6.882 Proposal: Bayesian Inference on Popularity

Andrew Xia, Kelvin Lu, Brandon Zeng

March 23 2018

1 Project Proposal

For our final project, we are interested in using Bayesian Analysis techniques to better understand how to predict tweet popularity from existing Twitter data.

We plan on implementing a paper by Zaman et al [3] which details a Bayesian approach for predicting the popularity of tweets, in which the popularity of a tweet is defined as the number of retweets it receives. The paper analyzes a tweet and examines properties of the retweet graph, such as the depth, the time delay between retweets, in addition to properties of the user (number of followers, etc) to create a generative model of the probability that a tweet will be retweeted.

Our main priority is to replicate the model with the data provided in [3]. In addition, the dataset in [2] contains over 120,000 tweets and their subsequent retweet graph, classified on whether the original tweet was "real" or "fake" news[1].

If time permits, we also hope to augment the model presented in [3] by including additional features such as the time of day of the tweet. One deficiency in [3] is that the model doesn't incorporate the actual text in the tweet as a feature. We hope to include text as a feature to our model, using models such as TF-IDF or word2vec. Finally, analyzing the popularity and spread of tweets may not be based on a single tweet; in fact, multiple tweets (such as trending topics) may affect the spread of a tweet. We can look at hashtags within a tweet to determine correlation between tweets in our model.

2 Plan and Deadlines

We plan to take the following steps to complete our project, as detailed in the following timeline:

1. Week of 3/26 (Spring Break): we plan on spending more time conducting background reading and exploring more papers to impwlement.

[1] Fake news is a feature that we initially won't be examining but it could also be an interesting feature to encorporate!

2. Week of 4/2 Complete explaratory data analysis to familiarize ourselves with the dataset (i.e., tweets, retweet graphs, reaction times). Implement generative model for retweet graph evolution.

3. Week of 4/9: Submit Progress Report. Implement log-normal model for reaction times and bionomial model for retweet graph structure.

4. Week of 4/16 Combine the models into an overall graphical log-normal-binomial model for the evolution of retweet graphs.

5. Week of 4/23 Implement the posterior distribution, and sample using MCMC.

6. Week of 4/30 Evaluate our model using absolute percent error between predicted and actual tweets. As a baseline, we can compare against a regression model that uses tweet counts.

7. Week of 5/7: Write the project paper and submit.

3 Division of Labour

For this project, we will be working on separate portions of the model concurrently (e.g. one person will implement the log-normal portions of the model and another person will implement the binomial model portions). All of our code will be available and public at `https://github.com/qandrew/6.882-fp`. Data collection, sampling, and writing the final report will be done collaboratively.

4 Project Risks

One major risk to the progress of our project is our speed of implementation. If implementing the hierarchical graphical model becomes time-consuming, we will consult references online on implementing MCMC samplers, etc.

While we have already located our data, finding additional data to test our model may also become problematic. The datasets from [2, 3] are formatted differently, so data processing may be time consuming. If we are sparse on time, we will focus on implementing our model on the data from [3] first.

References

[1] Wallach Guo, Blundell and Heller, *The bayesian echo chamber: Modeling social influence via linguistic accommodation*, International Conference on Artificial Intelligence and Statistics (2015).

[2] Roy Vosoughi and Aral, *The spread of true and false news online*, Science (2018), 1146–1151.

[3] Fox Zaman and Bradlow, *A bayesian approach for predicting the popularity of tweets*, The Annals of Applied Statistics (2014), 1583–1611.

Appendix

Pre-Proposal

We initially focused on applying Bayesian inference on the spread of fake news on Twitter data, in which our current project is a generalized approach to this problem space. Our pre-proposal is available below for reference (although not directly relevant).

We hope to build upon previous work in this space in our analysis. For example, the Bayesian Echo Chamber [1] uses a model that when two people interact, the first person's use of certain words can increase the second person's probability of subsequently using such word. This in turn can create a model on a social graph to determine which accounts carry more influence, and help us propagate our detection of fake accounts. Another vein of work we can consider, influenced by a recent paper on fake tweets by the Media Lab [2], is to infer the validity of a tweet based on the retweet structure (who retweets, how often, etc.).