A Comparison of Social Network Analysis and Text Regression using Twitter Data

Drew Conway and John Myles White

September 12, 2011

1 Introduction

Within the social sciences, interest in statistical methods for the automatic analysis of text has grown considerably in recent years. Specifically, within the field of political science there has been continued interest in using text analysis to extend well-known spatial models of political ideology (Grimmer, 2010; Monroe et al., 2008; Laver et al., 2003). Moreover, interest in how the structure of social networks affects political attitudes and outcomes has been both growing (Siegel, 2009; Hafner-Burton et al., 2009) and controversial (Fowler and Christakis, 2010; Lyons, 2010). In an attempt to bridge the gap between these two research areas we have collected a large sample of Twitter data related to the U.S. Congress. Twitter provides a rich data set containing both text and social network information about its members.

Here we compare the usefulness of text analysis and social network analysis for predicting the political ideology of Twitter users — two methods that are, in principle, applicable to both members of Congress (for whom roll call data and precise spatial estimates of political ideology already exist) and to the surrounding network of Twitter users (for whom precise estimates of political ideology do not exist). To compare text analysis methods with tools from social network analysis, we fit a variety of L1- and L2-regularized regression models that use word count data from individual tweets to predict the ideal points of members of Congress. We then compare the performance of the resulting text models with the performance of social network models that employ techniques developed to model diffusion over networked structures to estimate the ideal points for the same members of Congress. In this preliminary study we find that each method provides novel insight into the ideological spectrum of the U.S. Congress.

2 Methods

Using data from the Sunlight Foundation, we have identified 67 Senators and 313 Representatives who maintain a presence on Twitter.[1] For these members of Congress, we have generated a data set of nearly one-hundred thousand tweets posted to Twitter between the period of November 11, 2007 and August 15, 2011. The data were harvested using a spider written in Python that makes calls

[1] The Sunlight Foundation provides this information via its API at `http://services.sunlightlabs.com/docs/Sunlight_Congress_API/`.

to Twitter's API on a regular basis.[2] Each tweet in the data set contains the following information: full name of member of Congress, title (Senator or Representative), party affiliation, home state, gender, Twitter user name, text of tweet, date and time the tweet was published, and the unique ID number for the tweet within Twitter's database.

In addition to this source of text data, we have used Google's SocialGraph API to generate the directed social graph for the members of Congress who use Twitter.[3] Summary statistics for this graph are shown in Table 1.

Table 1: Summary Statistics for Main Component of Congressional Twitter Network

Nodes	Edges	Density	Mean Degree
418,941	926,385	2.21	$5.28e^{-6}$

Taken together, the text and social network data set provides a useful testbed for comparing methods for measuring political ideology. Because we can use ideal points measured for all of the members of Congress based on roll call data (Jackman, 2001), it is possible to test our predictions rigorously. And the bicameral nature of the Congress provides a natural division for testing a predictive model on held out data: we fit our models to data from the House and then test the models on data from the Senate.

Given both text data and social network data, there are two obvious models that we can fit: a text regression model in which the word counts for each tweet are used to predict the ideal point of the tweeting member of Congress and a social network model in which structural position is indicative of ideological position. In this case we can model this process as a diffusion of ideological broadcasts through the social network with a given rate of absorption and decay at each node. By comparing the RMSE of both models on data from the Senate after fitting the models to the House, we can determine the viability of both analytic methods.

2.1 Text Regression

The 96,829 tweets in our data set can be treated as 96,829 separate observations; for each of these tweets, we observe the number of occurrences of any of the 40,000+ words in our corpus of tweets. Because many words occur only once, we have removed the terms that occur in less than 90 documents. After pruning, we have only 1,361 terms. Given the measured word counts as a predictor matrix, we attempt to predict the ideal point of the member of Congress that wrote each tweet in a standard linear regression with either L1 and L2 regularization of the fitted coefficients. Additionally, we perform regressions using either (a) only the hashtags mentioned in tweets or (b) only the mentions of other Twitter users occurring in our tweets. Results from all four regressions are reported below.

We fit the model using standard convex optimization techniques: this is implemented in the R package, `glmnet`, which we have used for our analyses. Using this software, we fit several model

[2] The code used to generate this corpus, and all other data related to this project, can be inspected and downloaded here: `https://github.com/drewconway/Twitter-Congress`. Readers should note, however, that due to Twitter's data policies the raw data set of tweets could not be shared. Also, due to the moving window of data available in Twitter's search API any replication of the data using the code available here may not match exactly what is used in this following analysis.

[3] We have used Google's social graph service rather than Twitter's because our need to build the full social graph of all members of Congress exceeded the limits provided by the Twitter API, but was within Google's.

variants, including two L1-regularized models that use only the hashtags and mentions from the tweets. Even these substantially impoverished models (there are only 79 hashtags and 27 mentions) outperform the baseline model on held out data.

2.2 Network Model

To model political ideology on the Congressional network we construct a simple model of transmission using linear decay. We use the Jackman scores as assumed ideal points for one set of members of Congress, i.e., either Representative or Senators. These scores are then 'broadcast' over the network, and each node absorbs this score at a linear rate of decay based on geodesic distance from the broadcasting node. Equation 1 below describes the form of this model.

$$\hat{\pi_v} = \sum_{i=1}^{N} \frac{\pi_i}{k+1} \tag{1}$$

This equation states that the estimated ideology of some node, $\hat{\pi_v}$, is equal to the sum of all broadcast ideologies for some set N of nodes within the network. In the case of the Congressional Twitter network, the set N will either be the set of Representative or Senators. These ideological broadcasts decay at an linear rate given the geodesic distance k of v from i. There have been many different decay rates proposed for information transmission in networks within the literature literature (Wu et al., 2004). Given the scale of the network this simple linear additive model is preferable both theoretically and for computational tractability.

3 Results

The results from the text regression and network model analyses are shown in Table 2 below. For the text regression, the baseline model uses only the average ideal point to make predictions for Senators, while the L1 and L2 regularized regression models use coefficients for all 1,361 terms in our pruned corpus. The Hashtags Only model is an L1 regularized regression model using a subset of 79 of the predictors from the full L1 model; similarly, the Mentions Only model is an L1 regularized regression model using only 27 of the predictors from the full L1 model. As an example of the terms that are given large weights by our regression models, we have shown the ten hashtags and mentions with the largest weights: these terms represent either the most strongly Republican or most strongly Democratic terms in our corpus when restricted to hashtags or mentions.

With the network model, we see that the results are slightly better than from the text model. For the results reported in Table 2, we use the ideological scores for the opposing house to predict the other. For example, we use the scores for the House of Representatives to predict the ideology of Senators in the Twitter network. As you can see from this table, the model for the Senate performs quite well, explaining about 26% of the variance in our data relative to the baseline model.[4]

In Figure 1, we visualize the relationship between the true Senate ideal points and the predicted ones from the network model. In the right panel of Figure 1 we replace the graphed points with the surnames of the senators for ease of examination.

[4] A simple linear transformation was done to the raw network model scores to scale them with respect to the Jackman scores.

Table 2: Comparison of text regression and network models

	Model	RMSE	R^2
Text Regression			
	Baseline Model	1.062	
	L1 Regularization	0.9729	0.08390
	L2 Regularization	0.9771	0.07994
	Hashtags Only	1.037	0.02354
	Mentions Only	1.058	0.00377
Network Model			
	Sen. Baseline	1.010	
	Sen. Estimate	0.7470	0.26097
	Rep. Base	0.9949	
	Rep. Estimates	0.8577	0.1378

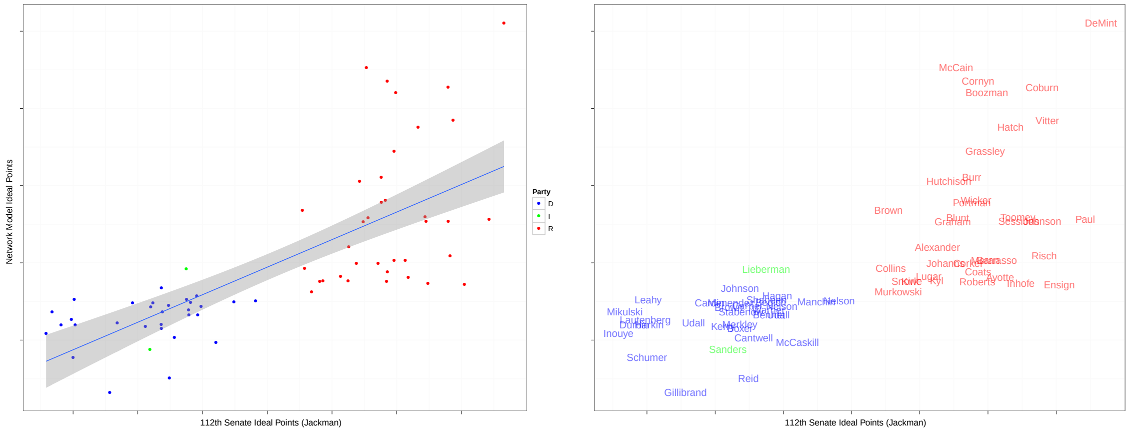
Table 3: Top 5 Republican and Democratic Hashtags and Mentions from Lasso

	Hashtag	Weight	Mention	Weight
Republican	#jobs	1.0288	@speakerboehner	0.7664
	#p2	0.6820	@foxbusiness	0.7126
	#dadt	0.6780	@foxnews	0.6955
	#reins	0.6307	@gopwhip	0.6776
	#tx17	0.6292	@thehill	0.6505
Democrat	#askobama	-1.2113	@cspan	-0.0743
	#pa11	-1.2181	@wsj	-0.1875
	#debt	-1.2209	@rephensarling	-0.2954
	#healthcare	-1.3877	@natresources	-0.5474
	#libya	-1.6383	@barackobama	-1.1266

4 Discussion

Our work provides a proof of concept demonstration of the viability of using both social network data and text derived from Twitter to model the political culture of the U.S. Future work will need to improve on the fine details of the methods we have employed. Another promising research project is to construct statistical models that employ social network and text data simultaneously as a twofold set of constraints for the estimation of spatial models of political ideology.

For the network model, we are not concerned with the content of a member of Congress's tweets, but rather with the way in which their position within the greater social network of Twitter exposes their political ideology. This is particularly interesting because Twitter is a *directed graph*, meaning that relationships are not mutual. The model, therefore, shows that given a set of fixed ideological broadcast nodes within the structure of Twitter, we can accurately estimate the ideology of all other nodes regardless of whether they actually follow anyone. Put another way, who follows you on Twitter tells us as much about your ideology as who you follow.



(a) Ideal Points

(b) With Sen. Surnames

Figure 1: Fit of Network Model Results for Senators

References

Fowler, J. H. and N. A. Christakis (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences 107*(12), 5334–5338.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis 18*(1), 1–35.

Hafner-Burton, E. M., M. Kahler, and A. H. Montgomery (2009). Network analysis for international relations. *International Organization 63*(03), 559–592.

Jackman, S. (2001). Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis 9*(3), 227–241.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(02), 311–331.

Lyons, R. (2010, July). The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. *ArXiv e-prints*.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16*(4), 372–403.

Siegel, D. A. (2009). Social networks and collective action. *American Journal of Political Science 53*(1), pp. 122–138.

Wu, F., B. A. Huberman, L. A. Adamic, and J. R. Tyler (2004). Information flow in social groups. *Physica A: Statistical and Theoretical Physics 337*(1-2), 327 – 335.