Project Proposal
PPOL 6801: Text as Data
Muhammad Saad (Net ID: ms4689)
Visualizing Privacy-Utility Tradeoffs in Differentially Private Topic Models

Problem Statement and Motivation

Latent Dirichlet Allocation (LDA) is an extensively used machine learning technique for text data applications. It allows researchers to discover latent themes within a corpus and to represent documents as a mixture of topics inferred from their word distributions. Because of its generative nature, LDA has become particularly useful for natural language processing (NLP) applications, including text summarization and topic retrieval. However, recent research has shown that there are tangible privacy risks associated with LDA. For example, Carlini et al. (2023) have shown that probabilistic large language models can *memorize* training data. More specifically, Manzonelli et al. (2024) have clearly demonstrated that these memorization risks can also be extended to LDA and are particularly susceptible to membership inference attacks, where an attacker can successfully deduce whether an individual's data was used to generate the LDA's output statistics or not.

Differential Privacy (DP) offers a way to not only quantify the privacy risks associated with LDA, but also protect against them by introducing calibrated noise at various points of the algorithm to perturb the outputs generated and mask individual contributions. The application of DP within LDA has grown over the past five years, where research has differed along three main dimensions: i). where noise is injected (topic-word distribution, document-topic distribution, or vocabulary); ii). how sensitivity is bounded (that is maximum contribution that an individual can make to the input data matrix); iii). what privacy accounting framework (pure-DP, approximate-DP or Renyi-DP) is used. Regardless of the various techniques adopted, the literature currently lacks a comparative perspective on how these different design decisions affect interpretability, utility, and computational efficiency associated with LDA.

This project aims to bridge that gap by simulating four major differentially privacy LDA approaches under identical conditions and visualizing their sensitivity-utility trade-offs. The motivation is two-fold: Firstly, to deepen conceptual understanding of both LDA and differential privacy algorithms. Secondly, to build a practical tool that helps analysts select an appropriate DP mechanism for their own analysis.

Data Source and Acquisition

The project will use the 20 Newsgroups dataset accessible here (scikit-learn developers, 2017), which contains around 18,000 discussion-board posts across 20 categories (politics, science, sports, religion, etc.). The dataset has been chosen because it is easy to access, has sufficient variability and amenability for qualitative interpretation, and is also used in two of the DP-LDA applications that the author wants to simulate (Huang and Chen, 2021; Manzonelli et al., 2024). It can be easily accessed through the scikit-learn environment. The author will include some light preprocessing for it, including tokenization, lowercasing, and stop-word removal.

Existing Research

Research on differentially private topic modeling has evolved along three major areas. Firstly, the type of privacy guarantee (that is, either pure-, approximate- or Renyi-DP), the unit of privacy protection (that is, whether the algorithm gives word-, document-, or author-level privacy), and

the learning algorithm (Collapsed Gibbs Sampling, Variational Inference, or Spectral Approximation). A detailed overview of the existing research has been included in Appendix A of this document.

These studies demonstrate that differentially private topic models differ not only in their mathematical definitions of privacy but also in where they inject noise and how they calculate sensitivity. This project will simulate four representative approaches, which are by Zhao (2020), Huang & Chen (2021), Huang et al. (2022), and Manzonelli et al. (2024), under identical preprocessing and dataset to evaluate how noise placement and sensitivity bounds impact utility and model stability. An overview of the four approaches that the project will simulate is included in Table 1.

Table 1: Selected Studies for Simulation

Authors (Year)	Method	Feature	Privacy Level / Mechanism	Data Source	GitHub Repository
Zhao et al. (2020)	DP-LDA via Collapsed Gibbs Sampling	Adds Laplace noise to word–topic counts at each iteration;	word-level, Laplace mechanism	Wikipedia subset, Enron Emails	Available
Huang & Chen (2021)	Subsampled Laplace DP (SUB-LDA)	Introduces Poisson subsampling before noise addition	word-level, Laplace subsampling	20 Newsgroups	Available
Huang et al. (2022)	Rényi-DP LDA	Replaces Laplace with Gaussian mechanism	word-level, Gaussian mechanism	Reuters dataset	Available
Manzonelli et al. (2024)	Vocabulary-level DP (DP Set Union)	Applies DP to vocabulary construction (feature selection) in addition to training	vocabulary-level, DP Set Union	20 Newsgroups	Available

Methodology and Analysis Plan

The project will re-implement simplified versions of the four DP-LDA algorithms using the same preprocessing pipeline and topic number (to be decided by the author). Privacy budgets will vary from epsilon equal to 0.1 (highest privacy protection) to 10 (lowest privacy protection with a fixed delta (δ). As far as possible, the author will also try to keep the clipping constant (C) same across all simulations. The project's evaluation will cover:

1. **LDA Utility:** Topic coherence
2. **Stability / Sensitivity:** Change in topic–word distributions, specifically the Φ parameter
3. **Efficiency:** Runtime and computational ease

The project's visualization outputs will include:

1. Privacy–utility curves which show changes in the privacy protection parameter and corresponding change in topic coherence
2. Heatmaps that depict changes in per-topic word distributions or changes in the Φ parameter
3. A comparative dashboard with interactive sliders for changes in the privacy parameter, the sensitivity bounds and the algorithm type.

Project Output and Success

The analytical output of this project will be an interactive decision-support tool that allows data analysts to visualize how differentially private-LDA mechanisms behave under varying privacy budgets and sensitivity thresholds. At the back-end will be python scripts which reproduce the four DP-LDA algorithms on the same preprocessed 20 Newsgroups corpus, with tunable parameters for the level of noise introduced, the sensitivity contribution, the choice of the learning algorithm, as well as the privacy budget accounting framework. The project output will include live visuals which display outputs like topic coherence vs. level of noise, and Φ change heatmaps. Overall, the aim is that analysts without a DP background should be able to use the tool to choose between DP mechanisms based on dataset size, sensitivity, and interpretability requirements. It will also be interesting to see whether some differential privacy noise mechanisms for LDA can actually improve topic stability or coherence, as a way to regularize LDA modelling.

Appendix A: *Summary of Existing Literature on DP Topic Modeling*

Authors	Notion of DP	Adjacency Definition	Learning Method	Other Technical Details
Zhu et al. (2016)	Pure-DP	Document-level	Collapsed Gibbs Sampling (CGS)	Early baseline DP-LDA; adds Laplace noise to word–topic counts; sensitivity bounded by per-doc token limit
Park et al. (2018)	Approximate-DP	Document-level	Variational Inference (VI)	Uses Moments Accountant and sub-sampling for tighter bounds; global sensitivity derived from doc contributions
Zhao et al. (2020)	Pure-DP	Word-level	CGS	Introduces hierarchical DP-LDA variants (HDP-LDA, LP-LDA, OLP-LDA); Laplace noise on word–topic counts
Huang & Chen (2021)	Approximate-DP	Word-level	CGS	Sub-sampled Laplace DP (SUB-LDA); exploits privacy amplification by sampling to reduce effective ε
Huang et al. (2022)	Rényi-DP	Word-level	CGS	RDP accounting with truncated Gaussian noise; non-negative projection for stability
DeCarolis et al. (2020)	Rényi-DP	Document-level	Spectral Approximation (SA)	Adds Gaussian noise to second-/third-order tensor moments; uses propose-test-release for local sensitivity
Manzonelli et al. (2024)	Approximate-DP	Vocabulary-level	CGS + DP Set Union	Protects vocabulary inclusion probabilities; splits ε across vocab and training phases

References

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). *Quantifying memorization across neural language models.* In *Proceedings of the International Conference on Learning Representations (ICLR 2023).* https://arxiv.org/abs/2302.07800

DeCarolis, P., Gaboardi, M., & Vadhan, S. (2020). *Rényi differential privacy mechanisms for tensor decomposition and topic modeling.* In *Proceedings of the Conference on Privacy in Machine Learning (NeurIPS Workshop).*

Huang, Y., & Chen, X. (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. 159–169). Association for Computational Linguistics. https://aclanthology.org/2021.findings-emnlp.14.pdf

Huang, Y., Chen, X., & Li, Y. (2022). *Improving parameter estimation and defensive ability of LDA under Rényi differential privacy. Entropy, 24*(2), 187. https://www.sciopen.com/article/10.1007/s11390-022-2425-x

Manzonelli, A., Zhang, J., & Vadhan, S. (2024). *Membership inference attacks and privacy in topic modeling. Transactions on Machine Learning Research.* https://openreview.net/pdf?id=NmWp5lFL7L

Park, M., Foulds, J. R., Chaudhuri, K., & Welling, M. (2016). *Private topic modeling*. arXiv. https://arxiv.org/abs/1609.04120

scikit-learn developers. (2017). *The 20 newsgroups text dataset.* scikit-learn documentation. https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

Zhao, T., Zhang, M., & Wang, J. (2020). *Latent Dirichlet allocation model training with differential privacy. Entropy, 22*(11), 1216. https://arxiv.org/pdf/2010.04391

Zhu, T., Li, G., Zhou, W., & Phillips, P. (2016). *Differentially private data publishing and analysis: A survey. IEEE Transactions on Knowledge and Data Engineering, 29*(8), 1619–1638. https://dl.acm.org/doi/abs/10.1109/tkde.2017.2697856