enforce sparsity in θ_d. Intuitively, a sparse document-topic distribution θ_d implies that documents are typically generated by a limited number of topics. Because there are k topic components in the model, there are k topic-word distributions φ_z that describe LDAs latent topics. In LDA, each φ_z is a V-dimensional multinomial distribution over the vocabulary specified with a sparse Dirichlet prior.* Like the parameter α, the parameter for each topic-word distribution β is symmetric, and we assume that $\beta < 1$ to capture the idea that a only small number of words should have a high probability of occurring in each topic.

To define probability distributions over the hidden and observed variables, we must describe two more corpus-wide representations of the hidden variables in LDA. Because there are M documents in a corpus, there are M document-topic distributions θ_d that describe a corpus. In LDA, A corpus' *document-topic distribution* is defined as $\Theta = \{\theta_1, ..., \theta_M\}$ where each θ_d is a distribution over k topics. This structure is easily represented as an $M \times k$ shaped matrix where each row corresponds to document's distribution over topics and sums to one: $\sum_{i=1}^{k} \theta_{d,i} = 1 \, \forall \, d \in D$. Describing the

*In Blei et al.'s original description of LDA's generative process, they do not assume a Dirichlet prior for φ_z. However, this concept is later introduced as Dirichlet smoothing and is common to all implementations of LDA (Blei et al., 2003).

latent variable for topic assignments is a bit more nuanced because LDA assumes that documents are generated by a mixture of topics. Therefore, every word in each document is assigned a topic. We can define $Z = \{z_1, ..., z_M\}$ as the corpus-wide topic assignments where $z_d = [z_d^1, ..., z_d^N]$ are the topic assignments for word all N words in document d. Using the corpus-wide notation for the latent variables Φ, Θ and Z, the conditional distribution of the hidden variables given the observed documents is:

$$p(\Phi, \Theta, Z \mid D) = \frac{p(\Phi, \Theta, Z, D)}{p(D)} \tag{2.1}$$

Our goal is to estimate the corpus' hidden topic structure by computing the posterior in Equation 2.1. However, the posterior distribution in Equation 2.1 is intractable due to the hardness of computing the marginal probability $p(D)$. Therefore, we face a Bayesian inference problem that requires methods to efficiently approximate the posterior.

Common implementations of LDA rely on two main approaches to approximate the posterior and estimate Φ: Sampling-based approaches and variational-based approaches. Sampling-based approaches use Markov Chain Monte Carlo (MCMC) algorithms. In particular, Gibbs-sampling is a popular technique used to estimate the posterior distribution over the assignments of words to topics $p(Z|D)$ and infer Θ and Φ after the sampling period (Griffiths & Steyvers, 2004, Porteous et al., 2008). Variational Bayesian inference methods assume that the true posterior can be estimated by a simpler distribution whose parameters can be learned via optimization (Blei et al., 2003, Hoffman et al., 2010). These variational inference methods tend to be more efficient than sampling algorithms and find implementations in the popular python libraries scikit-learn and gensim.[†‡]

Regardless of the algorithm used to estimate LDAs hidden variables, as per our formal definition of topic models, the final goal of implementing LDA is to learn Φ. The topic-word distribution Φ

[†] https://scikit-learn.org
[‡] https://radimrehurek.com/gensim

captures how likely each word is to appear in each topic and is used to summarize topics by a few related words. Additionally, Φ can be used to estimate the document-topic distribution θ_d for any document d regardless of its presence in the training corpus which allows for downstream tasks like document classification or clustering. Overall, defining LDA as the generative model for documents and learning Φ is a useful unsupervised machine learning technique for understanding a corpus. However, like most machine learning implementations, LDA can present privacy concerns by leaking information about the training corpus.

3

Membership Inference Attacks Against Topic Models

In this chapter, we explore the privacy vulnerabilities associated with topic modeling. Specifically, we demonstrate strong membership inference attacks on Latent Dirichlet Allocation (LDA). Before launching into membership inference, we introduce the general themes in research on privacy attacks against machine learning models and their training data. In order to understand the pri-

vacy associated with ML models, researchers primarily study four types of attacks that exploit ML models to learn about their training data: data extraction, model inversion, property inference and membership inference (Carlini et al., 2021b, Fredrikson et al., 2015, Ganju et al., 2018, Shokri et al., 2017).

In a data extraction attack the adversary recovers specific examples from the ML model's training data. Successful data extraction constitutes a privacy violation when the data contains people's sensitive information. Additionally, these attacks present legal concerns when used to generate copyrighted text or images. Data extraction attacks have become increasingly effective as ML models have become increasingly complex and more likely to memorize their training data. For instance, as mentioned in the Chapter 1, researchers demonstrated that large language models like GPT-2 can regurgitate verbatim text from the training data (Carlini et al., 2021b). Furthermore, recent results revealed that image diffusion models are vulnerable to data extraction attacks as researchers successfully generated images from the training data (Carlini et al., 2023a).

In a model inversion attack the adversary learns general details about sub-classes of the ML model's training data. For example, a model inversion attack on a face identification model could generate many images of faces that resemble members of the training data (Fredrikson et al., 2015, Wang et al., 2022b). Unlike data extraction which generates specific examples from the training data, model inversion generates near-examples of the training data. However, these attacks still present privacy concerns because they allow the adversary to infer sensitive information about individuals in the training set. For instance, Fredrikson et al. apply model inversion on a pharmacogenetic drug dosage algorithm to discover genomic information on members of the training set (2014).

In a property inference attack the adversary learns certain global properties of the training data that aren't explicitly defined in the data. For example, adversaries can apply property inference to reveal the underlying system that created the data or the proportion of the training data that belongs

to a certain class (Ganju et al., 2018, Parisot et al., 2021). Property inference attacks raise privacy concerns when the training data set and its global properties are sensitive.

Finally, in a membership inference attack (MIA) the adversary learns if a specific observation appeared in the model's training data. Prior membership inference on ML models training data, researchers explored exploiting aggregate statistics released in genome-wide association studies (GWAS) (Dwork et al., 2015, Homer et al., 2008). These tracing attacks prompted the Foundation for the National Institutes of Health (NIH) to consider the privacy concerns associated with releasing results from GWAS, and provide inspiration for applying membership inference to ML models.

Membership inference is a fundamental attack that serves as the foundation for other more powerful attacks like data extraction. However, when the adversary learns that an individual contributed to a data set that is sensitive in nature, successful membership inference constitutes a privacy violation itself. Because MIAs are a fundamental privacy attack, researchers often use MIAs to evaluate the privacy associated with ML models. Therefore, we will consider MIAs in detail in this thesis.

3.1 Membership Inference Attacks

We define membership inference attacks against ML models using a security game introduced by Carlini et al. and inspired by Yeom et al. and Jayaraman et al. (2021a, 2021, 2018).

Definition 3.1.1 (Membership Inference Security Game $\mathsf{Exp}(\mathcal{A}, \mathcal{C}, \mathbb{D}, f_\theta, \mathcal{L})$) *Let $\mathbb{D}$ be the distribution over a data. Let f_θ be a machine learning model parameterized by θ, and $f_\theta \leftarrow \mathcal{L}(D)$ indicate that learning algorithm $\mathcal{L}$ is applied to learn θ. The membership inference security game proceeds between an adversary $\mathcal{A}$ and a challenger $\mathcal{C}$:*

1. *The challenger $\mathcal{C}$ samples a data set $D_{train} \leftarrow \mathbb{D}$ from the underlying data distribution and trains a machine learning model $f_\theta \leftarrow \mathcal{L}(D_{train})$.*

2. *$\mathcal{C}$ selects a bit $b \leftarrow \{0, 1\}$ uniformly at random.*

    - *If $b = 0$, $\mathcal{C}$ samples an observation $x \leftarrow \mathbb{D}$ such that $x \notin D_{train}$*
    - *Otherwise, $\mathcal{C}$ samples an observation $x \leftarrow D_{train}$*

3. *$\mathcal{C}$ sends x to the adversary $\mathcal{A}$ and shares query access to $\mathbb{D}$ and to the outputs of f_θ.*

4. *The adversary returns a bit $\hat{b} \leftarrow \mathcal{A}^{\mathbb{D},f}(x)$.*

5. *Output* 1 *if $b = \hat{b}$, and* 0 *otherwise.*

In the Membership Inference Security Game when $\texttt{Exp}(\mathcal{A}, \mathcal{C}, \mathbb{D}, f_\theta, \mathcal{L}) = 1$ the adversary successfully determines whether or not an observation x is in the training set of of model f_θ. Therefore, $\texttt{Exp}(\mathcal{A}, \mathcal{C}, \mathbb{D}, f_\theta, \mathcal{L})$ successfully defines a MIA.

The membership inference game assumes that adversary is limited to query access on the model. This assumption implies that the MIA is black-box attack meaning that the adversary can only observe model predictions. Some MIAs assume white-box adversarial knowledge which implies that the adversary has access to the model, the model's architecture, learned parameters, and any hyperparameters (Nasr et al., 2019). For instance, Hayes et al. and Hilprecht et al. explore white-box attacks on generative adversarial networks (GANs) (2019, 2019). However, researchers primarily focus black-box attacks because they require less adversarial knowledge and apply to real-world model deployments. When the model is designed for classification, the strictest black-box attacks assume that adversary only has access to the model's predicted label while most others assume that adversary has access to the predicted confidence scores for each class (Li & Zhang, 2021).

The defined membership inference game often assumes that the adversary has access to the underlying data distribution $\mathbb{D}$. We note that not all MIAs make this assumption. For example, Yeom et al. introduced the LOSS attack which exploits the simple fact that observations in the training data typically return lower values when evaluated using the model's loss function (2018). However,

many MIAs rely on the assumption that the adversary can access the underlying data distribution to train "shadow models" (Shokri et al., 2017). In a shadow model based attack, the adversary trains many separate shadow models that imitate the target model using data from $\mathbb{D}$. Then, by examining the difference between the shadow models' behavior and the target model's behavior on specific observations, the adversary can infer the information on the observations' membership in the training set of the target model. Attacks that leverage training shadow models tend to be the most popular against ML models (Hu et al., 2022, Long et al., 2020, Sablayrolles et al., 2019).

While most MIAs exploit deep learning models, some studies investigate classical ML models, or propose model-agnostic methods for membership inference (Ruiz de Arcaute et al., 2022, Salem et al., 2019). In the realm of probabilistic topic modeling, there are no studies dedicated to evaluating or understanding MIAs. However, researchers have proposed a few attacks against topic models to validate their privacy-preserving data analysis.

Zhao et al. develop a privacy preserving gibbs-sampler for LDA, and use a topic-based attack to evaluate their privatized model's defense (2021a). However, their topic-based attack operates at the word-level and fails to identify membership for a document. Furthermore, the authors circumvent designing an MIA for topic models by suggesting an ad-hoc approach: exploit an LDA-based classification model model. The details on this MIA are sparing, but we assume that LDA aids in selecting features for a separate ML model. Nonetheless, this ad-hoc attack may address the privacy leakage related to the separate classification model, rather than the topic model.

Unlike Zhao et al., Huang et al. propose a MIA for specifically for topic models to evaluate their private variant of LDA (2021a, 2022). Their attack uses the topic-word distribution Φ to predict a target document's topic distribution $\hat{\theta}_d$. In line with Salem et al., the authors note that the maximum posterior estimate and variance on $\hat{\theta}_d$ is always larger for members of the training set, and the entropy is always smaller for members of the training set (2019). To exploit their observation, they directly threshold the maximum topic-probability, standard deviation, and entropy to predict

membership in three separate attacks. If the maximum topic-probability or standard deviation on a target document's $\hat{\theta}_d$ is larger than the threshold on the maximum posterior or standard deviation criterion, they conclude that the target document was part of the training data. Conversely, if the entropy on a target document's $\hat{\theta}_d$ is smaller than the threshold on the entropy criterion, they conclude that the target was part of the training data.

Huang et al. use their MIA to evaluate their privatized model's defensive mechanism, but the technicalities and underlying explanations for effective membership inference on LDA are not fully explored (2022). Furthermore the authors evaluate attack performance using precision and recall. We argue that alternative evaluation methods are better suited to assess MIA as privacy violations.

3.1.1 Evaluating Membership Inference Attacks

At its core, an MIA is a binary classification technique, and researchers initially used standard metrics such as balanced accuracy, precision, recall and area under the receiver operating curve (ROC-AUC) scores to measure their effectiveness. However, recent studies argue that these traditional classification metrics are insufficient in analyzing MIA attack efficacy (Carlini et al., 2021a). The primary issue with average case metrics, like balanced accuracy, is that they only reveal attack behavior on average, and they fail to reveal if the attack can confidently infer membership for observations.

Carlini et al. illustrate the flaw in average-case metrics is by considering two attacks as in (2021a). The first attack perfectly infers membership for 0.1% observations, but has a 50% chance of success for all other observations. The other attack has a 50.05% probability of successfully inferring membership for all observations. Although both attacks have the same balanced accuracy, the first attack raises privacy concerns by confidently inferring membership for a few observations, and the second attack is almost completely ineffective. Therefore, we should analyze the true positive rate (TPR) at low false positive rates (FPR) to assess an attacks propensity to confidently infer membership for

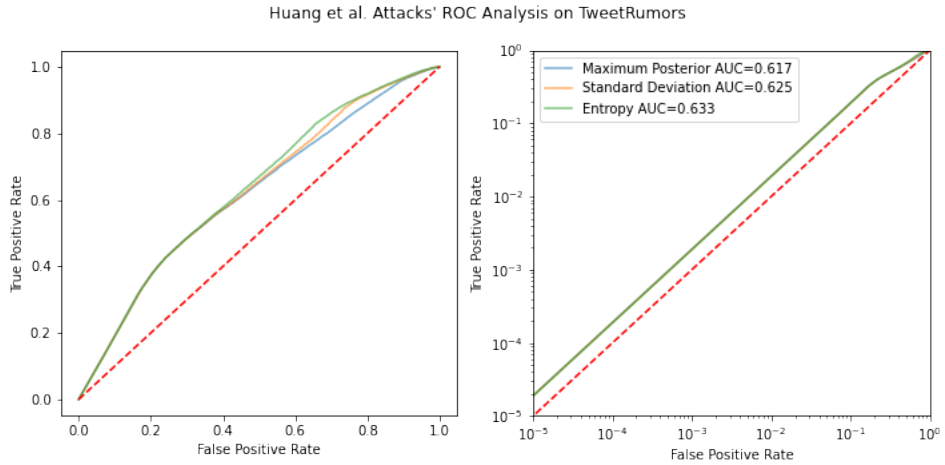

Figure 3.1: ROC-AUC Analysis for Attacks that Perform Poorly at Low FPRs

observations.

While we are interested at examining the TPR at low false positive regimes, we must be wary of traditional ROC and area under the curve (AUC) analysis. The ROC curve mostly high contains FPRs (above 10%), and we are mostly concerned with FPRs well below 10%. Therefore, visualizing ROC curves on their typical axises does not capture the TPRs at the FPRs we are interested in. For the same reason, AUC score inadequately summarizes attack performance. Consider figure 3.1. While the ROC curve for Huang et al.'s attacks on the left suggest that each attack performs fairly well most FPRs, we notice that the attack performs exceptionally bad at low FPRs (linearly decreasing) when we log scale the axises for the same ROC curves on the right Huang et al.. Therefore, when we use ROC analysis to report attack performance, we apply a log scale to the axises to better visualize the TPR at low FPRs.

MIAs on various deep learning models boast impressive performance at low FPRs. However, membership inference against topic models remains largely unexplored, especially when evaluated at low false regimes. Therefore, in the subsequent sections we propose a MIA extending the Likelihood Ratio Attack the against Latent Dirichlet Allocation (LDA) and evaluate attack performance

by analyzing the TPR at low FPRs (Carlini et al., 2021a).

3.2 The Likelihood Ratio Attack Against Topic Models

The Likelihood Ratio Attack (LiRA) is the state-of-the-art MIA that leverages training shadow models with the power of hypothesis testing to conduct membership inference (Carlini et al., 2021a). Researchers designed the LiRA to achieve high TPRs at low FPRs by leveraging the statistical power of the likelihood ratio test under the Neyman-Pearson lemma (Carlini et al., 2021a, Neyman & Pearson, 1933).* Therefore, we propose an extension of the LiRA to topic models. First, we define the adversarial assumptions underlying the LiRA against topic models. Then, we detail our extension of the LiRA and provide general intuition on MIA's efficacy against topic models. Finally, we show that our proposed attack significantly outperforms existing MIAs on topic models, and discuss the broader significance of our findings.

3.2.1 Adversarial Scenario

The adversarial assumptions accompanied with the LiRA against topic models follow directly from Definition 3.1.1: the Membership Inference Security Game. First, we assume that the adversary is given query access to the underlying data distribution $\mathbb{D}$. This assumption allows the adversary to train shadow topic models on data sets drawn from $\mathbb{D}$, and relates to a secondary assumption; the adversary has the requisite compute resources and knowledge to reliably train shadow topic models.

We also assume that the adversary is only given access to the output of the topic model Φ. For other attacks against ML models, we typically say that the adversary is given query access to the trained model f_θ to retrieve a prediction $\hat{y}$ on an observation x: $\hat{y} \leftarrow f_\theta(x)$. In the black-box set-

*The Neyman-Pearson lemma states that for a fixed significance level (FPR), the most powerful test (achieves the highest TPR) is obtained by comparing the likelihood ratio of the two hypothesises to a threshold

ting, we assume that the adversary can not observe intermediate training steps or the model's learned parameters θ directly. Our assumption for topic modeling is the same. However, instead of 'query access' to retrieve 'predictions', note that we simply state that the adversary is given access to the outputs of f_θ in step 3 of the Membership Inference Security Game. This extends nicely to our definition of topic models where we say that the topic model maps a corpus to topic-word distribution: $f_{\mathcal{M}} : D \rightarrow \Phi$. Therefore, under our black-box assumption, the adversary has access to the learned topic-word distribution Φ, but not other latent variables captured by $\mathcal{M}$.

To make our black-box assumption more concrete, let's consider Latent Dirichlet Allocation (LDA). In attack against LDA, we assume the adversary has access to Φ, but not the other latent variables (the topic-word assignments in Z and the document-topic distributions in Θ). Additionally, the adversary can not observe intermediate word count information or hyper-parameters used to learn LDAs latent variables. In fact, the adversary can not be certain that the assumed generative process follows LDA because they can not observe $\mathcal{M}$. Therefore, the adversary is left only with Φ and must use Φ to query some statistic from the learned topic model.

Hypothetically, we could consider a 'query' on a topic model as a request to predict the document-topic distribution $\hat{\theta}_d$ on an unseen document d such that $\hat{\theta}_d \leftarrow f_{\mathcal{M}}(d)$. However, $\hat{\theta}_d$ would hold no meaning without access to Φ because Φ summarizes each topic as a distribution over the vocabulary set. Furthermore, a clever adversary could 'query' $f_{\mathcal{M}}$ with various specially selected documents to reconstruct Φ. Therefore, we remove this level of abstraction and assume that the adversary is given direct access to Φ as the output of our topic model.

Consider the following theoretical scenario that summarizes our adversarial assumptions and highlights the potential privacy concerns associated with MIAs on topic models. Hospital XYZ regularly sees many patients with Condition X. In order to understand the social media behavior of patients with Condition X, hospital XYZ conducts a study by performing topic modeling on a set of tweets from patients with Condition X. Although the hospital does not reveal the tweets directly,

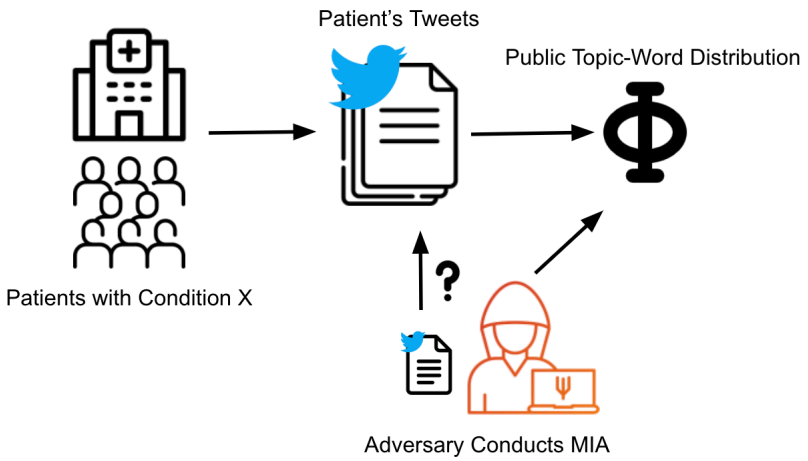

Figure 3.2: Theoretical Privacy Concern

they publicly release their code and learned topic-word distribution with the results of their study.

In the Hospital XYZ scenario captured in Figure 3.2, a potential adversary with only access to Φ and an understanding that Hospital XYZ used topic modeling on a set of tweets from patients with Condition X can conduct a MIA. Using the assumption that the adversary has knowledge on the underlying data distribution $\mathbb{D}$, which is valid because they understand that the data came from twitter, the adversary can collect auxiliary data from $\mathbb{D}$, train shadow topic models, and conduct a LiRA against Φ. If the adversary can confidently infer membership for a tweet, they learn that the author of the tweet has been seen at Hospital XYZ for Condition X which poses direct privacy violation.

3.2.2 A Likelihood Ratio Test on Topic Models

In a LiRA against topic models the adversary performs a hypothesis test to determine whether or not Φ was learned using a document d. By conducting the hypothesis test the adversary must differentiate between two distributions of learned shadow topic models. Formally, we consider $\mathbf{T}_{in}(d) = \{\Phi \leftarrow f_{\mathcal{M}}(D \cup \{d\}) | D \in \mathbb{D}\}$ and $\mathbf{T}_{out}(d) = \{\Phi \leftarrow f_{\mathcal{M}}(D \backslash \{d\}) | D \in \mathbb{D}\}$ where

$\mathbf{T}_{in}(d)$ is the distribution of Φ learned with the target document d, and $\mathbf{T}_{out}(d)$ is the distribution of Φ learned without the target document d. Given an observed Φ_{obs}, the adversary performs a hypothesis test to determine if Φ_{obs} was sampled from $\mathbf{T}_{in}(d)$ or $\mathbf{T}_{out}(d)$. Like Carlini et al., we frame this problem using the Likelihood Ratio Test with the test statistic $\Lambda(\Phi, d)$:

$$\Lambda(\Phi, d) = \frac{p(\Phi_{obs}|\mathbf{T}_{in}(d))}{p(\Phi_{obs}|\mathbf{T}_{out}(d))}, \tag{3.1}$$

where $p(\Phi_{obs}|\mathbf{T}_x(d))$ is the probability density function over Φ_{obs} under the fixed distribution of shadow topic models $\mathbf{T}_x(d)$ (2021a). However, the test in Equation 3.1 is intractable because the adversary can not observe the generative process for Φ. Therefore, the adversary must select an informative one-dimensional statistic that can be efficiently computed on Φ to perform the hypothesis test.

Instead of calculating $p(\Phi_{obs}|\mathbf{T}_x(d))$ directly, we query a statistic $\zeta : (\Phi, d) \rightarrow \mathbb{R}$ on all Φ to estimate $p(\Phi_{obs}|\mathbf{T}_x(d))$. Under the simplified test, we define $\tilde{\mathbf{T}}_x(d)$ as the distribution of statistics queried on shadow Φ with document d such that $\tilde{\mathbf{T}}_{in}(d) = \{\zeta(\Phi, d) \mid \Phi \leftarrow f_{\mathcal{M}}(D \cup \{d\}), D \leftarrow \mathbb{D}\}$ and $\tilde{\mathbf{T}}_{out}(d) = \{\zeta(\Phi, d) \mid \Phi \leftarrow f_{\mathcal{M}}(D \backslash \{d\}), D \leftarrow \mathbb{D}\}$. The test can proceeded by calculating the probability density of ζ queried on Φ_{obs} under the estimated distribution $\tilde{\mathbf{T}}_x(d)$: $p(\zeta(\Phi_{obs}, d) \mid \tilde{\mathbf{T}}_x(d))$.

Under the supervised machine learning setting, researchers can use the loss or a logit-scaled prediction score as their statistic (Carlini et al., 2021a). In language modeling, they opt for natural likelihood estimates such as perplexity (Carlini et al., 2021b). To apply the likelihood ratio test to Φ, we explored many alternatives for statistics queried on Φ, and selected the best based on a standard set of criteria and metrics to compare distributions (i.e. KL divergence). Appendix A contains an in-depth analysis of our selection criteria and analysis.

Our experiments indicate that a heuristic for the target document's maximum log-likelihood un-

der the topic model Φ is an effective statistic for conducting the a likelihood ratio test on topic models. To efficiently calculate said statistic for an unseen document, the adversary assumes an LDA-like generative process for documents, and directly maximizes the log-likelihood of the document over all fixed document-topic distributions such that:

$$\zeta(\Phi, d) = \max_{\theta} \log p(d|\theta, \Phi) = \max_{\theta} \log(\prod_{w \in d} \sum_{z} \theta_z * \Phi_{z,w}) = \max_{\theta} \sum_{w \in d} \log(\sum_{z} \theta_z * \Phi_{z,w}), \quad (3.2)$$

where θ is a document-topic distribution with constraints $\theta \in [0,1]^k$ and $\sum_z \theta_z = 1$. We estimate ζ directly by applying an out-of-the-box optimization method to minimize $-\log p(d|\theta, \Phi)$ (Gao & Han, 2012, Virtanen et al., 2020).

3.2.3 On Memorization and Per-Example Hardness

Understanding and explaining memorization is a popular trend in research on neural ML models. Generally, we see that these models seemingly memorize their training data and labels (Feldman, 2021, Feldman & Zhang, 2020). In language modeling, memorization refers to the model's ability to return exact sequences of text from the training data (Carlini et al., 2023b, 2019). Clearly, memorization presents an issue for training data privacy, and MIAs are inherently tied to a models propensity to memorize their training data.

In order to visualize our statistic's behavior and verify aspects of memorization and per-example hardness for probabilistic topic models, we replicate an experiment conducted on neural-networks by Carlini et al. and Feldman & Zhang using LDA (2021a, 2020). Their tests show that outlying observations tend to have a greater effect on the learned model than other observations, like inliers, when included in the training set of a neural-network. We show similar results by conducting the following experiment. First, we sample half of a data set $D_{train} \subset D$ and learn Φ using LDA such

that $\Phi \leftarrow f_{\mathcal{M}}(D_{train})$. Then, we calculate $\zeta(\Phi, d)$ for target documents d and note if $d \in D_{train}$. We repeat the experiment many times to empirically estimate $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$.

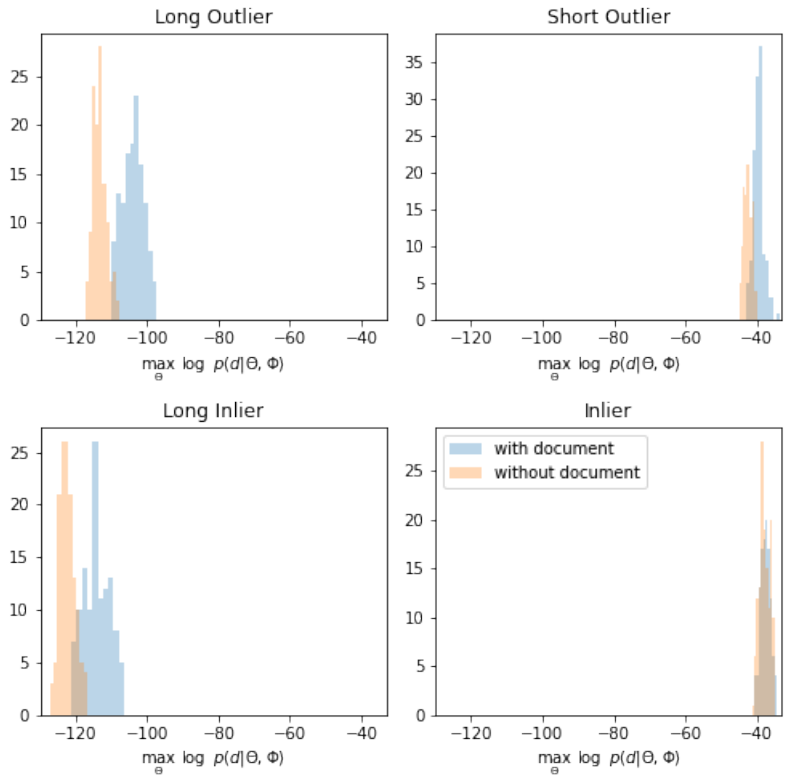

Figure 3.3: Statistic $\zeta(\Phi, d)$ Evaluated on Different Types of Documents

Figure 3.3 shows the results for the experiment by plotting the histograms of $\zeta(\Phi, d)$ on various documents when $d \in D_{train}$ (blue) and when $d \notin D_{train}$ (orange). We define an outlier as a document that contains many words that appear infrequently in D, and the length ("long" or "short") relates to the number of words in the document. For example, if we let C_d represent the word count for document d and C represent a set word counts for each $d \in D$, the word count for "long" documents satisfies $C_d \geq \text{mean}(C) + \sigma(C)$ and the word count for "short" documents satisfies $C_d \leq \text{mean}(C) - \sigma(C)$. The document word count labeled "inlier" satisfies $C_d \leq \text{mean}(C) \pm \sigma(C)$.

Our results in figure 3.3 reveal that although probabilistic topic models do not learn like neural-networks, under the queried statistic ζ they exhibit similar behavior in terms of memorization and per-example hardness. When included in the training data for Φ, longer documents significantly increase $\zeta(\Phi, d)$. We also notice that an outlying document affects $\zeta(\Phi, d)$ similarly. However, instead of relying on the document length to skew $\zeta(\Phi, d)$, the effect comes from infrequently used words. For these documents, we observe that the word probability $\Phi_{z,w}$ in certain topics dramatically changes for words that appear infrequently in D or words that occur many times in d. Collectively, these factors affect the learned topic model Φ and shift $\zeta(\Phi, d)$.

Although we hesitate to claim broadly that probabilistic topic models memorize their training data, the simple fact that the model is more likely to generate specific documents after they are incorporated in the training data is similar to memorization. However, because the generative process for documents in LDA is extremely simple compared to language models, they do not learn verbatim sequences of words. Instead, topic model's ability to "memorize" the training data is due to learning word co-occurrences.

Topic models learn their topics based on word co-occurrence patterns in documents which equates releasing Φ to providing a high dimensional query on word co-occurrence in a corpus. Unlike language models, which predict the next word in a sequence, topic models generate words based on a distribution of topics. For instance, a language model trained on many examples of "Mary had a little lamb" has an extremely high probability of predicting "lamb" given the starting sequence "Mary had a little." A topic model trained on many examples of "Mary had a little lamb" has a high probability of generating "lamb" given Φ, but that probability could also be equal to the probability of predicting "Mary" or "little." Nonetheless, the likelihood of observing the document "Mary had a little lamb" increases in either case which indicates some degree of memorization.

We note that longer documents have lower values for $\zeta(\Phi, d)$ and are less likely to be generated by the learned model. Because the log-likelihood for a document d is the sum of independent log-

probabilities for generating each word in d, we expect the magnitude of $\zeta(\Phi, d)$ to be higher for longer documents. Our experiments show that long documents contain lower $\zeta(\Phi, d)$ scores. This confirms our intuition for the magnitude of $\zeta(\Phi, d)$ and may suggests that the model struggles to capture a coherent topic structure for longer documents. We speculate that longer documents tend to have a more diverse vocabulary, making it harder to fit a meaningful document topic distribution under Φ. This notion helps explain why including longer documents in the learned model increases $\zeta(\Phi, d)$ because the model now fits the longer documents' word co-occurrence better. Furthermore, observing lower $\zeta(\Phi, d)$ for longer documents also suggests per-example hardness, i.e. some observations are inherently more difficult to fit than others.

The Likelihood-Ratio test exploits per-example hardness by distinguishing between the estimated distributions $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$ for each document d. Modeling separate distributions gives the LiRA the benefit of accounting for the natural differences in the test statistic on various documents. However, an attack with global thresholds can not differentiate between true non-members and hard-to-fit documents or true members and easy-to-fit documents. As noted by Carlini et al., the attack can only confidently identify non-members of the training data using global thresholds (2021a). Huang et al.'s use of global thresholds in their MIA significantly hinders attack performance at low false positive regimes (2022). Therefore, we propose the online and offline LiRA against topic models as a stronger alternative.

3.2.4 Online Likelihood Ratio Attack

The LiRA on topic models follows directly from the LiRA on supervised ML models proposed by Carlini et al. (2021a). However, the attack on topic models differs because we use statistic ζ queried on Φ. Other than using ζ on Φ, the process for conducting a LiRA on a topic model is essentially the same on supervised ML models.

First, the adversary trains $2N$ shadow topic models on an auxiliary data set D drawn from the

underlying data distribution $\mathbb{D}$. By planting d into the train set of N shadow models, the adversary ensures that half of the shadow topic models are learned using the target document d. We call the shadow topic models learned with or without the target document Φ_{in} and Φ_{out} respectively. The adversary queries $\zeta(\Phi, d)$ for all shadow Φ, and estimates a normal distribution for the *in* and *out* sets of queried statistics in $\tilde{\mathbf{T}}_{in}$ and $\tilde{\mathbf{T}}_{out}$. Finally, they query ζ on the target topic model's topic-word distribution Φ_{obs} and return the estimated Likelihood-ratio test statistic. Algorithm 1 provides the pseudocode for the online LiRA against topic models.

Algorithm 1 Online LiRA on Topic Models

1: **Inputs:** Target topic-word distribution: Φ_{obs}, target document: d, shadow model iterations: N, statistic function: ζ, data distribution: $\mathbb{D}$, and topic model $f_{\mathcal{M}}$.
2: **Returns** Likelihood Ratio Test Statistic: Λ
3: $\tilde{\mathbf{T}}_{in} = \{\}$ ▷ Initialize empty sets
4: $\tilde{\mathbf{T}}_{out} = \{\}$
5: **for** N times **do**
6: $\quad D \leftarrow \mathbb{D}$ ▷ Sample auxiliary data from distribution
7: $\quad \Phi_{in} \leftarrow f_{\mathcal{M}}(D \cup \{d\})$ ▷ Train shadow model with target document
8: $\quad \tilde{\mathbf{T}}_{in} \leftarrow \tilde{\mathbf{T}}_{in} \cup \{\zeta(\Phi_{in}, d)\}$
9: $\quad \Phi_{out} \leftarrow f_{\mathcal{M}}(D \backslash \{d\})$ ▷ Train shadow model without target document
10: $\quad \tilde{\mathbf{T}}_{out} \leftarrow \tilde{\mathbf{T}}_{out} \cup \{\zeta(\Phi_{out}, d)\}$
11: **end for**
12: $\mathcal{N}_{in} \leftarrow \mathcal{N}(\text{mean}(\tilde{\mathbf{T}}_{in}), \text{var}(\tilde{\mathbf{T}}_{in}))$ ▷ Estimate normals
13: $\mathcal{N}_{out} \leftarrow \mathcal{N}(\text{mean}(\tilde{\mathbf{T}}_{out}), \text{var}(\tilde{\mathbf{T}}_{out}))$
14: $\Lambda \leftarrow \frac{p(\zeta(\Phi_{obs}, d) \mid \mathcal{N}_{in})}{p(\zeta(\Phi_{obs}, d) \mid \mathcal{N}_{out})}$ ▷ Calculate test statistic
15: **Return** Λ

The LiRA leverages parametric modeling by estimating Gaussian's for $\tilde{\mathbf{T}}_{in}$ and $\tilde{\mathbf{T}}_{out}$ in lines 12 and 13. If the query statistic follows a known distribution, parametric modeling allows the attack to achieve similar performance as directly modeling $\tilde{\mathbf{T}}_{in}$ and $\tilde{\mathbf{T}}_{out}$ while training fewer shadow topic models. The query statistic ζ appears to be well approximated by a Gaussian (note the distributions in Figure 3.3). Therefore, we can use parametric modeling allows for simpler, more efficient estima-

tion of the Likelihood Ratio test statistic Λ.

Technically, the Likelihood-ratio test statistic Λ is the test statistic under the null hypothesis that the document is a member of the training set. However, the ordering of the null and alternative hypothesis does not matter for Λ in the online attack because we are simply comparing the ratio of the probability density at $\zeta(\Phi_{obs}, d)$. The purpose of including $\mathcal{N}_{in}$ in the numerator of Λ is to provide ease of interpretation; if $\zeta(\Phi_{obs}, d)$ is more likely to be sampled from $\mathcal{N}_{in}$ than $\mathcal{N}_{out}$, then Λ is larger than 1. In fact, as Λ approaches ∞ it becomes more likely that the document was sampled from $\mathcal{N}_{in}$ and is a member of the training set. Therefore, we can interpret $\Lambda \in [0, \infty]$ directly as a membership confidence score or threshold Λ at a desired significance level to determine if we reject or fail to reject the null hypothesis for a single document.

3.2.5 Offline Likelihood Ratio Attack

In the online attack, the adversary must train N shadow topic models each time they want infer the membership of document d. This process is computationally expensive, and poses usability limitation because it requires that the adversary trains the shadow topic models after deciding to test the membership of d. To ease the burden of training N new shadow topic models, Carlini et al. propose the offline LiRA. In the offline LiRA the adversary does not train shadow topic models on data sets with the target document. Algorithm 2 provides the pseudocode of the offline LiRA against topic models.

Because we only train shadow topic models on data sets without the target document, we can evaluate new target documents after learning a collection of N shadow topic models once by comparing $\zeta(\Phi_{obs}, d)$ to $\mathcal{N}_{out}$ for any d. This means that lines 3-9 can be executed regardless of d, and lines 11-17 can be repeated for any d after estimating $\mathbf{T}_{out}$ once. In other words, we do not need to train N shadow models for each document we want to evaluate.

In the LiRA's offline version, the likelihood ratio test in line 17 is switched to a one-sided hy-

Algorithm 2 Offline LiRA on Topic Models

Inputs: Target topic-word distribution: Φ_{obs}, target document: d, shadow model iterations: N, statistic function: ζ, data distribution: $\mathbb{D}$, and topic model $f_{\mathcal{M}}$.
Returns Likelihood Ratio Test Statistic: Λ
$\mathbf{T}_{out} = \{\}$ ▷ Initialize empty set
for N times **do**
    $D \leftarrow \mathbb{D}$ ▷ Sample auxiliary data from distribution
    $\Phi_{out} \leftarrow f_{\mathcal{M}}(D)$ ▷ Train shadow models without target document
    $\mathbf{T}_{out} \leftarrow \mathbf{T}_{out} \cup \{\Phi_{out}\}$
end for
$\tilde{\mathbf{T}}_{out} = \{\}$ ▷ Calculate Statistic Across $\mathbf{T}_{out}$
for Φ_{out} in $\mathbf{T}_{out} = \{\}$ **do**
    $\tilde{\mathbf{T}}_{out} \leftarrow \tilde{\mathbf{T}}_{out} \cup \{\zeta(\Phi_{out}, d)\}$
end for
$\mathcal{N}_{out} \leftarrow \mathcal{N}(\text{mean}(\tilde{\mathbf{T}}_{out}), \text{var}(\tilde{\mathbf{T}}_{out}))$ ▷ Calculate Test Statistic
$\Lambda \leftarrow 1 - \Pr[\mathcal{N}_{out} > \zeta(\Phi_{obs}, d)]$
Return Λ

pothesis test under the null hypothesis that the target document was not used to learn Φ_{obs}. In this case, Λ is 1 minus the survival function of $\mathcal{N}_{out}$ at $\zeta(\Phi_{obs}, d)$ which is equivalent to the cumulative density function. Therefore, we are estimating the probability that a random $\zeta(\Phi_{out}, d)$ draw from $\mathcal{N}_{out}$ will be as high as $\zeta(\Phi_{obs}, d)$. As the value of $\zeta(\Phi_{obs}, d)$ gets larger compared to the mean of $\tilde{\mathbf{T}}_{out}$, then Λ approaches 1 and we are more likely to reject the null hypothesis that the document is a non-member. Therefore, we can interpret $\Lambda \in [0, 1]$ as a membership confidence score.

3.3 Attack Evaluation

In this section, we evaluate LiRA against Latent Dirichlet Allocation (LDA). Although the LiRA operates like a hypothesis test, we do not treat our test statistic Λ like a standard test statistic that we wish to threshold for a desired significance level. Instead, because Λ is a ratio of probabilities in the online LiRA, thresholds on Λ at a given significance level are normalized for tests across each

document. Consequently, we can interpret Λ as a predicted membership score where higher values indicate that the document is more likely to be a member of the training data. In the offline LiRA, where Λ is a direct probability estimate, the same interpretation applies. As discussed in Section 3.1, we are interested in evaluating attack performance at low false positive regimes. Therefore, we estimate empirically the attack's TPR at low FPRs and plot the attack's ROC curve on log scaled axes.

3.3.1 Data

We evaluate our attacks against three data sets from various sources with different sizes, document lengths, vocabulary lengths, and optimal number of topics:

- *TweetRumors*[†] contains a collection of 5,802 tweets posted on five major news events. Zubiaga et al. originally collected the data to detect unverified information (i.e. rumors) online (2016). However, the data set extends nicely for topic modeling with limited topics because of the five distinct news events. We combine the labeled rumors and non-rumors into one large corpus for topic modeling.

- *20Newsgroup*[‡] is a collection of approximately 18,000 newsgroup documents across 20 different newsgroups. Though the author does not explicitly mention data collection, introduced the original data set. Numerous research studies and competitions use the data set for topic modeling, and it continues to be a popular choice for topic modeling research.

- *NIPS*[§] contains 1,500 research papers from Neural Information Processing Systems (also referred to as NeuralIPS). There are various NeuralIPS datasets available online, but we opt

[†]Data available at https://www.zubiaga.org/datasets
[‡]Data available at http://qwone.com/~jason/20Newsgroups
[§]Data available at https://archive.ics.uci.edu/ml/datasets/bag+of+words

Data Set	Number of Documents M	Average Document Length	Vocabulary Size V
TweetRumors	5,698	9	5,942
NIPS	1,494	893	10,346
20Newsgroup	18,037	84	74,781

Table 3.1: Data Set Profile After Pre-Processing

to the use bag-of-words corpus available on the UCI Machine Learning Repository due its popular use in topic modeling research (Dua & Graff, 2017). For example, Huang et al. use *NIPS* to evaluate their MIA and corresponding defense (2022).

We apply the following standard text pre-processing procedures for each data set: removal of all non-alphabetic characters, tokenization, removal of stop words, removal of tokens longer than 15 characters, removal of tokens shorter than 3 characters, and lemmentization.[¶]

After pre-processing we transform each corpus into it's corresponding matrix representation such that D is an $M \times V$ matrix. Each entry $D_{m,v}$ represents the count of the v^{th} term in the vocabulary set in the m^{th} document in the corpus. Table 3.1 provides the basic data profile for each data set after pre-processing.

3.3.2 Attack Simulations and Set-Up

We conduct the following experiments across each data set to evaluate the LiRA on topic models. We replicate Carlini et al. first set of experiments by randomly sampling half of the data set, training LDA and release Φ_{obs} (2021a). The adversary proceeds by conducting an online LiRA against the target document using Φ_{obs} by randomly sampling half of each data set to train shadow models. With this framework, we can evaluate membership for each document by repeatedly sampling half of the data set to train shadow models and ensuring that each document appears in exactly N

[¶]Tokenization, stop words, and lemmentization implemented via python package nltk https://www.nltk.org

shadow model training subset. Note that the adversaries auxiliary data and the target model's training data likely overlap. This is a strong assumption made to accommodate the smaller size of our data sets. As observed by Carlini et al., we do not expect attack performance to drop using disjoint data sets (2021a).

We compare the LiRA directly to Huang et al.'s attack in our evaluations (2022). We replicate their attacks by randomly sampling half of the data set, training LDA and release Φ_{obs}. Then, using Φ_{obs} we estimate each document's topic-distribution $\hat{\theta}$ for every document, compute the maximum posterior, standard deviation, and entropy on $\hat{\theta}$ (2022). We directly threshold the maximum posterior, standard deviation, and entropy to evaluate their attack for every document in each data set.

All of the proposed simulations ran on MIT Lincoln Laboratory Supercomputing Center's system TX-Green (Reuther et al., 2018). We implemented the algorithms and simulations in Python version 3.9. For each experiment, we learn Φ using sklearn's implementation of LDA with default learning parameters.[‖] For *TweetSet* and *20Newsgroup* we select the number of topics k in accordance with the underlying structure of the data: 5 and 20 respectively. For *NIPS* we vary k to study the efficacy of the attack at various topic numbers. Additionally, for the first set of experiments we test each target document separately which implies that we train a new N shadow topic models for each target document in the offline algorithm. Although not required under the offline LIRA, we opt to enforce separate tests in the offline variant to run online an offline attacks with overlapping data sets. We replicate each experiment 10 times and report our results across all iterations.[**]

3.3.3 Results

Overall, our results indicate that the LiRA is effective at inferring membership on LDA at low false positive regimes. First, we conduct the initial experiment on *TweetSet* while varying N to under-

[‖] https://scikit-learn.org/

[**] Code for all experiments and evaluations is available at https://github.com/nicomanzonelli/topic_model_attacks.

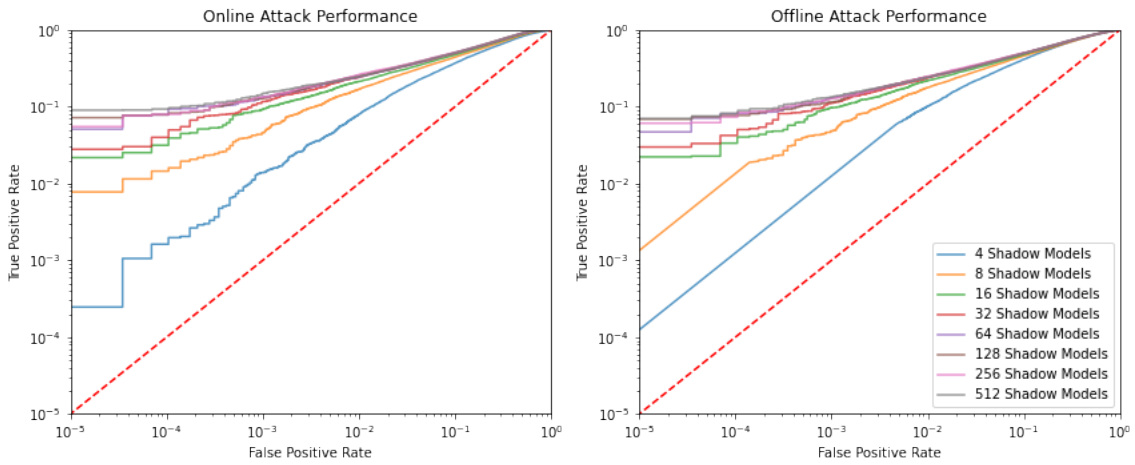

Figure 3.4: Online and Offline ROC Analysis While Varying N Shadow Models on *TweetRumors*

Number of Shadow Models N	Online LiRA	Offline LiRA
4	1.39%	1.25%
8	4.58%	4.83%
16	9.17%	9.67%
32	11.7%	11.2%
64	12.8%	11.5%
128	13.2%	12.9%
256	13.3%	12.8%
512	14.8%	13.4%

Table 3.2: Attack TPR at FPR of 0.1% While Varying N on *TweetRumors*

stand the LiRA on topic model's behavior using different numbers of shadow topic models. Figure 3.4 and Table 3.2 display our results.

Our results in Figure 3.4 and Table 3.2 indicate that increasing the number of shadow model's for the online attack generally increases attack performance, but with diminishing marginal returns. The relationship between the number of shadow and attack performance is similar for the offline attack, but the offline attack tends to perform slightly worse at each N. For the remainder of our attacks we fix $N = 128$ for efficiency reasons.

Number of Topics k	Online LiRA	Offline LiRA
5	31.8%	31.8%
10	44.9%	43.6%
15	60.5%	56.0%
20	67.9%	63.2%
25	72.5%	70.2%

Table 3.3: Attack TPR at FPR of 0.1% While Varying k on *NIPS*

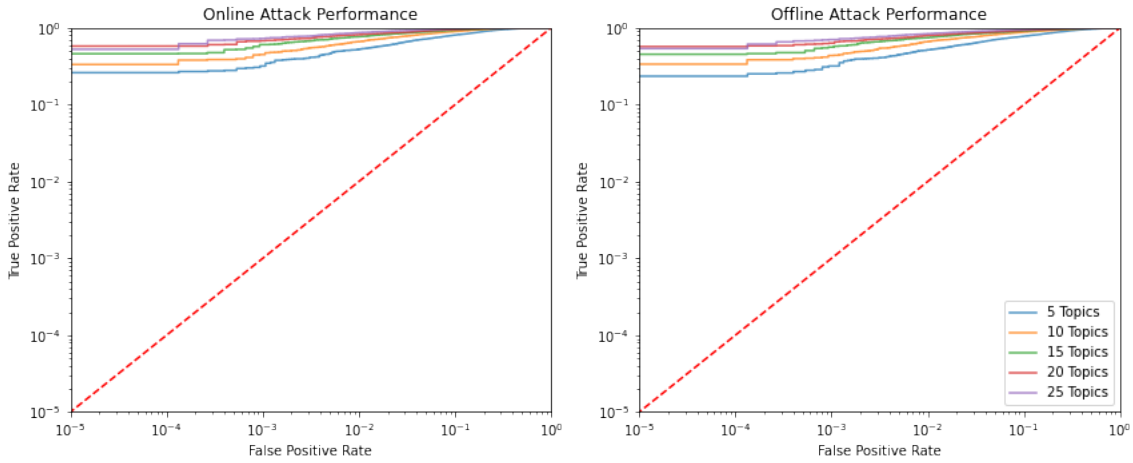

Figure 3.5: Online and Offline ROC Analysis While Varying k on *NIPS*

We also vary the number of topics k on *NIPS* and analyze the affect on attack performance. The ROCs in Figure 3.5 and TPRs in Table 3.5 display the attack performance. By analyzing the curves in Figure 3.5 we see that as we increase k on *NIPS*, the attack performance increases at all FPRs. To compare baseline attacks on *NIPS*, we fix $k = 10$.

We directly compare ROCs for our attack and Huang et al.'s in Figure 3.6 (2022). We can see that our attack strictly dominates Huang et al.'s attack at all false positive regimes for each data set. Furthermore, when we compare Huang et al.'s attack ROCs in Figure 3.6 to Figure 3.4, we see that the Online and Offline LiRA performs better regardless of N for *TweetRumors*.

Data Set	LiRA	Maximum Posterior	Standard Deviation	Entropy
TweetRumors	12.8%	0.19%	0.19%	0.18%
NIPS ($k = 10$)	44.9%	1.69%	1.69%	1.31%
20Newsgroup	21.1%	0.57%	0.57%	0.53%

Table 3.4: Attack TPR at 0.1% FPR with 128 Shadow Models

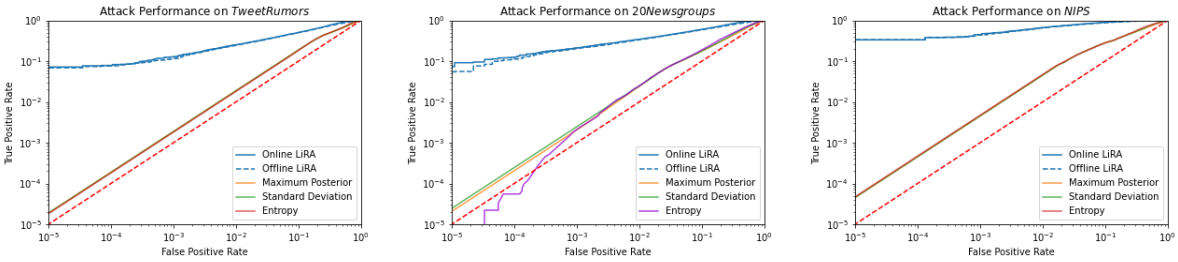

Figure 3.6: Online and Offline ROC Attack Comparison on Each Dataset (128 Shadow Models, NIPS k=10)

3.3.4 Discussion

Our results show that the LiRA is effective against LDA. In general, we reveal that BoW topic models are vulnerable against MIAs. Despite having a limited number of parameters and simple text representations, certain documents inclusion in the training data affect LDA's learned topic-word distribution Φ in a manner that leaks information on the documents presence in the training set.

In large deep-learning based language models, the success of MIAs is directly related to the model's ability memorize certain training examples. Although BoW topic models do not memorize the training data like large language models, their overall vulnerability to MIAs shows that topic models, despite their simpler architecture, exhibit behavior that resembles memorization. BoW topic models learn each topic distribution based on word co-occurrence frequency. If certain words in a document suddenly become more common to the training set, then Φ will reflect the word occurrences in a few specific topics. In many cases, this increases the likelihood that the model will generate said document. As discussed in Chapter 3.2.3, observing some increase in the likelihood of the docu-

ment reflects some degree of memorization.

Our extension of the LiRA turns this observation into a privacy attack by estimating the likelihood of the target document d in N shadow models learned with and without d, and estimates the relative probability of likelihood of d under the target model in either distribution. In the first experiment, we analyze how N affects attack performance. As expected, increasing N increases attack performance but only up to a certain point. Attack performance generally improves when we increase N because there are more samples to accurately estimate the distributions of the likelihood of d in models trained with and without d.

Because the LiRA considers the likelihood-ratio of d, it accounts for per-example hardness. Concurrent with Carlini et al.'s results, attacks that use global thresholds, like Huang et al., do not consider document level differences on the learned model and fail to confidently identify members of the training data (2021a, 2022). Our results in Figure 3.6 verify this intuition; the LiRA outperforms Huang et al.'s attacks at all FPRs (2022).

In our experiments on *NIPS* we considered attack performance while varying the number of topics k. We observed that as we increase k, attack performance increases. Intuitively, because increasing k allows the number of parameters in Φ to grow linearly by the length of the vocabulary set, there are more parameters to represent words' topic distribution. When we increase k, the words in a document are likely relegated to fewer topics which fit the document's word co-occurrence structure better and increase the overall likelihood estimate of the document. This allows certain document's to have a more observable affect to Φ and increases attack performance.

Although we don't directly analyze the relationship between data set size, vocabulary set size, and document-length on attack performance, our results allow us to make few basic observations. First, small data sets with long documents, like *NIPS* appear to be very vulnerable. Intuitively, this observation makes sense because we expect long documents to have a greater affect on the model as noted in in Chapter 3.2.3. Furthermore, longer documents typically yield to a larger, more diverse vocabu-

lary set. For *NIPS* we note that factor of vocabulary size to number of documents is approximately 1:7. Because there are many more parameters in Φ than documents, the model can characterize more complex relationships between the words in each document. We see that for *TweetRumors*, where the number of documents and the vocabulary size are approximately equal, the model is less to exhibit this behavior and the attack generally performs worse compared to the other data sets. Generally, we expect the attack performance to improve as the vocabulary size increases compared to the number of documents in the training set.

Overall, our findings demonstrate that the LiRA can effectively infer document membership in LDA models. The effectiveness of MIAs on BoW topic models has important implications for both researchers and practitioners. First and foremost, practitioners should be mindful of the privacy of their text data its authors before releasing topic models. Our work empirically proves that topic models are vulnerable to strong privacy attacks like the LiRA. When topic modeling with sensitive text data, practitioners should consider adopting privacy preserving methods to releases Φ. Researchers that develop private topic modeling and defenses against MIAs should analyze the empirical privacy guarantees of their methods using the strong attacks presented in this work, rather than relying on the basic attacks that fail to confidently identify members of the training data.

3.3.5 Limitations and Future Work

While our results provide insights on topic models' privacy, there are a few limitations of our methodology and experiments. First, our attack methodology is inherently limited by the BoW model assumptions. We demonstrate that our attack can confident identifying members of the training set for BoW topic models. However, because we rely on a BoW document representation, we can only identify that documents with certain word frequencies are part of the training set. For example, let's say the document d = "mary had a little lamb" was included in the training set. For any alternative document d' with the same combination of words in d, the attack would conclude that d' is in

the training set (i.e. d' = "lamb had a little mary"). This may cause the attack to realize more false positives by concluding that d' is in the training set. The data sets used in our attack evaluations did not include documents with overlapping word combinations. Additionally, most reasonable documents contain a very few or no practical alternative word combinations which eases the implications of this limitation.

Another limitation presents itself in our evaluation methodology by interpreting Λ as a membership confidence score rather than a test statistic to be thresholded by sampling from the null. To describe the statistical implications of interpreting Λ as a membership confidence score, lets consider the offline LiRA under null hypothesis is that the document is a non-member. For a single attack on one document, the significance level α is the probability that the test rejects the null hypothesis when the null is actually true (false positive) at some threshold τ. At the same τ, the statistical power of the test $(1 - \beta)$ is the probability that the test yields a true positive. For any given threshold τ, we can empirically estimate the significance level and power of our hypothesis test by executing a LiRA many times on the same document, obtaining different Λ, comparing each Λ to τ, and calculating the FPR and TPR compared to ground truth.

In our evaluations, we empirically estimate the significance level and power of the underlying likelihood ratio-test by running the LiRA many times on different documents. Because Λ is a probability estimate normalized across LiRAs, we posit that calculating the FPR and TPR at a fixed threshold τ also empirically estimates the significance level and power of the hypothesis test at τ for any document. When we typically perform a hypothesis test and empirically estimate the power of the test we choose τ based on a desired significance level by sampling from the null distribution. Because we interpret Λ like a membership confidence score, we do not follow the sampling procedure for each document. If we say the attack has a TPR of 20% at an FPR of 5%, then the empirically estimated power of the test should be approximately 20% if we choose a threshold based on a significance level of 0.05. However, we do not examine the calibration between evaluation methods. The

reported TPR at the low FPR is entirely empirically estimated. Future work should examine the calibration between the empirically estimated TPR at a FPR and the empirically estimated power of the test when sampling from the null and choosing τ in accordance with a significance level that corresponds to the empirically estimated FPR.

Additionally, our research is focused on cases where the topic modeling output is restricted to the entire topic-word distribution Φ. In many practical use cases, researchers will release the top-n words with or without word probabilities associated with each topic, word clouds of the top n words in each topic or interactive visualizations such as LDAvis (Sievert & Shirley, 2014). Determining the affect that alternative releases have on attack performance by examining ways that clever adversaries can reconstruct Φ given each release mechanism is reserved for future research.

While our results are promising, future work for MIAs against topic modeling should consider a more comprehensive abolation study like in Carlini et al. (2021a). First, researchers should examine closely whether the attack using ζ applies to other probabilistic topic models or neural topic models. Attack performance may be hindered by assuming LDAs generative process if the target topic model is not LDA. Regardless, we expect some similar form of the proposed LiRA to perform well against other topic models. Secondly, researchers should examine the LiRA on topic models under conditions when the shadow model training data is disjoint from the training set or synthetically generated using the model. Finally, future research should use synthetic data sets to directly control vocabulary size, document length, and data set size to directly observe the relationship between each and attack performance. We speculate the attacks would perform well under each suggestion, but this work is left for follow-on research.

Finally, in this work we attempt to draw parallels the between the memorization exhibited by deep-learning models and the affect that certain document's have on BoW topic models. Schofield et al. briefly recognize memorization in their work while quantifying the effect that duplicated documents have on semantic models (2017). However, characterizing memorization in topic models,

other probabilistic generative models, and state-of-the-art ML models remains an open question for the ML research community. Undoubtedly, memorization in ML allows for the privacy attacks discussed in this chapter. Future work on privacy in ML should focus on addressing memorization and developing adequate defenses against privacy attacks.

4

Privacy and Defenses Against Membership Inference

Designing adequate defenses against privacy attacks on ML models is an open field of research, but the primary solutions used to protect privacy are based on generalization techniques and differential privacy. In this chapter, we explore the various defenses to MIAs against topic model's and propose an improved solutions for protecting privacy in topic modeling.

4.1 Privacy Preserving Data Analysis and Machine Learning

The field of privacy-preserving data analysis, whether applied to answer simple statistical queries or train a ML model, is primarily focused on preserving the privacy of individuals while extracting insights from the data. For ML models, a natural definition of privacy stems from *Dalenius' desideratum*: access to a ML model should not reveal anything about the individual observations in the training set that cannot be learned without access (Dalenius, 1977). However, Dwork & Naor show that achieving Dalenius' for any useful data analysis is impossible (2010). Therefore, other notions of privacy arise in machine learning research.

Broadly, we say that a machine learning model should not leak sensitive information about its training data. For example, work on model inversion attacks assume that a privacy violation occurs if the model's outputs reveal information about sensitive attributes used in the training data (Fredrikson et al., 2014). However, this statement raises questions on what constitutes 'sensitive' data. Traditionally we say that, sensitive data encompasses information such as personally identifiable information (PII), financial information, or health information. As machine learning models become more complex and data sets become larger, it is increasingly difficult to identify 'sensitive' information (Brown et al., 2022). Furthermore, declared 'sensitive' information is often correlated to seemingly benign information in the training data or model output.

To ease the burden of defining sensitive attributes or to study privacy under conditions where the entire data set may be considered sensitive, researchers primarily study memorization in ML (Carlini et al., 2023b, 2019, 2021b, Leino & Fredrikson, 2020). In this work, privacy violations occur when the model allows the user to extract verbatim samples or entire examples from the training data. Clearly, learning verbatim information from the training data constitutes a privacy violation for the individuals in the training data. While memorization represents a narrow definition of privacy in ML models, it provides a clear and measurable goal for researchers to work towards solving.

The primary solutions used to protect privacy in ML are based on generalization techniques and differential privacy (DP). Generalization techniques and DP help elevate the memorization problem by encouraging the model to make similar predictions on data regardless of its inclusion in the training data. Some common generalization techniques include de-duplicating the training data, implementing drop-out, or adding regularization (Foret et al., 2021, Lee et al., 2022, Srivastava et al., 2014). While researchers find empirically that generalization techniques can prevent over-fitting and boost privacy, they do not offer the same provable guarantee offered by DP (Dwork et al., 2006b).

DP is particularly useful for answering statistical queries on data sets while preserving individual level privacy. A statistical query is simply a function that takes a data set as input and outputs a statistic or summary of the data. For instance, the U.S. Census recently implemented DP for one-shot releases of simple statistical queries (e.g. counting, sum, or mean queries) on 2020 census data (U.S. Census Bureau, 2021). However, this framework has been extended for a variety of purposes including training ML models.

Under the notion of privacy provided by DP, privacy is the degree to which an individual's information included in the training data can be distinguished in the output of the ML model. By attempting hiding the affect of any one individual on the learned model, DP acts as a direct defense against privacy attacks like MIAs and lessens memorization. Furthermore, DP provides clear, quantifiable method for reasoning about privacy loss.

4.1.1 Fundamentals of Differential Privacy

Differential privacy (DP) is a notion of individual privacy rooted in cryptography that provides strong theoretical guarantees of privacy. DP is not an algorithm, rather it is a mathematical definition that requires the output of some data analysis to be indistinguishable (with respect to a small multiplicity factor) between any two adjacent data sets x and x'.

Definition 4.1.1 (Differential Privacy (Dwork et al., 2006a,b)) Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a randomized algorithm. For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, we say that $\mathcal{M}$ is (ε, δ)-differentially private if for all neighboring databases $x, x' \in \mathcal{X}^n$ and every $S \subseteq \mathcal{R}$

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') \in S] + \delta.$$

When $\delta = 0$, the algorithm $\mathcal{M}$ satisfies ε-DP which is often called *pure differential privacy*. For DP algorithms, ε and δ are often referred to as the *privacy loss parameters*.

The notion of adjacency is critical in DP. Typically, we say that data sets x and x' are adjacent if they differ **by** one observation. Often referred to as *user-level* adjacency, this notion enforces DPs promise of individual level privacy by considering adjacency at the user level (Dwork & Roth, 2014). However, there are other relaxed notions of adjacency. For instance, some definitions of adjacency require that x and x' are adjacent if they differ **in** one observation. This relaxed notion is often referred to as *event-level* adjacency (Dwork & Roth, 2014).

For any text data set, we say that two corpora D and D' are user-level adjacent if they differ by one author's documents. Naturally, user-level adjacency for text documents can be referred to as *author-level adjacency*. If two corpora D and D' differ by one document, then adjacency is at the *document-level*. When designing DP algorithms on text-data, we often make the simplifying assumption that each document in a corpus has a unique author. In this case, author-level adjacency is equivalent to document-level adjacency. For corpora the most relaxed notion of adjacency is *word-level adjacency*; two corpora D and D' are word-level adjacent if they differ by one word in one document. Assuming that each document belongs to a unique author or using word-level adjacency are common methods used to bound the sensitivity for DP text analysis.

Satisfying DP relies defining adjacency and adding noise scaled to the sensitivity of the statistical query of interest f.

Definition 4.1.2 (Global Sensitivity) Let $f : \mathcal{X}^n \rightarrow \mathbb{R}^k$ be a real valued function. The global sensitivity of the function GS_f is the maximum change in f evaluated across all adjacent data sets $x, x' \in \mathcal{X}^n$ such that:

$$GS_f = \max_{x,x' \in \mathcal{X}^n} \|f(x) - f(x')\|.$$

Adding noise scaled to the global sensitivity allows us to mask an individuals contribution to the query f, even if the individual is an extreme outlier. Considering sensitivity across possible all adjacent data sets gives us a worst-case estimate of sensitivity. By adopting event-level adjacency, researchers attempt to control the sensitivity. However, we lose the strong individual privacy guarantee offered under user-level privacy; the DP algorithm is indistinguishable regardless of any one user's data included in the analysis. Another method used to reason about the sensitivity of a statistical query is to consider the local sensitivity of f.

Definition 4.1.3 (Local Sensitivity (Nissim et al., 2007)) Let $f : \mathcal{X}^n \rightarrow \mathbb{R}^k$ be a real valued function. The local sensitivity of said function $LS_f(x)$ is the maximum change in f with respect to a data set $x \in \mathcal{X}^n$:

$$LS_f(x) = \max_{x \sim x'} \|f(x) - f(x')\|.$$

Definition 4.1.4 (Smooth Sensitivity (Nissim et al., 2007)) For $\beta > 0$ and function $f : \mathcal{X}^n \rightarrow \mathbb{R}^k$, the β-smooth sensitivity of a function is

$$S^*_{f,\beta}(x) = \max_{x \in \mathcal{X}^n} (LS_f(x') * e^{-\beta\, d(x,x')})$$

where $d(x, x')$ is the distance between x and x'.

By using smooth sensitivity, researchers protect privacy without addressing worst-case adjacent data sets. Relying on local senstivity relaxes DPs worst-case privacy guarantees, but is useful for

preserving utility when the global sensitivity is possibly unbounded or the actual data set does not realize global sensitivity or both.

Along with strong theoretical guarantees of privacy, DP mechanisms have nice properties convenient for designing private algorithms: DP mechanisms are closed under post-processing and they compose adaptively. Because DP mechanisms are closed under post-processing, we can compute any function based on the output of the mechanism without leaking additional privacy. We define post-processing formally in Theorem 4.1.1.

Theorem 4.1.1 (Post-Processing (Dwork et al., 2006b)) Let $M : \mathcal{X}^n \to \mathcal{R}$ be an (ε, δ)-DP algorithm and let $g : \mathcal{R} \to \mathcal{S}$ be a function that maps the set of all possible outputs of M to some arbitrary set $\mathcal{S}$. Then, $g \circ M$ is (ε, δ)-DP.

Because DP composes adaptively, we can build private algorithms by iteratively combining outputs from DP mechanisms. We refer to to this process as composition, and define Basic Composition formally in Theorem 4.1.2.

Theorem 4.1.2 (Basic Composition (Dwork et al., 2006a)) For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, let M_i be an $(\varepsilon_i, \delta_i)$-DP algorithm for $i \in [k]$. Their composition $M_{[k]} = (M_1(x), \ldots M_k(x))$ is $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-DP.

A strength of composition in DP is that DP algorithms compose adaptively: the privacy guarantees of a sequence of DP mechanisms can be combined, even if the choice of each mechanism depends on the outcome of the previous ones.

The Basic Composition Theorem stipulates that privacy degrades linearly in the number of mechanisms combined k. However, there are various composition theories, such as advanced composition, that attempt to achieve sub linear privacy loss with k (Dwork et al., 2010). Additionally, other variants or relaxations of DP, like Rényi Differential Privacy (Rényi-DP) or Zero Concentrate

Differential Privacy (*zCDP*), that allow for improved composition (Bun & Steinke, 2016, Mironov, 2017).

In the ML context, researchers apply DP to learn neural-network based models by privatizing the learning algorithm stochastic gradient decent (SGD). DP-SGD protects privacy by perturbing the gradients with Gaussian noise before performing the gradient descent update (Abadi et al., 2016). Various implementations of DP-SGD allow researchers to apply privacy using popular deep-learning libraries like tensorflow and pytorch (Abadi et al., 2016, Yousefpour et al., 2022). However, there is also an effort to develop DP variants of other learning algorithms designed for topic models.

4.1.2 Trust Models of Diferentially Privacy

Before introducing DP topic models, we first define the two primary trust models that dictate the operational assumptions and requirements for implementing DP. In the *central model* a trusted data curator holds a data set and performs all private computation (Dwork et al., 2006b). Under the *local model* each individual party is responsible for private computation before sharing their data with an untrusted curator. A good example of DP in the local model is the randomized response algorithm where participants in a survey randomize their survey response in a DP-manner before reporting to the data curator (Warner, 1965).

In this thesis, we are primarily concerned with investigating private topic models under the central model. Operationally, the central model for topic models is captured by the scenario posed in Chapter 3.2.1 where hospital XYZ collects data on patients with condition X and learns a topic model based on their data. In this scenario, hospital XYZ acts as the 'trusted' curator whose goal is to release a DP topic-word distribution.

Model	Notion of DP	Notion of Adjacency	Learning Method	Other Technical Details
Zhu et al. (2016)	ε-DP	document-level	CGS	
HDP-LDA	ε-DP	word-level	CGS	
SUB-LDA	(ε, δ)-DP	word-level	CGS	sub-sampling
RDP-LDA	Rényi-DP	word-level	CGS	
Park et al. (2018)	(ε, δ)-DP	document-level	VI	moments accountant and sub-sampling
Decarolis et al. (2020)	Rényi-DP	document-level	SA	local sensitivity from propose-test-release
Topic-DP	Rényi-DP	author-level	PH	pain-free smooth sensitivity

Table 4.1: A brief summary of the existing literature on DP topic modeling. Learning methods are Collapesed Gibbs Sampling (CGS), variational inference (VI), spectral algorithm (SA), Model Agnostic or Post-Hoc (PH). References for named model's contained paragraph body.

4.1.3 Diferentially Private Topic Modeling

At its core, private topic modeling presents a Bayesian inference problem that requires private computation of the posterior topic-word distribution. In this section, we provide an overview of the existing DP topic modeling algorithms that operate under the central model and highlight their strengths and limitations. Table 4.1 provides a brief summary of the reviewed DP topic models.

Collapsed Gibbs Samplers (CGS) are by far the most popular method of learning learning DP topic models. To privatize the CGS process for LDA, researchers take a few different approaches. First, Zhu et al. recommend learning LDA using CGS, then perturbing the final word-count statistics to satisfy-DP (2016). Zhao et al. introduce HDP-LDA (2021b) which perturbs the word-count statistics in each iteration of CGS (2021b). Because HDP-LDA composes over many iterations, managing the privacy budget is difficult. To address this, SUB-LDA leverages sub-sampling procedures to boost privacy over many iterations, and RDP-LDA uses Rényi-DP to provide tighter bounds on privacy loss (Huang & Chen, 2021, Huang et al., 2022).

Although variational inference is extremely popularly in the non private setting, there are limited DP implementations for LDA. Park et al. propose a document-level private process for learning topic models via stochastic variational inference (2018). By using variational inference and the moments accountant method from Abadi et al., composes better over iterations 2016.

Unlike previous approaches that use MCMC or variational inference, Decarolis et al. propose learning private LDA with a spectral algorithm based on matrix decomposition (2020). Overall, their proposed algorithm benefits over variational inference and MCMC methods because it is more efficient and does not require composition over many learning iterations. However, the efficiency does not scale with data set size, and they rely on local sensitivity using the Propose-Test-Release framework (Dwork & Lei, 2009).

Instead of focusing on privatizing the learning algorithm in LDA, Wang et al. introduce a model-agnostic approach that privatizes the topic-word distribution after estimation (2022a). TopicDP benefits over LDA specific DP learning algorithms because it can be applied to any learned topic-word distribution. However, the authors use the pain-free sensitivity sampler and use smooth sensitivity in their privacy guarantee.

While each DP topic modeling algorithm has their advantages and disadvantages, we notice a few common themes across implementations. First, the iterative nature of learning algorithms for LDA posses a difficult composition issue in terms of managing the privacy loss. Secondly, because topic models operate on BoW text data, reasoning about adjacency is critical to ensuring privacy. We speculate that privatized CGS procedures that rely on word-level adjacency to control sensitivity may be insufficient to protect against strong privacy attacks like the one's discussed in Chapter 3. Finally, all of the proposed methods emphasize releasing the topic-word distribution Φ, but do not address the privacy concerns with the releasing the vocabulary set that corresponds to Φ.

In the subsequent sections, we argue that without privatizing the vocabulary set, any privatized release for Φ does not satisfy DP. To fill this gap in the literature, we propose Fully DP Topic Modeling (FDPTM) as a framework for DP releases of topic-word distributions and their vocabulary sets. We analyze the privacy associated with FPDTM, and empirically evaluate utility and privacy with the strong attacks introduced in Chapter 3.

4.2 Private Vocabulary Selection and Topic Modeling

When performing topic modeling and analyzing the resulting topic-word distribution Φ, each column of Φ corresponds to a word indexed from a vocabulary set. Without the vocabulary set, Φ loses practical interpretability because the user is simply left with a collection of indices and their probabilities of appearing in each topic. Topics are best summarized by a collection of high-probability words, rather than high-probability indices. Therefore, Φ is almost meaningless without the associated vocabulary set.

Under the central model of DP, we assume that a data analyst queries the trusted data curator, who holds a private data set, to release a topic-model where the number of topics k is publicly known. In doing so, the analyst is essentially submitting two queries on the curator's data set: a query for Φ and a query for the accompanying vocabulary set. However, the literature on private topic modeling largely disregards privacy associated with releasing the vocabulary set by assuming a fixed, perhaps publicly known vocabulary set.

In this case, releasing the topic-word distribution Φ without sanitizing the accompanying vocabulary set in a DP manner could cause a privacy violation. Even if the values in Φ are private, as a whole the released topic-word distribution does not satisfy DP because the vocabulary set accompanied by Φ is not private.

To illustrate this point, let's consider the following scenario. Suppose we have a corpus D with just one document d. If we were to naively release Φ and the vocabulary set for the corpus D, then we must release all of the words in d. Clearly, we leak the privacy of the document in D. We can address this privacy violation by studying methods for DP vocabulary selection.

4.2.1 Private Vocabulary Selection

The issue of DP vocabulary selection can be formalized as the DP Set-Union (DPSU). Our work does not focus on solving the DPSU problem. Instead, we simply focus on applying DPSU to private vocabulary selection for topic models.

Definition 4.2.1 (Differentially Private Set Union (DPSU) (Gopi et al., 2020)) Let U be some universe of items, possibly of unbounded size. Suppose we are given a database D of users where each user i has a subset $W_i \subseteq U$. We want an (ε, δ)-differentially private Algorithm A which outputs a subset $S \subseteq \cup_i W_i$ such that the size of S is as large as possible.

From the definition, we notice that DP vocabulary selection is simply an instance of the DPSU problem. Instead, we say that we consider a universe U of terms and a corpus D of documents with authors. Each author i contributes a subset $W_i \subseteq U$ of terms in their documents. Our goal is to design a (ε, δ)-DP algorithm A to release the vocabulary set $S \subseteq \cup_i W_i$ such that the size of S is as large as possible.

Researchers first considered the DPSU where each user only contributed one item. Under this framework, researchers sought to release as many approximate counts of as many items as possible in $\cup_i W_i$ (Korolova et al., 2009, Wilson et al., 2019). In vocabulary selection terms, the authors' sought to release approximate word frequencies for as many words as possible. Their algorithms guarantee privacy by constructing a histogram of $\cup_i W_i$, adding noise from a Gaussian or Laplace to each word frequency in the histogram, and releasing all word frequencies that fall above a certain threshold. This approach satisfies (ε, δ)-DP for large enough thresholds, but is not so useful for vocabulary selection because we assume that every user only contributes one word.

Korolova et al. and Wilson et al. consider the case of users contributing more than one word to the vocabulary by upper-bounding the sensitivity based on the maximum the number of words

contributed by all users $\Delta_0 = \max_i |W_i|$. However, most users contribute significantly less than Δ_0. Therefore, building the histogram directly on counts of words in W_i wastes sensitivity. To address said issue, Gopi et al. directly address the DPSU problem using weighted histograms and update policies (2020).

Gopi et al. propose two algorithms, *Policy Laplace* and *Policy Gaussian*, to address the DPSU algorithm (2020). The proposed algorithms build weighted histograms using update policies with certain contractive properties which allow them to control sensitivity. The author's prove that under certain hyperparamter selection, their algorithms satisfy (ε, δ)-DP, and empirically outperform previous DP vocabulary selection mechanisms.

Gopi et al.'s solution is an important step in the right direction, but their mechanism have a few limitations for vocabulary selection (2020). Namely, they consider that each user contributes a set of words without multiplicity discounting the fact that some words appear multiple times in the same document. Carvalho et al. propose the algorithm GW to solve DPSU with item frequency (2022). They prove that their algorithm satisfies (ε, δ)-DP, and empirically outperforms Gopi et al.'s solution (2020). Carvalho et al. also present a solution which boosts performance using term frequency information from public data sets (2022). However, we note that incorporating public data complicates privacy and usability for language modeling (Tramèr et al., 2022).

While the described solutions for DPSU address DP vocabulary selection, the issue of DP vocabulary selection can also be framed as the t-Heavy Hitters problem (**?**). However, we opt to primarily consider the DPSU as a more general problem aimed to release as many terms as possible, not just terms that appear more than t times, and explore how we can apply existing DPSU algorithms for topic modeling.

4.2.2 Fully Differentially Private Topic Modeling

Now that we've introduced DP algorithms for topic modeling and vocabulary selection, we combine them to propose Algorithm 3 as a high-level procedure for fully differentially private topic modeling (FDPTM).

Algorithm 3 Fully Differentially Private Topic Modeling (FDPTM)

1: **Require:** Corpus D, DP vocabulary selection algorithm $\mathcal{M}_1$ and DP topic modeling algorithm $\mathcal{M}_2$.
2: Apply standard pre-processing and tokenization: $D_{pre} \leftarrow \texttt{PRE}(D)$
3: Apply DP vocabulary selection to corpus: $S \leftarrow \mathcal{M}_1(D_{pre})$
4: Remove all words $w \in D_{pre}$ if $w \notin S$: $D_{san} \leftarrow \texttt{SAN}(D_{pre}, S)$
5: Learn Φ using DP topic modeling: $\Phi \leftarrow \mathcal{M}_2(D_{san})$
6: **Release:** Topic-word distribution Φ and corresponding vocabulary set S

FDPTM composes the privately selected vocabulary set S and the private topic-word distribution Φ. Therefore, Theorem 4.2.1 holds due to the theory of composition outlined in Theorem 4.1.2. Applying $\mathcal{M}_2$ is not a simple post-process of S because Φ is learned using D_{san} which encodes word co-occurrences in the original data set.

Theorem 4.2.1 (Privacy Guarantee of FDPTM) If the algorithm $\mathcal{M}_1$ for selecting the vocabulary set S satisfies $(\varepsilon_1, \delta_1)$-DP, and the algorithm $\mathcal{M}_2$ for learning Φ satisfies $(\varepsilon_2, \delta_2)$-DP, then the overall release for the topic model satisfies $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.

Proof (sketch). First, let $\mathcal{M}_1$ be a $(\varepsilon_1, \delta_1)$-DP vocabulary selection algorithm that returns a private vocabulary set S, and $\mathcal{M}_2$ be a $(\varepsilon_2, \delta_2)$-DP topic modeling algorithm that returns a private topic-word distribution Φ. Now, let PRE be a function that takes a corpus D and applies standard pre-processing procedures, and SAN be a function that takes a corpus D and a vocabulary set S and removes all words $w \in D$ if $w \notin S$. Finally, let $\mathcal{M}_x$ represent the FDPTM algorithm such that for a

corpus D

$$\mathcal{M}_x = (\mathcal{M}_1(\texttt{PRE}(D)),\ \mathcal{M}_2(\texttt{SAN}(\texttt{PRE}(D),\ \mathcal{M}_1(\texttt{PRE}(D))))).$$

Definition 4.2.2 (k-Stability (Thakurta & Smith, 2013)) A function $f : U^* \rightarrow \mathcal{R}$ is k-stable on input D if adding or removing any k elements from D does not change the value of f, that is, $f(D) = f(D')$ for all D' such that $D \triangle D' \leq k$. We say f is stable on D if it is (at least) 1-stable on D, and unstable otherwise.

Lemma 4.2.1 (Composition with Stable Functions (Thakurta & Smith, 2013)) Let f be a stable function, and let $\mathcal{M}$ be an (ε, δ)-DP algorithm. Then, their composition $\mathcal{M}(f(x))$ satisfies (ε, δ)-DP.

The functions PRE and SAN are stable algorithms because each document is processed independently of the others based on a standard set of rules. Simply, adding or removing a document from the corpus does not affect the functions behavior on other documents,. Therefore, PRE and SAN are stable functions.

Via our definition of $\mathcal{M}_1$, and using the fact that PRE is a stable function, then the first term in $\mathcal{M}_x$, $\mathcal{M}_1(\texttt{PRE}(D))$, satisfies $(\varepsilon_1, \delta_1)$-DP. The second term of $\mathcal{M}_x$ applies $\mathcal{M}_2$ which directly depends on $\mathcal{M}_1$ and the stable functions PRE and SAN. Therefore, via adaptive composition, $\mathcal{M}_x$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP. □

The first step of FDPTM, the algorithm PRE, can include standard text pre-processing procedures such as removal of all non-alphabetic characters, tokenization, removal of stop words, removal of tokens based on length, lemmentization, and tokenization. Each of these tasks are stable, and do not affect the privacy of FDPTM. We note that some pre-processing steps, like removing corpus specific stop words based on corpus frequency, are not stable, and should be avoided in PRE for FDPTM. Additionally, any additional vocabulary set cleaning after applying $\mathcal{M}_1$ maintains $(\varepsilon_1, \delta_1)$-DP via Theorem 4.1.1 on post-processing.

Successfully implementing FDPTM requires a few key considerations. First, the process depends on carefully tuning the privacy parameters and other parameters within M_1 and M_2. The data curator must ensure that each mechanism satisfies differential privacy (DP) independently and understands the privacy loss associated with each algorithm. To this end, the data curator must ensure that M_1 and M_2 satisfy the same notion of differential privacy. This framework may be difficult to apply when M_1 satisfies pure DP and M_2 satisfies Rényi-DP. Finally, the curator should pay close attention to the authorship of each document and how it affects the privacy guarantees of each mechanism. For instance, in our evaluations we make the assumption that each document has a unique author to ensure author-level privacy.

Overall, FDPTM is a modular framework for releasing meaningful topic-word distributions under the central model of DP. Although our evaluations focus on LDA, other topic models tailored to specific use cases can easily fit into the proposed procedure. Furthermore, as state-of-the-art DP vocabulary selection and DP topic modeling algorithms develop, the quality of the released topic-word distribution will improve due to FDPTM's flexible framework.

4.3 Evaluations

In this section, we empirically evaluate the utility and defensive ability of FDPTM. We investigate how the vocabulary selection mechanism and DP learning mechanism change the quality of the learned topic model. Additionally, we empirically evaluate the privacy provided by tuning M_1 and M_2 using membership inference attacks.

We evaluate the utility of FDPTM using topic coherence and the log-likelihood of the data normalized by vocabulary length. Natural likelihood measures are inherently skewed by the size of the vocabulary length; corpses with a larger vocabulary set tend to have higher likelihoods. Therefore, to account for the varying size of the vocabulary set in FDPTM, we divide the log-likelihood of the data

by the length of the sanitized vocabulary set.

While likelihood-based measures assess model fit or the probability of the model generating the corpus, they fail to capture the quality or coherence of the model from a human perspective. In fact, Chang et al. demonstrate that likelihood metrics are negatively correlated with human-measured topic interpretability (2009). Instead, topic coherence is a measure designed to capture human interpertabilty of topics.

Definition 4.3.1 (Topic Coherence (Mimno et al., 2011)) Let $D(v)$ be the document frequency of word v (i.e. the number of documents where word v occurs), $D(v, v')$ be the co-document frequency of words v and v' (i.e. the number of documents where v and v' both occur), and $V^{(t)} = \{v_1^{(t)}, ..., v_M^{(t)}\}$ be a list of the M most probable words in topic t. Topic coherence for a topic t is

$$coherence(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})},$$

where 1 is a smoothing factor to avoid taking the logarithm of zero.

Mimno et al. show that topic coherence measures correlate to human evaluations of topic quality (2011). Topic's with higher coherence are considered better to be more interpretable. In our analysis, we select the topic $M = 10$ words and report the average topic coherence across each topic.

We empirically evaluate the privacy associated with FDTM by conducting the online LiRA for topic models presented in Chapter 3. Similar to previous attack evaluations, we are interested in analyzing the True Positive Rate at low False Positive Rates and plot the ROC curve on log scaled axes.

4.3.1 Experimental Set-Up

Using the *TweetRumors* data set from Chapter 3.3.1, we conduct a few different experiments to evaluate the FDPTM. For each experiment we use Carvalho et al.'s implementation of GW for DP vocabulary selection, and our implementation of Zhu et al.'s DP LDA algorithm (2020, 2016).* For DP vocabulary selection using GW, we fix the privacy parameter $\delta = 10^{-5}$ and choose the cut-off value Γ with the parameter $\alpha = 3$ like in Gopi et al. and Carvalho et al. (2022, 2020).

With those basic settings for GW, we run the following experiments to evaluate utility. First, we vary the privacy loss parameter for GW $\varepsilon_1 \in \{1, 2, 3, 4, 5, 10, 15, 20\}$ and hold the privacy loss parameter for DP LDA fixed at $\varepsilon_2 = \infty$. When $\varepsilon_2 = \infty$ LDA is not private. For each setting of ε_1 we evaluate the perplexity and average topic coherence. To compare performance DP vocabulary selection and DP LDA, we also fix $\varepsilon_2 = 3$ and vary ε_1. Then, we fix $\varepsilon_1 = 3$, vary ε_2, and conduct the same evaluations.

Because automated topic evaluation is inherently imperfect, we choose to display the top 10 words in a specific topic and their corresponding word probabilities. We display the top 10 words and probabilities over 3 different privacy settings with the same global privacy budget ($\varepsilon_1 + \varepsilon_2 = 6$): $\varepsilon_1 = \varepsilon_2 = 3$, $\varepsilon_1 = 2$ & $\varepsilon_1 = 4$, and $\varepsilon_1 = 4$ & $\varepsilon_1 = 2$. We'd like to note that the data set *TweetRumors* is a collection of tweets over five major news events. We choose to display a topic concerning the 2014 shooting of Micheal Brown in Ferguson, Missouri because of the event's widespread notoriety and distinct impact on social media.†

To empirically evaluate the privacy of FDPTM, we test FDPTM against the Online LiRA presented in Chapter 3. For all attacks we learn 64 shadow models using sklearn's implementation

*https://github.com/ricardocarvalhods/diff-private-set-union

†We acknowledge the tragic nature of Michael Brown's death and its lasting impact on our society, particularly in terms of raising awareness on police brutality, racial injustice, and systemic inequality in the United States. Other events included in the data set, such as the Germanwings Flight 9525 crash in 2015, the Charlie Hebdo Shooting in 2015, the Lindt Cafe siege in 2015, and the 2014 shootings at Parliament Hill in Ottawa, were also devastating occurrences.

of LDA with default learning parameters.[‡] Like before, we conduct attacks while varying $\varepsilon_1 \in \{1, 5, 10\}$ and fixing $\varepsilon_2 = 5$. Then, we fix $\varepsilon_1 = 5$ and vary ε_2.

All of the proposed simulations ran on MIT Lincoln Laboratory Supercomputing Center's system TX-Green (Reuther et al., 2018). We implemented the algorithms and simulations in Python version 3.9. For the the *TweetRumors* data set we fix the number of topics $k = 5$. We repeat each experiment 10 times and report the results across each iteration.[§]

4.3.2 Results

In our first experiment, we vary the privacy loss parameter for GW ε_1. Figure 4.1 displayws the model fit and topic coherence as we increase ε_1. The plot that displays topic coherence contains error bars that show 1 standard deviation away from the mean topic coherence score across all iterations. We also plot the average vocabulary size with Φ as we increase ε_1 in Figure 4.2.

In general, we see that as we increase ε_1, the log-likelihood of the model increases, even when we control for vocabulary size. We expect to see this result because increasing the privacy loss involves adding less noise addition into the vocabulary selection process. We do not observe the same results for the topic coherence. Only at $\varepsilon_1 = 1$ is the average topic coherence more than a standard deviation away from each other.

For the second experiment, we hold either ε_1 or ε_2 constant while varying the privacy loss for the alternative ε. Figure 4.3 displays the model fit and topic coherence as we increase the varying ε. As expected, when we increase ε we see an increase in the log-likelihood of the model. Topic coherence increases with ε. We note that topic coherence when varying ε_1 begins to plateau around $\varepsilon_1 = 3$ while topic coherence continues to increase as ε_2 increases. Therefore, we realize diminishing utility returns faster as we increase ε_1.

[‡] https://scikit-learn.org/

[§] Code for the experiments and evaluations is available at https://github.com/nicomanzonelli/topic_model_attacks/dp_defenses.

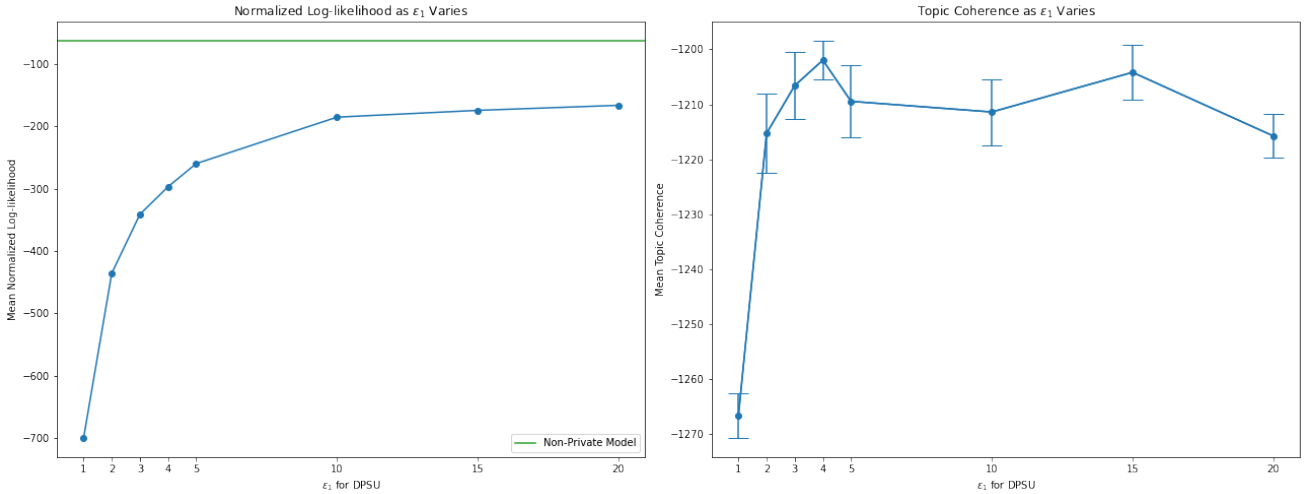

Figure 4.1: Log-likelihood and topic coherence as ε_1 increases and LDA is not private. The green horizontal line represents a non-private vocabulary selection model log-likelihood. We choose not to display the non-private baseline for topic coherence because it is very small compared to the figure (≈ -117).

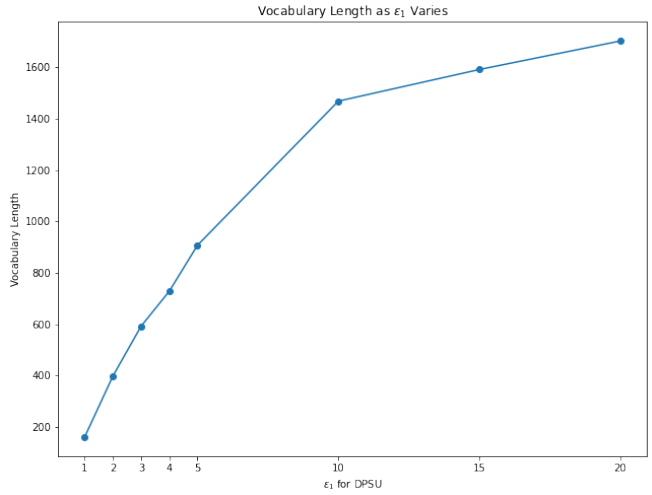

Figure 4.2: Vocabulary Length as ε_1 Varies. The non-private vocabulary size is 5,942.

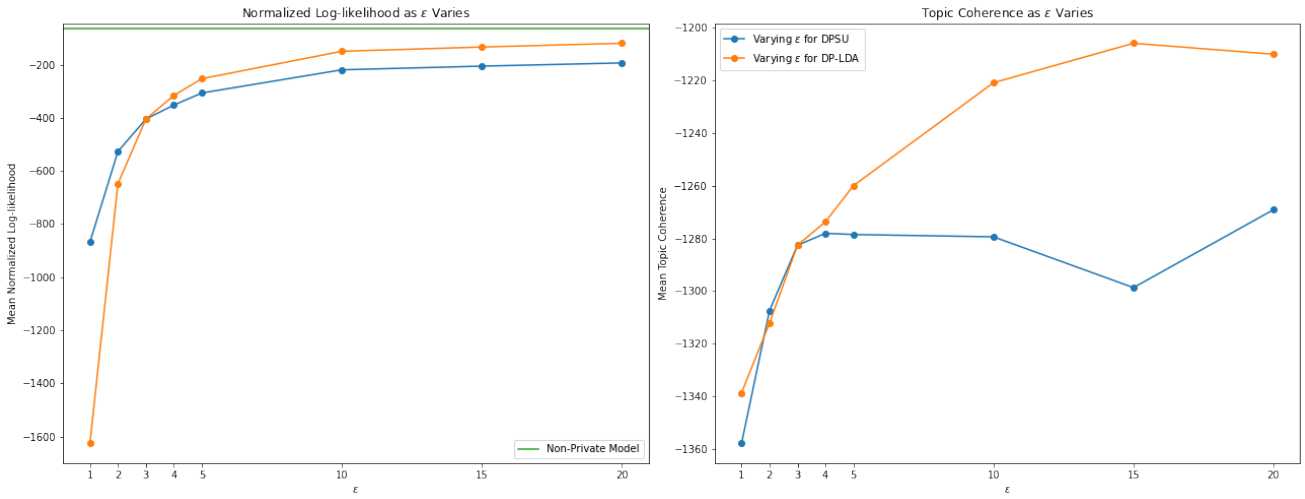

Figure 4.3: Log-likelihood and Topic Coherence as ε Increases. The blue line shows the the results while varying ε_1 for DPSU and holding $\varepsilon_2 = 3$. Orange displays results for varying ε_2 for DP LDA while holding $\varepsilon_1 = 3$.

When visually inspect our topics in Table 4.2, we see that at the same privacy total privacy budget $\varepsilon_1 + \varepsilon_2 = 6$ the privacy loss parameter for DP LDA has a more significant effect on the model. When $\varepsilon_1 = \varepsilon_2$ we see a minor amount of word intrusion because "ottowa", and "parliament" sneak their way into the top 10. When $\varepsilon_1 = 2$ and $\varepsilon_2 = 4$, the ferguson topic becomes more coherent because word intrusion ("cheif" and "charliehebdo") only appear near the bottom of the topic. Finally, when $\varepsilon_1 = 4$ and $\varepsilon_2 = 2$ topic coherence decreases significantly. Word intrusion propagates through many words in the topic because most of the noise is included in the LDA learning process.

In our final set of experiments we evaluate attack performance by varying $\varepsilon_1 \in \{1, 5, 10\}$ and fix $\varepsilon_2 = 5$. Then, we fix $\varepsilon_1 = 5$ and vary ε_2. Figure 4.4 displays the ROC curve's for each attack. Generally, we see that attack performance decreases as we decrease either privacy loss parameter.

4.3.3 Discussion

Overall, our empirical results reveal that FDPTM is a practical approach for DP topic modeling that protects against MIAs. From the utility perspective, we see that increasing the privacy loss parame-

$\varepsilon_1 = \varepsilon_2 = 3$		$\varepsilon_1 = 2$ & $\varepsilon_2 = 4$		$\varepsilon_1 = 4$ & $\varepsilon_2 = 2$	
Word	Prob.	Word	Prob.	Word	Prob.
ferguson	0.106	ferguson	0.178	ferguson	0.077
police	0.061	police	0.070	charliehebdo	0.043
ottowa	0.046	mikebrown	0.040	police	0.029
shooting	0.034	officer	0.02	jesuischarlie	0.017
mikebrown	0.027	black	0.018	right	0.015
shot	0.019	brown	0.018	attack	0.014
officer	0.013	right	0.016	hebdo	0.011
brown	0.014	chief	0.016	like	0.011
parliament	0.013	charliehebdo	0.014	mikebrown	0.010
michael	0.012	shot	0.014	charlie	0.010

Table 4.2: Top 10 Word Probabilities for Various Settings of ε_1 for DPSU and ε_2 for DP LDA.

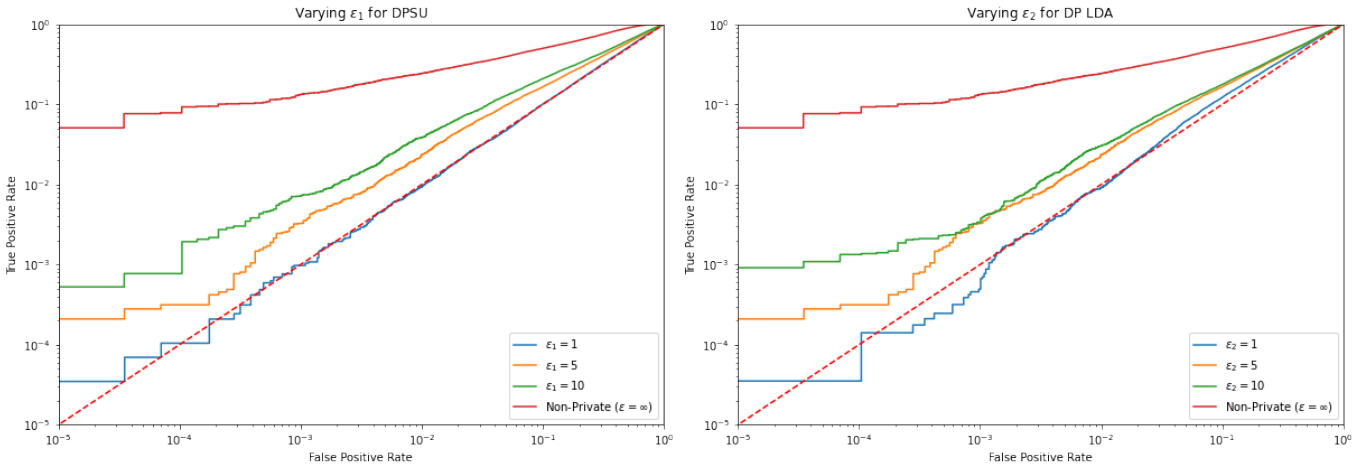

Figure 4.4: Attack ROC Curve's While Varying Privacy Parameters.

ter for DP vocabulary selection led to an increase in the log-likelihood of the model, while the topic coherence does not significantly change after $\varepsilon = 3$. Therefore, noise injection during the vocabulary selection process may decreases model fit when accounting for vocabulary size but does not necessarily affect the coherence of the topics generated for $\varepsilon_1 > 3$. Additionally, when we compare the coherence measured while increasing ε_2, the affect of varying ε_1 seems less significant at most privacy loss parameters.

Although we observe a decrease in topic coherence and model fit due to DPSU, we gain empirical privacy. DP vocabulary selection decreases attack accuracy because we shrink the size of the vocabulary set. By decreasing the length of the vocabulary set, we limit the number of parameters in Φ. In turn, each documents' topic distribution becomes more similar to other documents' topic distributions, and the likelihood of the document tends to change less based on it's inclusion in the training data. Pruning the vocabulary set also relates to the affect that outlying words have on the model. In Chapter 3, we noticed that outlying words (words used infrequently) inclusion in the training set affected learned model. By simply removing outlying words with DP vocabulary selection, we limit the ability for them to have large effects on the topic model and protect privacy.

Overall, the addition of DP vocabulary selection to the DP topic modeling work-flow is important for guaranteeing private, interpretable releases of Φ. Not only does DP vocabulary selection allow for fully DP releases of Φ, it also provides an effective defense against MIAs. Our results indicate that dedicating most of the privacy budget to DP topic modeling, rather than vocabulary selection, increases utility at the same privacy level. Intuitively, we can inject less noise into DP learning procedure, and more into the vocabulary selection algorithm ($\varepsilon_1 < \varepsilon_2$) to increase model interpretability at the same global privacy loss ($\varepsilon_1 + \varepsilon_2$).

4.3.4 Limitations

While our results are promising, our experiments and methodology has a few limitations. First, FDPTM's assumptions under the central model of privacy consider the number of topics k to be publicly known. There may be instances where k gives the adversary insight on the individuals in the data. However, by incorporating DP vocabulary selection into DP topic modeling, FDPTM takes an important step in the right direction toward complete privacy.

Second, we are limited by the breadth of our experiments. Additionally, similar to the limitations mentioned in Chapter 3.3.5, we are limited by the structure of the data set evaluated. To pose more robust experiments we could create synthetic corpses with different lengths, vocabulary sizes, and word co-occurrence structures to concretely understand the effect of each on utility and privacy. However, performing our attacks on a real data set helps build interpretation of FDPTM in practical applications where the data distribution is similar.

Additionally, when assessing utility, we inherently encounter the limitations associated with automatic topic evaluation. There is a large body of existing and active research on studying automatic topic model assessment. The overall consensus is that topic model evaluation is imperfect. While different measures give us different insights into the topic model, the best evaluation is typically based on human inspection (Hoyle et al., 2021). However, it's often best to use as many measures as possible. We address this limitation by using two different topic assessment measures.

Finally, one significant limitation of the presented work is that we do not evaluate generalization techniques on topic models. Comparing our DP topic modeling methods to generalization would provide a better picture of privacy preserving ML in the context of topic models. In the future, we would be interested in studying how each technique could decrease attack efficacy and increase model generalization ability.

4.3.5 Other Defenses and Future Work

While DP remains the gold-standard in privacy protection, implementing sound DP algorithms and reasoning about privacy loss requires time and resources. For this reason, ML researchers study a number of generalization techniques to increase usability and protect against privacy attacks. In fact, in some cases generalization techniques out perform DP (Liu et al., 2021).

Research on improving model generalization for privacy reasons is mostly focused on deep learning models. However, we expect generalization techniques to assist in protecting privacy for probabilistic topic models as well. Although we do not evaluate them, it's worth mentioning a few generalization techniques that could be applied to topic models in follow-on research.

Data de-duplication has been shown to alive memorization in deep learning models, and we expect it to have similar effects topic models (Lee et al., 2022). A study by Schofield et al. showed that under certain conditions, LDA will sequester repeated documents into a small number of topics. We would expect this repeated document to very vulnerable to the attack proposed in Chapter 3 because the documents concentrated topic distribution would yield high likelihood estimates when the document is included in the train set. In addition to data de-duplucation, we expect that simply pruning the vocabulary set and removing terms that appear infrequently would help protect privacy for topic models for the same reason that DP vocabulary selection decreases attack efficacy.

Finally, while generalization techniques incorporated in the learning process like drop-out and regularization are do not see common implementation in topic models, they could assist in protecting privacy. While drop-out is designed for gradient-descent, creative topic modeling learning algorithms could incorporate drop-out or a regularization term. We'd expect these generalization techniques to help protect privacy.

5

Conclusion

Our research motivates ethical considerations in the design and application of ML systems. While topic models seem vintage, LDA is still a popular choice understanding collections of text. Additionally, despite the current focus on researching deep neural network ML models, we gain valuable perspective from studying simple probabilistic generative models like LDA.

In our work, we show that topic models exhibit aspects of memorization and successfully implement membership inference attacks against them. To combat memorization, we propose a modu-

lar framework for implementing better private topic models using differential privacy. This work highlights the greater need for continued development in the field of privacy-preserving ML, and informs current trends in research.

The direction of current research on memorization in ML focuses on exploring the extent that large deep learning models memorize their training data. One recent result shows that memorization increases with model scale (Carlini et al., 2023b). Although we do not fully analyze memorization in topic models, our results indicate even these smaller, simpler models show signs of memorization. We leave the full analysis of memorization in topic models to future research.

Memorization presents many challenges in ML. From a privacy perspective, memorization exposes user training data and creates vulnerabilities for the privacy attacks studied in this work. If simpler generative models memorize their training data, then it's only inevitable that large neural models memorize their training data.

Let's consider a parallel scenario presented by the development of the internet. Many problems in cybersecurity exist because the internet was designed without security in mind. Similarly, ML models were designed to create the most accurate predictions, or generalizations, without privacy in mind. Observing notions of memorization in simple probabilistic models developed nearly 20 years ago highlights our trend toward favoring utility in the fundamental trade-off between privacy and utility. As ML models continue to become more sophisticated and widely used, privacy concerns become increasingly relevant. Like the early days of the internet, the field of ML is now grappling with the challenges of securing technology that was not originally designed with privacy in mind.

A

One-Dimensional Query Statistics for Topic Models

In this appendix, we explore the many candidates for querying the learned topic model's topic-word distribution Φ. As mentioned in Section 3.2.2, the goal is to reduce Φ to an informative one-dimensional statistic given d to conduct. Under our adversarial assumptions defined in Section 3.2.1, the adversary can compute standard operations to reduce Φ or assume the generative struc-

ture and use Φ to infer the topic-word distribution for d and calculate natural likelihood estimates.

A.1 Requirements

To conduct the LiRA on topic models without modifying the underlying algorithm presented by Carlini et al., we have a few critical requirements (2021a). These requirements can apply to any model queried with an observation and can help define a principled approach to conducting the LiRA. The follow list of requirements outlines our one-dimensional statistic 'wish-list:'

1. The queried statistic has an intuitive interpretation related to topic models.
2. When document d is included the training data, the queried statistic increases on average.
3. The distribution of statistics queried on Φ is approximately normal and bound from $[-\infty, \infty]$.

The first requirement allows us reason about the why the attack may work. We opt to use metrics that mimic or follow natural likelihood estimates (the likelihood of observing the target document given the model Φ) because the interpretation allows us to reflect on memorization.

The next requirement is related to the first requirement, and is most important for the offline test. If we require that the queried statistic is greater on Φ when d is included in the training set, then under a natural likelihood measure we have have nice interpretation: the likelihood of the document increases when the document is in the training set. Additionally, this requirement is essential to the offline variant of the LiRA. Because the offline test statistic estimates a random draw from the estimated out-distribution on statisitcs ($\tilde{\mathbf{T}}_{out}$) is as high as the statistic on the target Φ. This is made more apparent by the test statistic in Algorithm 2 ($\Lambda \leftarrow 1 - \Pr[\mathcal{N}_{out} > \zeta(\Phi_{obs}, d)]$). This requirement is not necessary if we change direction of the test statistic in the offline test. However, our primary goal is to extend the LiRA as-is to topic models.

The final requirement allows us to implement parametric modeling in the LiRA. Although this requirement can be relaxed to "must follow a known distribution," estimating a different known distribution requires modifying the LiRA. Additionally, this requirement can be satisfied using simple data transformations. For instance, Carlini et al. use a logit-scaled predicted probability to enforce this constraint in their introduction of the LiRA (2021a).

A.2 Candidates

Most candidate statistics are based on the log-likelihood of the document under the model Φ. There are many methods to evaluate the log-likelihood of a held-out document given the model, but we must operate under the assumption that the adversary only has access to Φ for the target model. This limits our options for likelihood estimates. Additionally, because the adversary does not know for certain that the target Φ came from LDA, they must assume LDA's generative process to derive an informative statistic. Wallach et al. details many methods for estimating the likelihood of a held-out document (2009). We opt estimate the likelihood of the document by estimating the document's topic distribution θ and empirically evaluating the log-likelihood.

$$p(d|\Phi,\theta) = \sum_{w\in d} \log(\sum_{z} \theta_z * \Phi_{z,w}) \tag{A.1}$$

There are multiple methods for estimating θ under Φ. For instance Hofmann reference "folding-in" (1999). However, "folding-in" requires access to the model's underlying topic-word assignments. Assuming the adversary only has access to Φ, we estimate θ using a gibbs-sampler. First, we randomly sample topic assignments for each word $\mathbf{z}$, and estimate θ based on those counts. For each word in the document, we compute the probability of the word getting assigned to each topic given the document d, θ and Φ, and the topic assignments of all other words $\mathbf{z}_{\neg w}$:

$$p(z = w_k|d, \mathbf{z}_{\neg w}, \Phi, \Theta) \propto \Phi_{z,w_k} * \theta_k. \tag{A.2}$$

We sample a new topic for the current word from $p(z = w_k|d, \mathbf{z}_{\neg w}, \Phi, \Theta)$ and update $\mathbf{z}$ accordingly. We iterate through each word in the document (while shuffling words) until convergence or we reach a maximum number of iterations. This process tends to converge by 10 iterations. Finally, we can estimate Φ by counting and normalizing topic assignments in $\mathbf{z}$.

Once we estimate θ, we can estimate the log-likelihood of the document in accordance with Equation A.1. Another useful heuristic for quickly estimating the log-likelihood of the document is to treat θ as a fixed variable and maximize over all θ such that:

$$\zeta(\Phi, d) = \max_{\theta} \sum_{w \in d} \log(\sum_{z} \theta_z * \Phi_{z,w}). \tag{A.3}$$

Note that this our statistic ζ used in Chapter 3. While this heuristic is useful for estimating log-likelihood, but it is not a true unbiased estimate because we treat θ as a fixed variable rather not a random variable.

As another natural likelihood estimate, we can estimate the perplexity of the document. The perplexity of the document"is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood" (Blei et al., 2003). Once we calculate the log-likelihood of the document, the perplexity is very simple to implement given the number of words in the document C_d:

$$\textit{perplexity}(d|\Phi, \theta, C_d) = \exp \frac{-p(d|\Phi, \theta)}{C_d} \tag{A.4}$$

Finally, as a simple estimate for the log-likelihood of a document the adversary could attempt to estimate the *conditional log-likelihood* $p(d|z, \Phi)$. The conditional log-likelihood is essentially

the log-likelihood of the document under the assumption that each word in a document is only generated from one topic. Similar to the heuristic used in A.3, we can estimate this by maximizing over topics. This maximization is simple and does not require optimization over all possible topic-word distributions θ.

$$\xi(\Phi, d) = \max_{z} \sum_{w \in d} \log(\Phi_{z,w}) \tag{A.5}$$

A.3 Evaluation

To evaluate the candidate statistics, we run the experiment described in Section3.2.3. To reiterate the experiment proceeds as follows: first, we sample half of a data set $D_{train} \subset D$ and learn Φ using LDA such that $\Phi \leftarrow f_{\mathcal{M}}(D_{train})$. Then, we calculate $\zeta(\Phi, d)$ for target documents d and note if $d \in D_{train}$. We repeat the experiment many times to empirically estimate $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$. We repeat this experiment for each data set described in Section 3.3.1 by estimating 100 samples for each $\tilde{\mathbf{T}}_{x}(d)$ to replicate conducting the online LiRA with $N = 100$ shadow models. We conduct the experiment using 1000 random samples from each of the data sets described in Chapter 3.3.1.

As a simple metric to estimate the difference in our distributions we first inspect the difference in means of $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$ for each document. Figure A.1 displays the differences in means of $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$ for each document.

We can see from Figure A.1, we see that that the average value for each statistic except perplexity increases when the document is included in the training data for each document. Intuitively these results make sense because the lower the perplexity, the more likely the document is to be generated by the model. We see that perplexity will not satisfy requirement 2. However, we note that it could work well if we used the reciprocal or reversed the direction of the hypothesis test statistic in the LiRA.

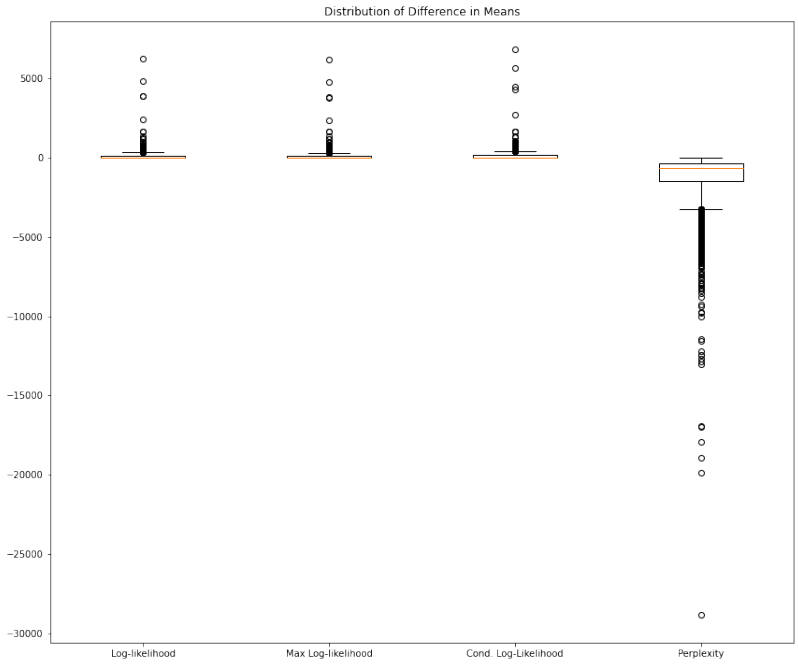

Figure A.1: Each Statistics' Difference in Mean for Each Document

Another metric to estimate the difference in distributions is the KL-divergence between normals estimated from $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$. We expect that the statistics with high KL-divergence between normals estimated from $\tilde{\mathbf{T}}_{x}(d)$ will be more vulnerable to the LiRA. Figure A.2 displays the KL-divergences $\tilde{\mathbf{T}}_{in}(d)$ and $\tilde{\mathbf{T}}_{out}(d)$ for each document.

Based on KL-divergence, it appears that either the log-likelihood or the max log-likelihood estimate $\zeta(\Phi, d)$ would make good statistics for an attack. While it appears that the conditional log-likelihood estimate has more outliers and may out-perform the other statistics, the overall distribution of KL-divergences is much smaller compared to the other statistics. We note that perplexity appears to perform the best in terms of KL-divergence, but does not satisfy requirement 2.

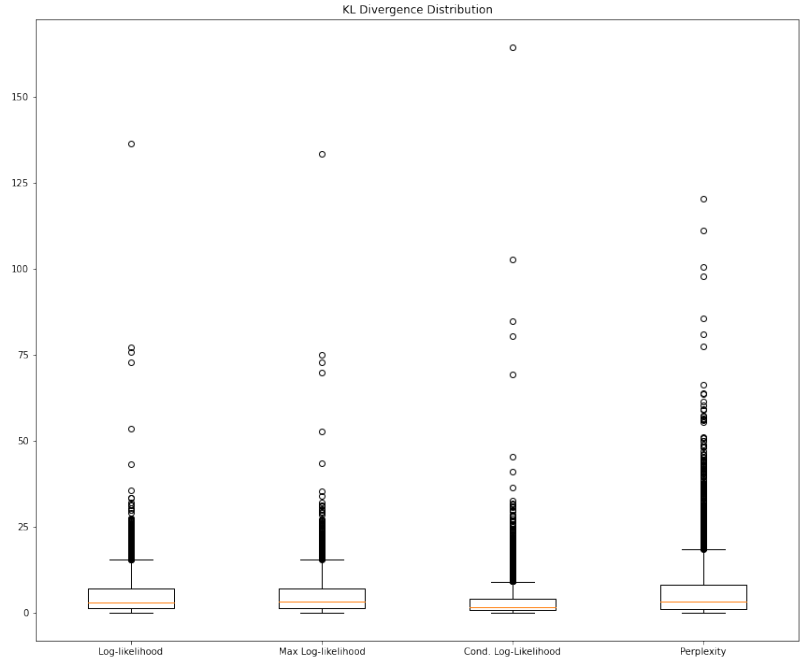

Figure A.2: Each Statistics' KL Divergence for Each Document

A.4 Discussion

After running the experiment outlined in section A.3, we determined that the statistic ζ is best fit for the LiRA. Compared to the conditional log-likelihood statistic ξ, our heuristic ζ can capture document's complex topic structure better by imitating LDA's generative process. Therefore, ζ realizes better performance on long documents with many words from various topics, and benefits from interpretation in relation to LDA.

Classical likelihood measures like perplexity and log-likelihood seem to satisfy the interpretability requirement. However, perplexity does not satisfy requirement 2 because when the document d is included in the training data, the perplexity decreases. Once again, we note that perplexity could work well as a statistic for the LiRA if we used the reciprocal or reversed the direction of the hypothesis test statistics.

The log-likelihood seems to be a suitable candidate. However, there are many methods of calculating the log-likelihood (Wallach et al., 2009). The stochastic method we used to calculate the log-likelihood estimate without the underlying topic-word counts may be snot robust or stable for long documents.

The statistic ζ estimates the log-likelihood of the document using deterministic optimization. We choose ζ for the LiRA because it satisfies all the 3 requirements in section A.1, and does not rely on stochastic methods of estimating the log-likelihood. However, we note that the LiRA could easily be modified to work with the log-likelihood given an accurate, consistent likelihood estimate.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318).: ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022). What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (pp. 2280–2292). New York, NY, USA: Association for Computing Machinery.

Bun, M. & Steinke, T. (2016). Concentrated differential privacy: simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference* (pp. 635–658).: Springer.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramèr, F. (2021a). Membership inference attacks from first principles. *CoRR*, abs/2112.03570.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023a). Extracting training data from diffusion models.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023b). Quantifying memorization across neural language models.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)* (pp. 267–284). Santa Clara, CA: USENIX Association.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021b). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633–2650).: USENIX Association.

Carvalho, R. S., Wang, K., & Gondara, L. S. (2022). Incorporating item frequency for differentially private set union. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9), 9504–9511.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, volume 22: Curran Associates, Inc.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15, 429–222.

Decarolis, C., Ram, M., Esmaeili, S., Wang, Y.-X., & Huang, F. (2020). An end-to-end differentially private latent Dirichlet allocation using a spectral algorithm. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (pp. 2421–2431).: PMLR.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41, 391–407.

Dua, D. & Graff, C. (2017). UCI machine learning repository.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay (Ed.), *Advances in Cryptology - EUROCRYPT 2006* (pp. 486–503). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C. & Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09 (pp. 371–380). New York, NY, USA: Association for Computing Machinery.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Lecture notes in computer science*, Lecture Notes in Computer Science (pp. 265–284). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C. & Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 93–107.

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4), 211–407.

Dwork, C., Rothblum, G. N., & Vadhan, S. (2010). Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* (pp. 51–60).

Dwork, C., Smith, A., Steinke, T., Ullman, J., & Vadhan, S. (2015). Robust traceability from trace amounts. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*.

Feldman, V. (2021). Does learning require memorization? a short tale about a long tail.

Feldman, V. & Zhang, C. (2020). What neural networks memorize and why: Discovering the long tail via influence estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 Red Hook, NY, USA: Curran Associates Inc.

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15 (pp. 1322–1333). New York, NY, USA: Association for Computing Machinery.

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., & Ristenpart, T. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14 (pp. 17–32). USA: USENIX Association.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18 (pp. 619–633). New York, NY, USA: Association for Computing Machinery.

Gao, F. & Han, L. (2012). Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1), 259–277.

Gopi, S., Gulhane, P., Kulkarni, J., Shen, J. H., Shokouhi, M., & Yekhanin, S. (2020). Differentially private set union. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (pp. 3627–3636).: PMLR.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1), 5228–5235.

Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2019). Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019, 133–152.

Hilprecht, B., Härterich, M., & Bernau, D. (2019). Monte carlo and reconstruction membership inference attacks against generative models. volume 2019.

Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, volume 23: Curran Associates, Inc.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99 (pp. 50–57). New York, NY, USA: Association for Computing Machinery.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4(8), e1000167.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, volume 34 (pp. 2018–2033).: Curran Associates, Inc.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s), 1–37.

Huang, T. & Chen, H. (2021). Improving privacy guarantee and efficiency of Latent Dirichlet Allocation model training under differential privacy. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 143–152). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Huang, T., Zhao, S., Chen, H., Liu, J., & Xu, Y. (2022). Improving parameter estimation and defensive ability of latent dirichlet allocation model training under rényi differential privacy. *Journal of Computer Science and Technology*, 37(6), 1382–1397.

Jayaraman, B., Wang, L., Evans, D., & Gu, Q. (2021). Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2021(3), 32–53.

Korolova, A., Kenthapadi, K., Mishra, N., & Ntoulas, A. (2009). Releasing search queries and clicks privately. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09 (pp. 171–180). New York, NY, USA: Association for Computing Machinery.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better.

Leino, K. & Fredrikson, M. (2020). Stolen memories: Leveraging model memorization for calibrated White-Box membership inference. In *29th USENIX Security Symposium (USENIX Security 20)* (pp. 1605–1622).: USENIX Association.

Li, Z. & Zhang, Y. (2021). Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21 (pp. 880–895). New York, NY, USA: Association for Computing Machinery.

Liu, J., Oya, S., & Kerschbaum, F. (2021). Generalization techniques empirically outperform differential privacy against membership inference. *ArXiv*, abs/2110.05524.

Long, Y., Wang, L., Bu, D., Bindschaedler, V., Wang, X., Tang, H., Gunter, C. A., & Chen, K. (2020). A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 444–459).: IEEE.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11 (pp. 262–272). USA: Association for Computational Linguistics.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)* (pp. 263–275).

Murshed, B. A. H., Mallappa, S., Abawajy, J., Saif, M. A. N., Al-ariki, H. D. E., & Abdulwahab, H. M. (2022). Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*.

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 739–753).

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706), 289–337.

Nissim, K., Raskhodnikova, S., & Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC'07*, Proceedings of the Annual ACM Symposium on Theory of Computing (pp. 75–84). STOC'07: 39th Annual ACM Symposium on Theory of Computing ; Conference date: 11-06-2007 Through 13-06-2007.

Parisot, M. P. M., Pejo, B., & Spagnuelo, D. (2021). Property inference attacks on convolutional neural networks: Influence and implications of target model's complexity. *CoRR*, abs/2104.13061.

Park, M., Foulds, J., Chaudhuri, K., & Welling, M. (2018). Private topic modeling.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08 (pp. 569–577). New York, NY, USA: Association for Computing Machinery.

Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., Jones, M., Klein, A., Milechin, L., Mullen, J., Prout, A., Rosa, A., Yee, C., & Michaleas, P. (2018). Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)* (pp. 1–6).: IEEE.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04 (pp. 487–494). Arlington, Virginia, USA: AUAI Press.

Ruiz de Arcaute, G. M., Hernández, J. A., & Reviriego, P. (2022). Assessing the impact of membership inference attacks on classical machine learning algorithms. In *2022 18th International Conference on the Design of Reliable Communication Networks (DRCN)* (pp. 1–4).

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., & Jegou, H. (2019). White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning* (pp. 5558–5567).: PMLR.

Salem, A., Zhang, Y., Humbert, M., Fritz, M., & Backes, M. (2019). ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed Systems Security Symposium*: Internet Society.

Schofield, A., Thompson, L., & Mimno, D. (2017). Quantifying the effects of text duplication on semantic models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2737–2747).

Shalit, U., Weinshall, D., & Chechik, G. (2013). Modeling musical influence with topic models. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research* (pp. 244–252). Atlanta, Georgia, USA: PMLR.

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)* (pp. 3–18).

Sievert, C. & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Baltimore, Maryland, USA: Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958.

Thakurta, A. G. & Smith, A. (2013). Differentially private feature selection via stability arguments, and the robustness of the lasso. In S. Shalev-Shwartz & I. Steinwart (Eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research* (pp. 819–850). Princeton, NJ, USA: PMLR.

Tramèr, F., Kamath, G., & Carlini, N. (2022). Considerations for differentially private learning with large-scale public pretraining.

U.S. Census Bureau (2021). Disclosure avoidance for the 2020 census: An introduction. https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.htmls.

Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2020). Interoperable pipelines for social cyber-security: Assessing twitter information operations during nato trident juncture 2018. *Comput. Math. Organ. Theory*, 26(4), 465–483.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (pp. 1105–1112). New York, NY, USA: Association for Computing Machinery.

Wang, H., Sharma, J., Feng, S., Shu, K., & Hong, Y. (2022a). A model-agnostic approach to differentially private topic mining. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1835–1845).

Wang, K.-C., Fu, Y., Li, K., Khisti, A., Zemel, R., & Makhzani, A. (2022b). Variational model inversion attacks.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63–69. PMID: 12261830.

Wilson, R. J., Zhang, C. Y., Lam, W., Desfontaines, D., Simmons-Marengo, D., & Gipson, B. (2019). Differentially private sql with bounded user contribution.

Xun, G., Li, Y., Zhao, W. X., Gao, J., & Zhang, A. (2017). A correlated topic model using word embeddings. In *IJCAI*, volume 17 (pp. 4207–4213).

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, (pp. 268–282).

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., & Mironov, I. (2022). Opacus: User-friendly differential privacy library in pytorch.

Zhao, F., Ren, X., Yang, S., Han, Q., Zhao, P., & Yang, X. (2021a). Latent dirichlet allocation model training with differential privacy. *IEEE Transactions on Information Forensics and Security*, 16, 1290–1305.

Zhao, F., Ren, X., Yang, S., Han, Q., Zhao, P., & Yang, X. (2021b). Latent dirichlet allocation model training with differential privacy. *IEEE transactions on information forensics and security*, 16, 1290–1305.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021c). Topic modelling meets deep neural networks: A survey.

Zhao, W., Zou, W., & Chen, J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. *BMC bioinformatics*, 15 Suppl 11, S11.

Zhu, T., Li, G., Zhou, W., Xiong, P., & Yuan, C. (2016). Privacy-preserving topic model for tagging recommender systems. *Knowledge and information systems*, 46(1), 33–58.

Zubiaga, A., Liakata, M., & Procter, R. (2016). Learning reporting dynamics during breaking news for rumour detection in social media. *CoRR*, abs/1610.07363.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.