Undergraduate Research Opportunity Research (UROP) Project Report

Community Interpretation in Network Data with Covariates

## Abstract

## 1. Introduction

Community detection traditionally relies on network topology, but textual metadata can offer deeper insights into node similarities. Our core research question concerns how well advanced topic modeling algorithms can extract semantically meaningful structures, which could later inform or enrich community detection. In particular, we chose to implement hLDA for its ability to grow the topic structure organically through the nested Chinese Restaurant Process (nCRP). This approach enables the model to adapt to the complexity of the corpus without fixing the number of topics in advance. While community detection remains an intended application, this report primarily details the rationale, implementation, and preliminary outcomes of hLDA on a well-known text dataset.

## 2. Literature Review

Early methods like Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) applied matrix factorization (Singular Value Decomposition) to capture latent structures in the term-document matrix. However, LSA lacks a fully probabilistic interpretation, prompting the introduction of Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), which models each document as a mixture of latent topics, each represented by a word distribution. Extensions such as the Author-Topic Model (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) further enrich LDA by capturing how authors influence topic distributions. Hierarchical LDA (hLDA) (Blei, Griffiths, Jordan, & Tenenbaum, 2004) extends this framework by organizing topics into a tree structure whose depth and branching expand as needed. Compared to LSA, these probabilistic approaches offer clear interpretability of topic-word distributions and handle larger, more complex datasets with a flexible number of topics.

## Background

## Introduce hLDA, NCPR, etc…

## 3. Methodology

## 3.1 Implementation

We utilized a Python environment to implement hLDA, leveraging existing libraries or custom scripts for the nested CRP. Given the dataset’s size, computational complexity is higher than standard LDA, as each document contributes to the branching structure. The main steps include:

• **Data Loading**: Reading the 20 Newsgroups text files into memory and performing preprocessing.

• **Vocabulary Construction**: Creating a vocabulary of terms above a certain frequency threshold.

• **Model Initialization**: Setting hyperparameters, initializing topic assignments, and establishing a hierarchical structure.

• **Inference and Monitoring**: Running iterative sampling, monitoring convergence by inspecting topic distribution changes, and adjusting parameters if needed.

## 3.2 Data

### 3.2.1 The 20 Newsgroups Dataset

We use the 20 Newsgroups dataset, a well-known collection of roughly 20,000 newsgroup documents organized into 20 distinct categories. It is commonly used as a benchmark in text classification and clustering research. Each document is associated with a single newsgroup topic, covering diverse domains such as sports, politics, and technology. This makes 20 Newsgroups an ideal testbed for evaluating how well hierarchical topic modeling captures thematic structure.

• **Data Size and Format**: Approximately 20,000 documents split across the 20 categories.

• **Preprocessing**: We performed standard text cleaning steps, including tokenization, stop-word removal, and either stemming or lemmatization.

• **Representation**: Each document was converted into a bag-of-words format suitable for input to the hLDA algorithm.

### 3.2.2 Synthetic datasets

## 3.3 Results and Analysis

### 3.3.1 Time complexity

### 3.3.2 Hyperparameters

## 4. Discussion

Talk about the linkage with community detection?

## 5. Conclusion

## 7. References

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems, 16*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3(Jan), 993–1022.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6), 391–407.*

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI 2004), 487–494.*