Mapping the Political Twitterverse: Candidates and Their Followers in the Midterms

Alexander Hanna[1], **Ben Sayre**[2], **Leticia Bode**[3], **JungHwan Yang**[2], **Dhavan Shah**[2]
[1]Department of Sociology
[2]School of Journalism and Mass Communication
[3]Department of Political Science
University of Wisconsin-Madison
ahanna@ssc.wisc.edu, {bgsayre, lbode, jyang66, dshah}@wisc.edu

Abstract

Twitter provides a new and important tool for political actors. In the 2010 midterm elections, the vast majority of candidates for the U.S. House of Representatives and virtually all candidates for U.S. Senate and governorships used Twitter to reach out to potential supporters, direct them to particular pieces of information, request campaign contributions from them, and mobilize their political action. Despite the level of activity, we have little understanding of what the political Twitterverse looks like in terms of communication and discourse. This project seeks to remedy that lack of understanding by mapping candidates and their followers according to their use of hashtags (keywords) and user mentions (direct mentioning of other Twitter users). We have a unique data set constructed from tweets of most of the candidates running for the U.S. House of Representatives in 2010, all the candidates for the Senate and governorships, and a random sample of their followers. From this we utilize multidimensional scaling to construct a visual map based on hashtag and user mention usage. We find that our data have both local and global interpretations that reflect both political leaning and strategies of communication. This study provides insight into innovation in new media usage in political behavior in particular and a bounded topic space in general.

Introduction

Twitter as a medium of communication has rapidly come of age since its inception. It has opened the door for understanding communication, both at-large and in certain bounded topic spaces. For researchers, the medium of Twitter offers a unique context in which content and connections can be studied simultaneously.

This study takes a focused and unique sample from the political Twitterverse – candidates in the U.S. Congress in 2010 and their followers – and attempts to create a map of this space based on elements of their speech, namely hashtags. Through this, We are able to gain a more nuanced understanding of what these universes of discourse look like, and how political users are connecting with one another in this new medium.

With the dramatic growth in popularity of Twitter, political actors have increasingly begun to use tweets as one of many campaign tools. The vast majority of candidates for the U.S. Congress in 2010, for instance, employed Twitter at least marginally in their campaign strategy. Political use of Twitter by candidates included efforts to reach out to potential supporters, direct them to particular pieces of information, request campaign contributions from them, and mobilize their political action. As a result, a large amount of valuable information on political behavior is embedded in the political Twitterverse.

Prior Literature

The vast majority of research to date on the political use of Twitter has focused on members of Congress. Scholars have considered both what encourages members of Congress to adopt use of Twitter, and what helps them to be "successful" in such use. Lassen and Brown (2010) found members are more likely to adopt Twitter if their party leaders urge them to, if they are young, or if they serve in the Senate, whereas Gulati and Williams (2010) determined that party and campaign resources were the most important predictors. Chi and Yang (2010a) suggest that adoption is driven by a desire for constituency outreach, rather than a transparency motivation. Adoption may be accelerated by evidence of past users' success with the medium (Chi and Yang 2010b), and factors including vote share, funding, usage and influence may help to explain why some congressional users have more followers than their colleagues.

A single study to date has examined Twitter use within the electoral context, in an attempt to predict election outcomes. Tumasjan et al. (2010) searched for mentions of political candidates and political parties in tweets. Word count analysis of this sample of explicitly political tweets revealed that the more frequently a candidate or party was mentioned, the more likely electoral victory for that entity.

Our study hopes to further this literature by incorporating a global understanding of the political Twitterverse and using elements from this understanding as a way to gain insight into political outcomes. This is by design an exploratory study, not aiming at explanation of any explicit outcome, but attempting to describe the shape of a communicative space for a particular subject matter.

Towards Building a Map

In our mapping, we attempt to identify users who are more similar to each other by virtue of what they actually say – the elements of speech that they share. At the same time, we are able to identify which articles of political speech are more alike since they will share the same users. This means that there is a *duality of actors and political speech*, very similar to the duality of persons and groups as conceptualized by Breiger (1974).

We have two expectations of what the analysis of the map will produce. Much like Adamic and Glance's famous work mapping the political blogosphere (2005), we expect to find a sharply divided Twitterverse along party lines. Those who are on the Left will tend to say similar things and use similar hashtags. Second, we expect to pick up not only a global division based on partisanship, but also to identify local clusters of users who engage in types of political behavior specific to the political Twitterverse. As a nascent communication medium, people continually attempt to exploit that medium for their own purposes. In the political space, this is often to disseminate information as widely and effectively as possible. Therefore, we expect to detect in our data local clusters of users who engage in similar diffusion-maximizing behavior.

Data

Data for this project were gathered in two waves. The initial wave began on Labor Day 2010 and was based on a list of 404 candidates in 103 races for seats in the House of Representatives. At the time when this sample was started, the total number of Twitter accounts we could follow was too low for us to be able to follow all candidates for House races along with samples from their follower lists, so we had to select a subset of races to focus on. The strategy employed was to include all candidates in races that were either tossups or leaning to one side or the other (as judged by the New York Times in the last week of August), along with a handful of noncompetitive races chosen at random. 16 of these races had not held their primaries by Labor Day, and for these races all candidates running in the primary were included in the sample. Of those 404 candidates, 253 were from one of the two major parties and the rest were independents or third-party candidates. Out of this list of candidates, 233 were found to have Twitter accounts, with 201 of those being major-party candidates.

A random sample of followers was taken for each candidate such that it proportionally decreased as the sample approached the maximum sample size of 50. The size of the sample per candidate was calculated by

$$n_c = \frac{50}{1 + (50-1)/F_c} \tag{1}$$

in which F_c is the total number of followers for candidate C at the beginning of the measurement period.

The second wave of data collection was started at the beginning of October and included all gubernatorial candidates and all US Senate candidates as well as a replication of the first wave which resampled the House candidates using the same sampling formula. Among the races for governor, only three included a third-party or independent candidate in the race, and only the two candidates in the Nebraska race had not identifiable Twitter account. The Senate races were very similar, with only occasional races with third-party candidates and few candidates without identifiable Twitter accounts.

A new random sample of all candidates (N = 409) Twitter followers was added to the existing sample of users being followed, resulting in total follow sample of 23,466. Collection of tweets went until one month after the November 2 election.

Over this time period, nearly 9 million tweets were gathered, either directly tweeted by users in the sample or repeated (retweeted) by those users. The data were collected by using Twitter's Streaming API. A feature of the collection with the Streaming API is that the data structure returned by the API has a number of different elements included along with the actual tweet. This includes information such as all public user information and geolocation, and most relevant for the current project, the what the API labels as "entities", parts of the text which can be identified as either a URL, a hashtag (e.g. `#icwsm`), or user mention (e.g. `@UWMadison`). The current project takes advantage of the distinct enumeration of these "entities" which allows us to parse out entities from the data structure itself.

Methodology

In order to identify our variable of interest, we performed a multidimensional scaling (MDS) analysis (Kruskal and Wish 1981) for hashtags against every unique user in our dataset. We constructed a two-mode matrix of users by hashtags. Entries in the matrix were the number of times the user used the hashtag U_{ei} normalized by the user's total usage of hashtags $U_{e\bullet}$, then weighed by the population's total usage of that hashtag $P_{\bullet i}$. Equation 2 expresses this in mathematical notation.

$$M = [\frac{U_{ei}}{U_{e\bullet}} P_{\bullet i}] \tag{2}$$

The MDS analysis was performed using non-metric MDS. Non-metric MDS attempts to retain rank order of entries as ordered by distance while at the same time attempting to minimize the badness-of-fit (stress) iteratively (Kruskal and Wish 1981). We used the Kruskal's Non-metric Multidimensional Scaling function included in the R MASS package (Venables and Ripley 2002). Input to the MDS was a dissimilarity matrix calculated from Euclidean distance between rows of matrix M in equation 2. We allowed for two dimensions in order to most easily interpret the results graphically and substantively. Using additional dimensions did not dramatically decrease the stress.

As Breiger notes, we gain two things from this analysis. First we can see how actors cluster together in a two-dimensional space by virtue of what they say. This means we can discern distinct groupings of individuals based on their Twitter behavior. Second, we distinguish which entities are substantively closer to each other. Presumably, due to the concordance of entities uttered by a user, these entities share in being similar by some unobservable or latent variable. In the context of our study, we suggest that the latent variable is political sentiment. For example we expect to see a clustering of hashtags such as `#tcot` and `#teaparty`, and on the other end `#tlot` (Top Liberals on Twitter) and `#p2`.

The local interpretation of the analysis relies on the ability to observe clustering in the MDS output. This lends itself to more nuanced interpretations that do not accord to what may be considered only political leaning. We can also pick up on the variations in the types of political behavior in which users engage. Visually, we can discern clusters and attempt to assign meaning to them based on our knowledge of the case. We also can identify cluseters systematically using hierarchical cluster analysis (HCA). We used the output of the MDS analysis to generate a fitted distance matrix based on Euclidean distance between rows and used this as the input to HCA using Ward's method. We use this method to minimize the loss of information we get from the clustering process. This results in compact, spherical clusters of actors.

Making and Interpreting the Map

To generate the MDS, we used the 4979 users who used the most hashtags and the 4016 top hashtags. We found that using more users and hashtags did not change the analysis dramatically. What we expect in the political Twitterverse is a polarization based on the left/right or Democrat/Republican dichotomy such as in the political blogosphere (Adamic and Glance 2005). We can this both visually and computationally.

We need to be able to understand the map in a meaningful way and say more about its variance. There are two ways we can interpret the mapping, the global and local interpretation. The global interpretation lends itself to interpretation upon the axises. To assess the significance of any global interpretation, we created a variable based on the number of candidates the user followed, separated into five categories: *Democrat*, *Republican*, *Independent*, *Third-Party Left*, and *Third-Party Right*. The first three categories are self-explanatory, while last two were generated by categorizing various third-party groups according to their political leaning, i.e. Green for 3rd Party Left, and Libertarian and Tea Party for 3rd Party Right. We regressed the coordinates from the MDS using generalized linear modeling (GLM) on this variable. In addition, we created a separate variable from the two most popular hashtags in the political Twitterverse: `#tcot` and `#p2`. Again, we used GLM to assess the direction to which elements in the map lean.

Figure 1 displays the results for the the map based upon hashtags. We see the `#p2` and `#tcot` curves approaching the point of being orthogonal to each other, which is expected given the way that the map is constructed. It seems as though `#p2` line accords with the users on the left of the map (convenient since it is the hashtag that is supposed to represent the Left of the political spectrum) and the `#tcot` with those on the bottom and the right of the map. This is not terribly surprising. We also see that the lines of the Democrats and the Third-Party Left are nearly identical, and that they are the closest to `#p2`.

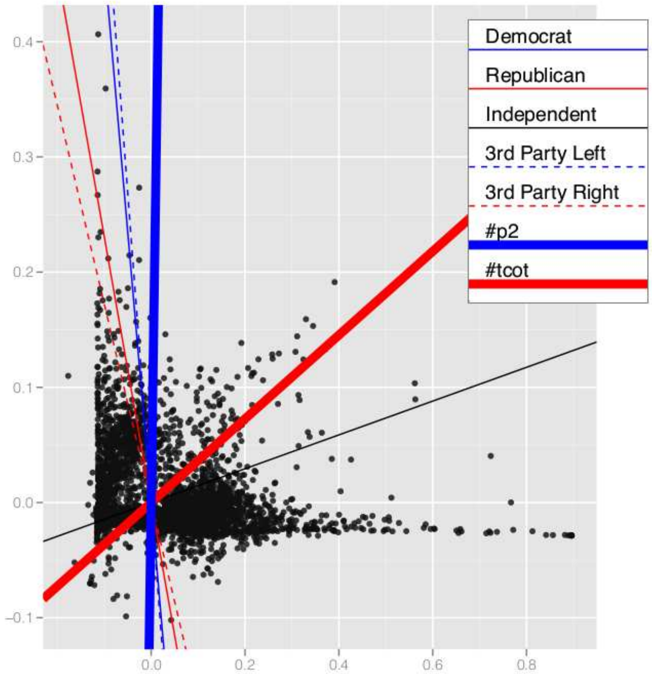

Figure 1: Multidimensional scaling with trend lines. Stress: 2.55

We would expect that the lines for the Republicans and Third-Party Right approach the `#tcot` line. But oddly enough, their slopes actually have the *opposite sign* of their respective hashtag. Even more strangely, the line for Independents seems to be far removed from all the other plots.

This analysis is revealing in two ways. It confirms that there is polarization in the political Twitterverse and describes how disparate users are with regard to hashtag usage. Secondly, and more importantly to understanding political behavior, we cannot understand how political actors behave in this online space merely from a global partisan dichotomy. There seem to be other mechanisms at work here. That is what the cluster analysis attempts to get at.

The results of this analysis are reported in Figure 2. We chose to separate the map into six clusters, although using five or seven clusters would not have changed the analysis dramatically. We can see distinct clusters which change consistently with the X axis, but do not change too much on the Y axis.

The regression analysis of this map is in Table 1. Clusters 1 and 2 are clearly in the Republican and Right camp, utilizing `#tcot` the most, showing high correlation in favor of Republicans and third-party Right candidates, and negative correlation with regard to Democrats. Similarly, cluster 6 exhibits strong affinities towards Democrats and `#p2`. However, what happens in the clusters between these two seems less clear. Cluster 3 correlates with both hash-

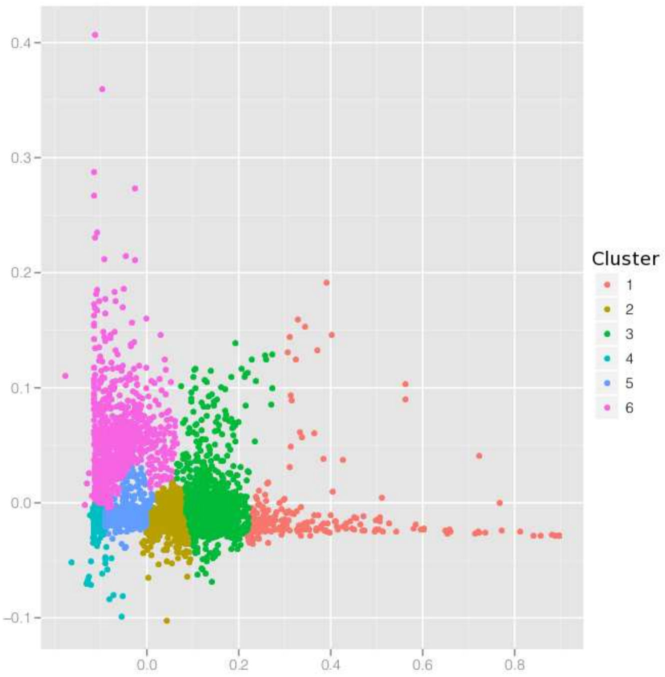

Figure 2: MDS with clusters

Cluster	1	2	3	4	5	6
N	223	708	1011	1603	616	817
Dem	-2.27	-2.37	-2.57	0.26 (ns)*	-1.07	6.27
Rep	6.84	7.5	6.14	-7.04	4.82	-8.65
Indep.	0.01 (ns)	-0.04 (ns)	-0.05	NA	0.08	0.02 (ns)
3rd R†	0.26	0.17	0.15	-0.16	0.08	-0.22
tcot	354.12	96.08	189.84	-171.47	-75.54	-87.27
p2	-15.33 (ns)	-32.09	35.42	-66.78	-42.16	131.12

* (ns) denotes non-significance at the p $\leq$ 0.01 level

† all coefficients for third-party left were insignificant

Table 1: Local interpretation of clusters

tags and leans towards Right candidates, 4 positively correlates with Democrats although nonsignificantly; more importantly it actually exhibits the second highest negative correlation with Republicans. Even more strange, it correlates negatively with both hashtags. Lastly, cluster 5 correlates negatively with both hashtags but otherwise has partisan leanings that look like cluster 3.

How are we to characterize these clusters, then? Clusters 1 and 2 could solidly be called conservatives and cluster 6 as progressives or liberals. Cluster 3, however, is a group that leans Right but uses both hashtags. One explanation for this finding is that these actors maybe exhibit a sort of "poaching" behavior in which they post tweets using both hashtags in order to enter the other side's discourse. If a user is interested in Left politics, she may search on the #p2 keyword. In order to get their message into that discussion, users in cluster 3 will post their messages with both hashtags. Clusters 4 and 5 seem to be those against and for Republicans, respectively, who are not distinguished based on their use of the two hashtags analyzed here. Therefore these users are not marked by their usage of the most popular hashtags, but possibly by other, more specialized hashtags. Further work within this framework could be done to find which hashtags these users are more inclined to use.

Conclusion

In this paper we have developed a method of creating a map of the political Twitterverse from the built-in functionality of Twitter. We found that solely Left/Right distinctions inadequately describe political behavior on the platform, and it is much more fruitful to discuss how users use Twitter in terms of actual political strategy, such as "encroaching" on others' keywords.

This analysis only chose to look at the use of hashtags in mapping the political Twitterverse. By the same token, we could have used the other entities, such as URLs and user mentions. We do not have to stop there, however. With computer-aided content analysis software such as InfoTrend or YoshiKoder, we can create an entirely new set of attributes from which to categorize tweets. The construction of this map will ultimately be useful in attempting to explain political outcomes such as elections and referenda. We hope that this method can be generally extrapolated to mapping any bounded space of discourses in the social media sphere, and attempting to explain outcomes in that space by virtue of elements of the map.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. WWW-2005 Workshop on the Weblogging Ecosystem, May 1014, 2005, Chiba, Japan.

Breiger, R. L. 1974. The duality of persons and groups. *Social Forces* 53(2):181–190.

Chi, F., and Yang, N. 2010a. Twitter adoption in congress. *Review of Network Economics*.

Chi, F., and Yang, N. 2010b. Twitter in congress: Outreach vs. transparency.

Gulati, G. J., and Williams, C. B. 2010. Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the united states congress. Chicago, Illinois: Midwest Political Science Association.

Kruskal, J. B., and Wish, M. 1981. *Multidimensional scaling*. Beverly Hills: Sage Publications.

Lassen, D. S., and Brown, A. R. 2010. Twitter: The electoral connection? *Social Science Computer Review*.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Venables, W. N., and Ripley, B. D. 2002. *Modern Applied Statistics with S*. New York: Springer, fourth edition. ISBN 0-387-95457-0.