Generating Status Hierarchies from Meeting Transcripts using the Author-Topic Model

David A. Broniatowski
Engineering Systems Division
Massachusetts Institute of Technology
Cambridge, MA 02139
david@mit.edu

Abstract

Topic models may be applied to solve problems of interest to many sub-fields of social-science. This paper expands the social-science uses of topic modeling to the analysis of group decision-making. In particular, we study committees of experts in the U.S. Food and Drug Administration. The output of the analysis is a set of directed social networks that reveal meaningful status hierarchies within these committees.

1 Introduction

Topic models have been applied to the social sciences in a limited fashion, with examples having largely taken the form of time-series studies of the evolution of specialized corpora (for example, [1]). The advantage of using topic models over more traditional social-science content analysis methods include the repeatability and consistency of the approach, as well as time- and labor-saving advances in automation. Quinn et al. [2] provides a compelling justification for the adoption of topic models by scholars of agenda-setting in political science. We feel that this argument may equally be extended to other domains of interest to social scientists. This paper demonstrates the application of the Author-Topic model [3] to the analysis of decision-making by small groups, a problem of interest in economics, sociology, organizational and social psychology, and engineering systems. In particular, we extend the methodology introduced in [4] to generate a set of directed graphs with the intention of examining status hierarchies on FDA Panels.

2 Literature Review

Research in the social sciences suggests that the determination of interpersonal affinity might be identified through the use of common language and jargon (see [5] for example). In previous work, we have operationalized these insights using the Author-Topic model, to construct social networks that represent whether actors within a committee meeting are using similar terminology, and therefore similar perspectives, to discuss the common problem to be solved [4]. In this paper, we incorporate status effects into the analysis by drawing upon the Expectation States literature in sociology (see [6] for example), which notes that, within small groups, status is often linked to frequency of speech and capacity to affect a topic shift. For example, a high status speaker may change the subject, whereas a lower-status speaker will remain on the subject introduced by the higher-status speaker.

3 Methodological Overview

A social network is created by first generating a term-document matrix from a meeting

transcript. Words are stemmed using PyStemmer [7], and function-words are removed. The AT Model is then used to assign each word token to a topic. Following a procedure similar to that outlined in [3], each document (i.e., each paragraph utterance) by a committee voting member is assigned to both that individual and to a "fictitious author", named "committee". This assigns words that are common to all voting members, such as procedural words, to "committee" enabling a focus on individual speakers' idiolects. Hyperparameters for the AT Model are set such that $\alpha=50/T$ and $\beta=200/W$, where T is the number of topics and W is the number of words. This is consistent with guidelines set in the Topic Modeling Toolbox [8]. A collapsed Gibbs sampler, using the algorithm outlined in [8], is used to generate samples from the AT Model's posterior distribution. Given the different subject matter for each meeting, we should not expect the same number of topics to apply for each analysis. The number of topics is therefore chosen independently for each transcript as follows: 35 AT Models are fit to the transcript for T = 1 … 35 topics. For each model, 20 samples are generated from one randomly initialized Markov chain after a burn-in of 1000 iterations. We find the smallest value, t_0, such that the 95th percentile of all samples for all larger values of T is greater than the 5th percentile of t_0. Given fitted priors of the sort recommended by Griffiths and Steyvers [8], the asymptotic behavior displayed in Figure 1 is typical of AT Model fits. We set the value of $T = t_0 + 1$ so as to ensure that the model chosen is beyond the knee in the curve.

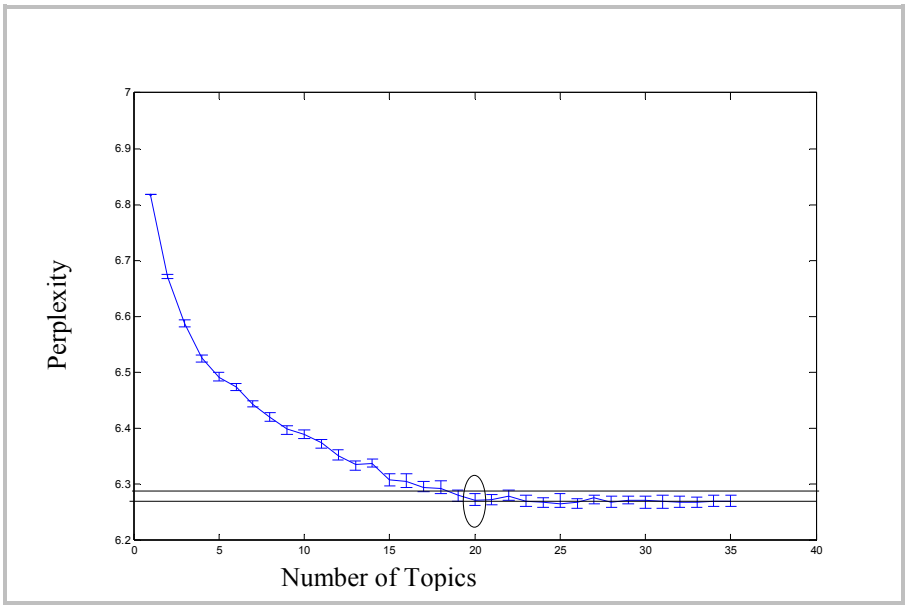

Figure 1: Perplexity vs. number of topics for the meeting of the FDA Circulatory Systems Devices Panel held on July 9, 2001. T, the number of topics, is equal to 20, using the procedure described above. Horizontal lines indicate the 5th and 95th percentiles for perplexity for a 19 topic model fit.

Once the number of topics has been chosen, a T-topic AT Model is again fit to the transcript. Ten samples are taken from 20 randomly initialized Markov chains, such that there are 200 samples in total. These form the basis for all subsequent analysis. A network is generated for each sample by examining the joint probability distribution for each author-pair in that sample (where X_i is the ith author, and Z_j is the jth topic):

$$P^s(X_1 \cap X_2) = \sum_i^T P^s(Z = z_i \mid X_1) * P^s(Z = z_i \mid X_2)$$

If the joint probability distribution exceeds 1/T, the lower bound on $P(X_1 \cap X_2)$ given completely unstructured topics, then we say that this author-pair is *linked* in this sample. Averaging over multiple samples drawn from the posterior distribution balances the impact of this relatively low threshold. In particular, , two authors are connected by an edge in the transcript-specific social-network if they are linked at least 125 times out of 200 samples.

This is consistent with the use of Bonferroni's criterion for a binomial distribution with family-wise error rate of 5%, and assuming ~15 voting committee members. An example network is shown in Figure 2.

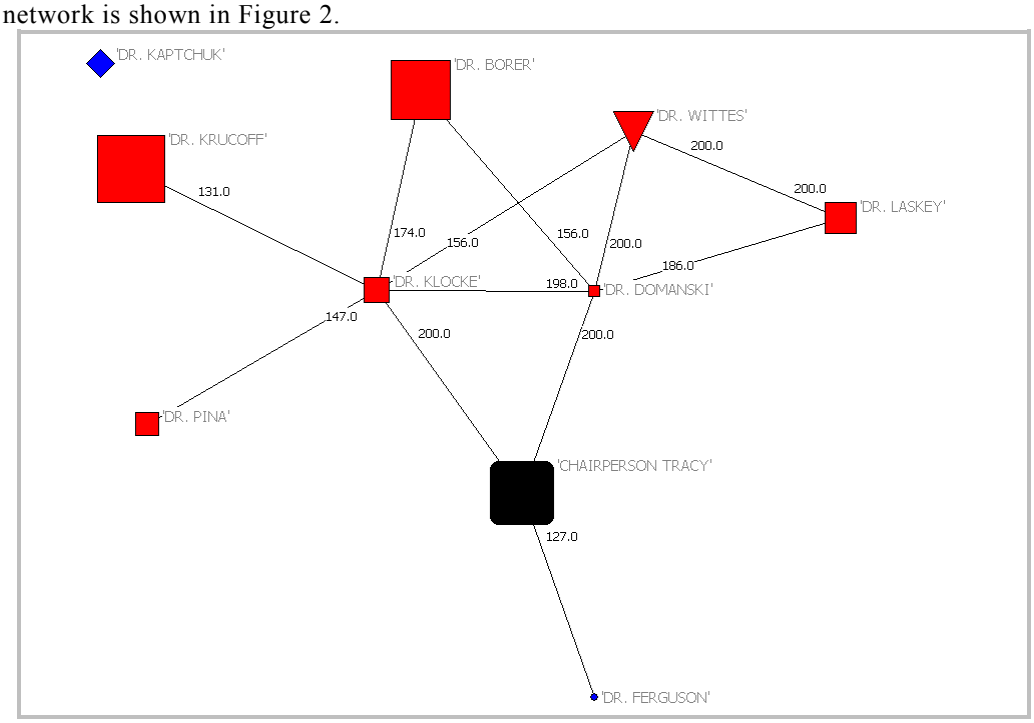

Figure 2: Network representation of the FDA Circulatory Systems Advisory Panel meeting held on July 9, 2001. Node size reflects number of words spoken; node shape represents medical specialty. Dr. Tracy was the committee chair. Non-approval votes are red; approval votes are blue; non-voters are black. Network generated by UCINet software [9].

We incorporate insights from the Expectation States literature by including information about who follows whom within a given topic, as follows: For each sample, s, from the AT Model's posterior distribution, we examine each author pair, X_i, X_j. If edge $e_{i,j}$ exists in the transcript-specific social-network defined above (i.e., if authors X_i and X_j are linked in at least 125 samples), then for topic, t, the topic cross-correlation function is $(f_{i,t}^{s} * f_{j,t}^{s})[\delta] = \sum_{d=-\infty}^{\infty} f_{i,t}^{s^*}[d] f_{j,t}^{s}[\delta + d]$ where $f_{i,t}^{s}(d)$ is the number of words spoken by author i and assigned to topic t in document d, in sample s. We examine the k^{th} peak of the topic cross-correlation function $m_k = \arg\max_{\delta}(f_i^t * f_j^t)[\delta]$, where k ranges from 1…n, the total number of maxima in the cross-correlation function. For each peak, if $m_k > 0$, we say that author i *lags* author j in topic t, at point m_k (i.e., $l_{i,j,t,m_k}^{s} = 1$). Similarly, we say that author i *leads* author j in topic t at point m_k (i.e., $l_{i,j,t,m_k}^{s} = -1$) if $m_k < 0$. Otherwise, $l_{i,j,t,m_k}^{s} = 0$. We define the *polarity* of authors i and j in topic t to be the median of the $l_{i,j,t,.}^{s}$. Next, we define the direction of $e_{i,j}$ in sample s as $d^s(e_{i,j}) = \sum_{t=1}^{T}(p_{i,j,t}^{s} * P^s(X_i \cap X_j))$. The *net edge direction,* $d(e_{i,j})$ is determined by partition of the unit interval into three equal segments. In particular, we examine the proportion of $d^s(e_{i,j})$ that are greater than 0. If more than 66% of $d^s(e_{i,j}) > 0$ then $d(e_{i,j}) = 1$ (the arrow points from j to i). If less than 33% of $d^s(e_{i,j}) > 0$ then $d(e_{i,j}) = -1$ (the arrow points from i to j). Otherwise, $d(e_{i,j}) = 0$ (the arrow is bidirectional). The result is a directed network, an example of which is seen in Figure 3.

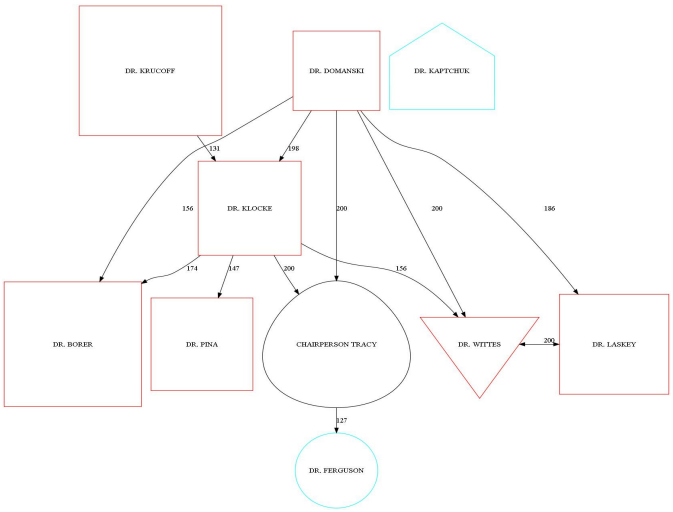

Figure 3: Directed network representation of the FDA Circulatory Systems Advisory Panel meeting held on July 9, 2001. Node size increases with the number of words spoken by that author; node shape represents medical specialty. Dr. Tracy was the committee chair. Non-approval votes are red; approval votes are blue; non-voters are black. This diagram is generated using the dot algorithm [10].

4 Results

The method described above can be used to generate meaningful representations of hierarchy within a committee meeting. We provide a preliminary validation of this claim by examining the set of 17 meetings in which the panel did not reach consensus. Analysis indicates that nodes representing voting members who are in the minority are 1.8 times as likely to have no children than are nodes representing voting members who are in the majority (χ^2=7.05; dof =1; p=0.008). These results indicate that topic-models may be used to analyze group decision-making. Future work will focus on validating these results against agent-based models in sociology and social psychology, and on incorporating the above method into a unified latent-variable model which might then be directly compared to state-of-the-art topic models.

References

[1] D. Hall, D. Jurafsky, and C.D. Manning, "Studying the History of Ideas Using Topic Models," *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing}*, Association for Computational Linguistics}, 2008, pp. 363-371.

[2] K.M. Quinn, B.L. Monroe, M. Colaresi, M.H. Crespin, and D.R. Radev, "An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th US Senate," 2006.

[3] M. Rosen-Zvi, T. Griffiths, P. Smyth, and M. Steyvers, "Learning Author Topic Models from Text Corpora," Nov. 2005.

[4] D.A. Broniatowski, "A Method for Generating Social Networks from Meeting Transcripts," International Joint Conference on Artificial Intelligence (IJCAI), Workshop on Modeling Intercultural Collaboration and Negotiation (MICON), Pasadena, CA: Springer-Verlag, 2009.

[5] J.A. Brown, "Professional language: words that succeed," Radical History Review, vol. 34, 1986, pp. 33-51.

[6] D.R. Gibson, "How the Outside Gets in: Modeling Conversational Permeation," Annual Review of Sociology, vol. 4, 2008.

[7] M.F. Porter, Snowball: A language for stemming algorithms. October 2001.

[8] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, Apr. 2004, pp. 5228-5235.

[9] S. Borgatti, M. Everett, and L. Freeman, "UCINET 6 For Windows: Software for Social Network Analysis," 2002.

[10] E.R. Gansner and S.C. North, "An Open Graph Visualization System and Its Applications to Software Engineering," SOFTWARE - PRACTICE AND EXPERIENCE, vol. 30, 1999, pp. 1203--1233.