Undergraduate Research Opportunity Research (UROP) Project Report

Community Interpretation in Network Data with Covariates

## Abstract

## 1. Introduction

**Motivation:** Adopting topic modelling techniques in network data.

**Current stage:** Exploring hLDA more concretely.

## 2. Literature Review

### 2.1 Background

Previous models (LSA, pLSA, LDA, author-LDA…)

Other paper that talks about integration of topic modelling and community detection?

## 3. Introduce hLDA

1. nCRP

Formula

2. hLDA

Symbols and definitions

Generative process

3. Gibbs sampling process

Path sampling, z

Level sampling, c

4. Hyperparameter

Effects on the result and why (base on the formula)

## 4. Implementation

1. Code?

2. Preprocessing steps of the corpus.

## 5. Result

### 5.1 The 20 Newsgroups Dataset

1. Dataset description

2. Test results:

i. Speed

ii. Tree structure.

iii. Examine if the hierarchy make sense. i.e. more general words at the root and more specific words at the leaves.

### 5.2 Synthetic datasets

1. Synthetic data generation

1.1. Run a hLDA model on some corpus and extract its posterior tree with the settings:

* L = 3
* Alpha = …

1.2. Draw 10 corpus:

* Each data sets is made up of 1000 document of 250 words.
* Take a subset of the posterior tree and generate each document base on this tree.
  + The proportion of word allocated to each level is simple based on a Dirichlet distribution (parameter = ?)

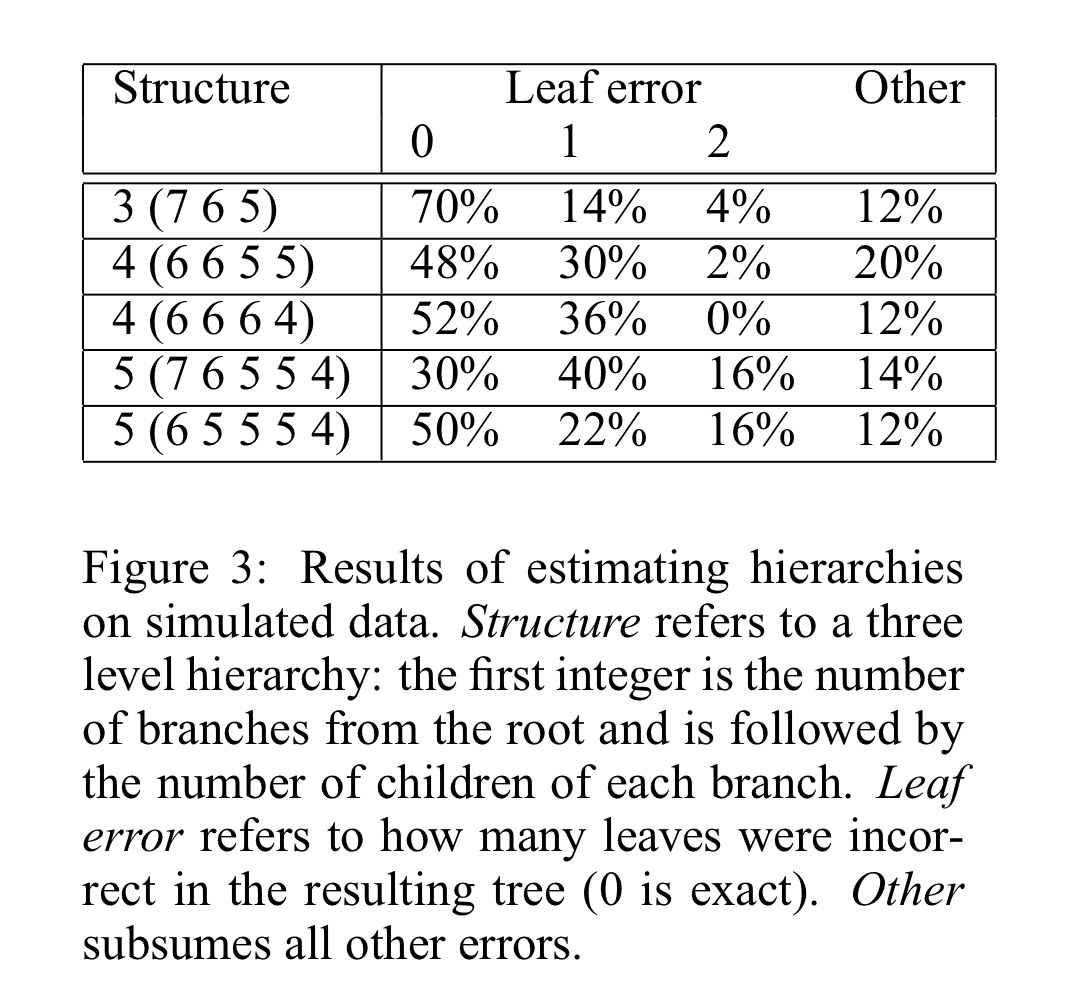
2. Test and analysis of result:

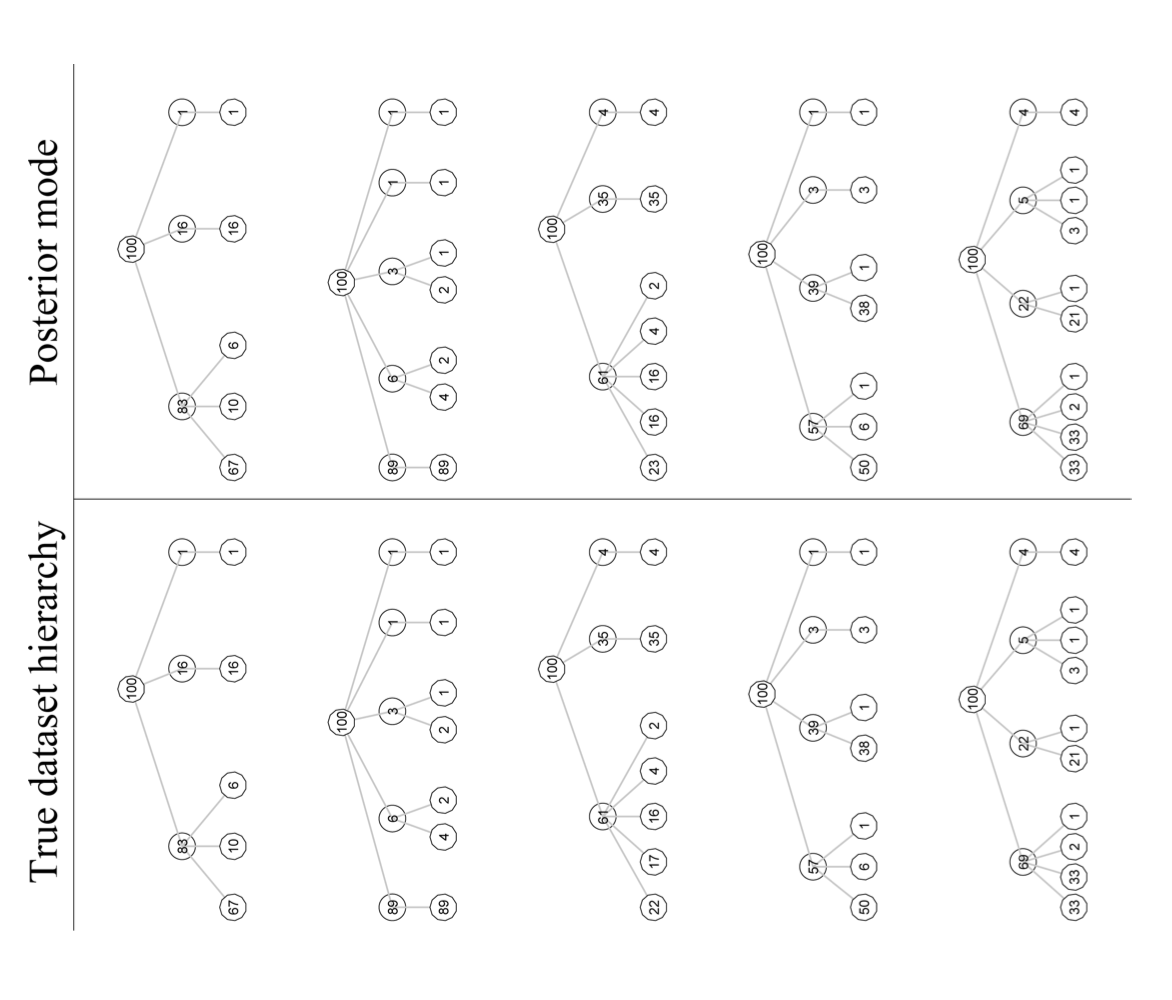
2.1 Rerun a new hLDA model on the synthetic data and extract the posterior tree

2.2 Compare the new posterior tree with the ‘ground truth’ tree

2.2.1 Compare the entire tree first. See if the same structure is obtained.

2.2.2 Examine the similarity of the nodes base on the top-n words.





### 5.3 Hyperparameter testing

1. gamma

2. alpha

3. eta

## 6. Discussion

Talk about the linkage with community detection + further consideration for hLDA.

## 7. Conclusion

## 8. References