Proposal for BDSS Research Rotation (2015-2016)

Trainee: Cindy Cook
Host: Dave Hunter
Meetings: Wednesdays 9:30-10:30
Monthly Reports: Will be written by the trainee and turned in the last week of each month.

Purpose: This rotation was inspired by a project encountered from the trainee's summer externship experience. The problem involved matching and ranking documents from a corpus to a given document outside of the corpus based on similarity along with other potential covariates. The corpus consists of unstructured text documents. The number of topics that the documents could have in common is not known, and the 'bag of words' assumption may not be correct. For the 2015-2016 research rotation, the ultimate goal is to understand both the advantages and limitations of current probabilistic methods for topic modeling. Emphasis will be put into how one can compare different approaches to the problem along with consideration of computational issues involved in each method. We will consider how one can approach the subject of topic modeling from a network perspective using a Bayesian framework, in particular we will also review current methods used for relational topic models. We will begin with a literature review of Latent Dirichlet Allocation (LDA) and conclude with an empirical evaluation of current methods along with the potential development of new methods.

Social/Big Data: Text data is inherently social; people communicate to others through the use of the written word. Unstructured text data is becoming increasingly popular in the social sciences, where information comes in the form of surveys, comments, articles, and even twitter posts. Many hours are spent going through text data and turning it into the 'typical' form which fits nicely into a table. This process, not only time consuming and expensive, can miss some of the meaning in translation. Text data is also inherently 'BIG'. We will take into consideration computational aspects of all methods with respect to both corpus size(the number of documents to be analyzed) and documents size(the length of these documents). Possible datasets we may work with include tweets from twitter, the simple English Wikipedia articles, the Association for Computational Linguistics (ACL) Anthology, and the CRAN repository.

Publication Strategy:

The results of this project will be communicated through the following means:

1. IGERT Spring Poster Session
2. Document equivalent to a Master's Project, but geared towards inclusion in a PhD. Dissertation