Week 10 Progress Report

Zhang Mingyuan

March 24, 2025

1 Introduction to the Updated Network

This week, I have extended my previous work by enhancing the document network with additional coauthorship edges. The resulting network thus combines both citation relationships and coauthor interactions.

Updated Network Summary:

- Nodes represent documents (abstracts).
- Edges represent citation and coauthorship.
- **node count:** 3206, **edge count:** 21356 where 15687 of them are coauthorship edges

2 Characteristics of the Network

After constructing the enhanced network, I run SBM and applied filtering to remove small communities (less than 30 nodes) to focus on significant clusters:

- **Nodes after filtering:** 2670
- **Edges after filtering:** 12020
- **Total communities identified:** 41

Community Size Statistics:

- Minimum community size: 30
- Maximum community size: 401
- Median community size: 58.0
- Average community size: 65.12

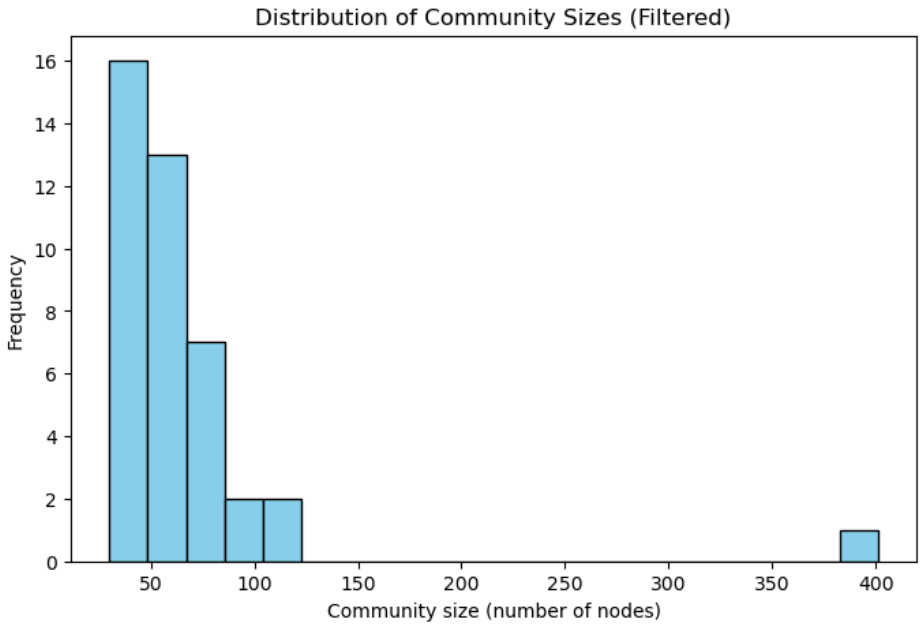

Figure 1: Distribution of communities

3 Detailed Analysis of Largest Community

The largest community identified (Community 942) was analyzed in depth:

- **Community size:** 401 nodes
- **Intracommunity edges:** 375
- **Word count statistics (abstracts):**
  - Minimum: 22 words
  - Maximum: 378 words
  - Mean: 152.40 words

Example Abstract Titles:

- Robust recovery of multiple subspaces by geometric methods
- Erich Leo Lehmann—a glimpse into his life
- Bayesian analysis of variable-order, reversible Markov chains
- On optimality of the Shiryaev-Roberts procedure
- Kernel density estimation via diffusion

4 Corpus Preprocessing

Corpus preprocessing remained consistent:

1. Tokenized text (converted to lowercase).
2. Removed abstracts with fewer than 20 words.
3. Applied lemmatization.
4. Removed stop words.
5. Generated Bag-of-Words representation.

Number of unique vocabulary: 9743

5 Latent Dirichlet Allocation (LDA)

LDA was applied to the entire corpus with six topics (k = 6). Below are the inferred topics based on top words:

Inferred Topics:

- **Topic 0: Multivariate and Functional Analysis**
  - *Keywords:* model, matrix, covariance, component, spatial, regression, linear, functional
- **Topic 1: Bayesian and Causal Modeling**
  - *Keywords:* prior, Bayesian, posterior, causal, inference, treatment, likelihood, Dirichlet
- **Topic 2: Statistical Estimation Techniques**
  - *Keywords:* estimator, estimation, regression, asymptotic, likelihood, variance, nonparametric
- **Topic 3: Survival Analysis and Time-dependent Models**
  - *Keywords:* hazard, survival, time, conditional, regression, simulation, probability
- **Topic 4: Experimental Design and Sampling Methods**
  - *Keywords:* design, algorithm, treatment, Markov, Monte Carlo, sampling, optimal
- **Topic 5: Hypothesis Testing and Multiple Comparisons**
  - *Keywords:* test, hypothesis, procedure, statistic, control, false discovery rate, simulation

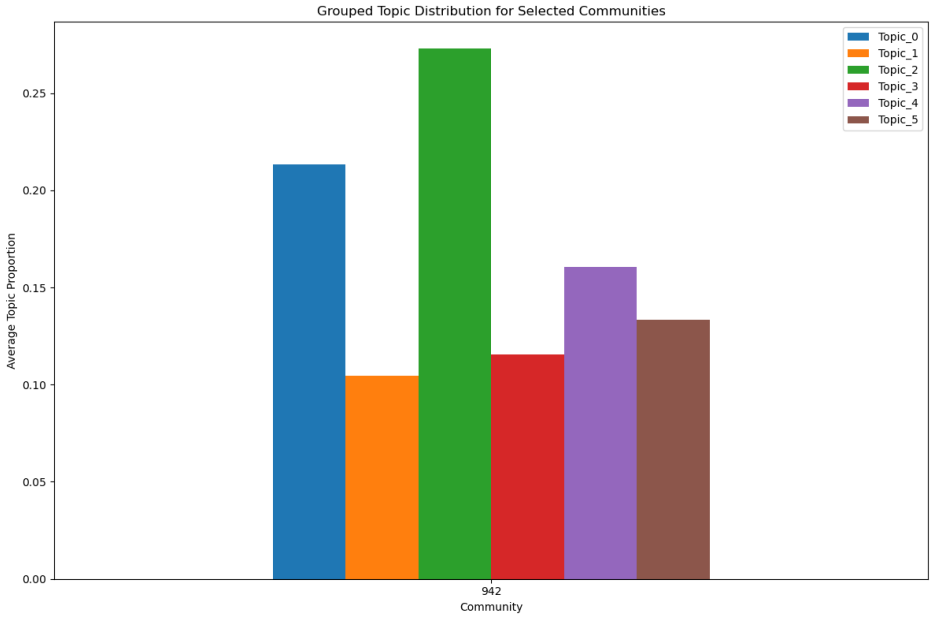

Figure 2: LDA result for community 942

6 Hierarchical LDA (hLDA)

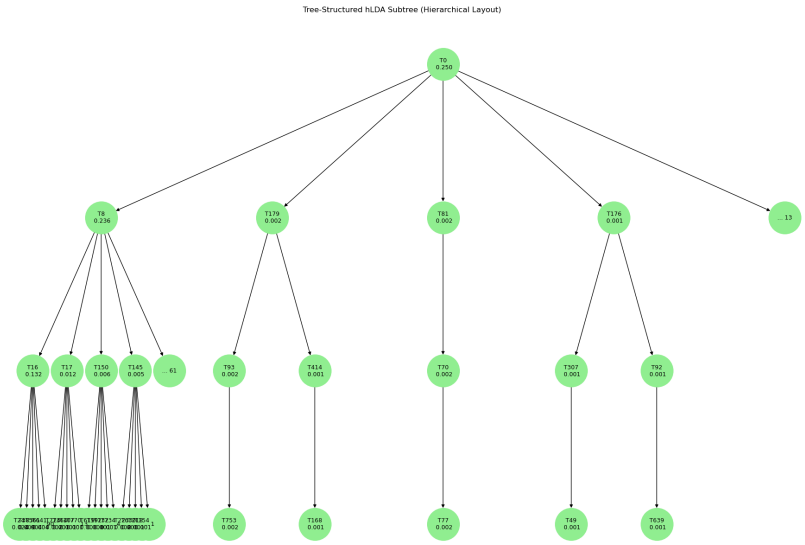

Figure 3: hLDA result for community 942

Selected Topic Summary

Below is a summary of selected topics extracted from different levels of the hLDA tree:

Level 0

- **Topic 0:** General research methodology — model, method, data, analysis, simulation, parameter

Level 1

- **Topic 121:** Geometric and structural modeling — structure, geometric, definition, maximal, compatible
- **Topic 14:** Banking efficiency and systems — inefficiency, cost, bank, allocative, technical, firm
- **Topic 176:** Aging and survival analysis — incidence, aging, cohort, mortality, survival, survey
- **Topic 124:** Community-level epidemiology — individual, vaccine, susceptibility, vaccination, household, outbreak
- **Topic 179:** Missing data and imputation — imputation, missing, fractional, multiple, estimator, value
- **Topic 87:** Experimental design and causal inference — experiment, outcome, randomized, bias, adjustment

Level 2

- **Topic 112:** Theoretical probability and computability — word, computable, mathematical, false
- **Topic 415:** Generic statistical modeling terms — model, data, method, estimator, test, regression
- **Topic 307:** Causal assumptions in data procedures — whereby, based, likelihood, proposed
- **Topic 92:** Time series and stationarity — sullivan, stationarity, centered, rapidly, resolve

- **Topic 19:** Homogeneity in statistical models — minor, homogeneous, estimation, procedure
- **Topic 414:** Public data and computational access — publicly, computational, distribution, likelihood
- **Topic 93:** Repetition of base modeling terms (possibly generic/default topic) — model, method, test
- **Topic 69:** Weighting and stability in models — summarized, weighting, stable, poorly

Level 3

- **Topic 36:** Visual perception and principles — principle, image, alignment, meaningful, prove
- **Topic 755:** Estimation technique (possibly least squares) — least, method, data, regression
- **Topic 49:** Biodiversity and species occurrence — specie, biological, community, richness, composition
- **Topic 639:** Life expectancy and demographic studies — life, disability, expectancy, rate, assumption
- **Topic 63:** Infectious disease modeling — contact, infection, vaccinated, surface, transmission
- **Topic 168:** Variable treatment in software/surveys — variable, location, augmentation, implemented
- **Topic 753:** Survey bias and population health — nonresponse, selection, bias, bmi, nhanes
- **Topic 720:** Educational interventions and outcomes — school, child, treatment, peer, assigned