A Network Perspective on LDA

Cindy Cook
Network Analysis Final Project
April 2015

---

1. Introduction

With an increase in the amount of digitized text, topic modeling has become an important area of research. One of the most popular methods for topic modeling is Latent Dirichlet Allocation (LDA) first described in [1]. LDA, a probabilistic hierarchical generative model, uses the Bayesian framework to discover hidden structure, or *topics*, within a set of documents referred to as a corpus. Given a document, the distribution of latent topics is found through estimation of the posterior distribution. Although LDA is commonly used, the methods are called into question by Lancichinetti in [2]. More info here...

This paper begins by describing a data set, referred to as Science using both descriptive and network statistics, that can be found in [2]. It then replicates the main findings from [2] involving the Science dataset. Lastly, we focus on expanding the study by...

1.1. The Data

All data and scripts can be found on github here . There are two main 'real world' datasets used to test the TopicMapping algorithms: Science and Wikipedia. The Science corpus contains $28,838$ documents from the web of science, where each document contains the title and abstract of a paper published in one of six top journals from different disciplines. After preprocessing the data, there are $106,143$ unique words and over 8.7×10^6 edges. The Wikipedia dataset contains over 4×10^6 nodes with over 800×10^6 edges. Due to the size of the Wikipedia data set, I was unable to receive this information from the authors. Therefore, I will focus on the Science data set only. Although the Science corpus was easily sent to me, it is still considered 'large.' In R, the adjacency matrix is too large to be stored. In particular, the program crashes while trying to plot the network using the plot function in the network package. In order to visually see the network, I used only the first 10 documents to create a network.

What we immediatley see in this plot are very distinct clusters along with very distinct bridges. Explain what this is informing us here... The article also makes use of several synthetic datasets: list and explain here for what I will use...

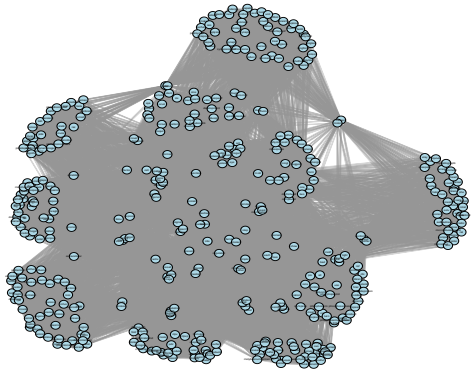

Figure 1: Network Plot of the first 10 Documents

2. Replicate the Study

3. Further the Study

4. Conclusions

4.1. Discussion

4.2. Future Directions

5. Bibliography

[1] Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation." *Journal of Machine Learning Research*: 3 993-1022.

[2] Lancichinetti, A., Sirer, M., Wang, J., Acuna, D., Kording, K., and Amaral, Luis. (2015), "High-Reproducibility and High-Accuracy Method for Automated Topic Classification." *Physical Review X*: 5, 0011007, 2160-3308.