The Party is Over Here: Structure and Content in the 2010 Election

Avishay Livne[1], Matthew P. Simmons[2], Eytan Adar[1, 2], Lada A. Adamic[1,2]

[1]Computer Science and Engineering, [2]School of Information,
University of Michigan, Ann Arbor
Ann Arbor, MI, USA, 48109
{avishay, mpsimmon, eadar, ladamic}@umich.edu

Abstract

In this work, we study the use of Twitter by House, Senate and gubernatorial candidates during the midterm (2010) elections in the U.S. Our data includes almost 700 candidates and over 690k documents that they produced and cited in the 3.5 years leading to the elections. We utilize graph and text mining techniques to analyze differences between Democrats, Republicans and Tea Party candidates, and suggest a novel use of language modeling for estimating content cohesiveness. Our findings show significant differences in the usage patterns of social media, and suggest conservative candidates used this medium more effectively, conveying a coherent message and maintaining a dense graph of connections. Despite the lack of party leadership, we find Tea Party members display both structural and language-based cohesiveness. Finally, we investigate the relation between network structure, content and election results by creating a proof-of-concept model that predicts candidate victory with an accuracy of 88.0%.

1. Introduction

Much has been made of the importance of social media in modern politics. Political parties and individual candidates have come to regard their online presence as so fundamentally important that they have hired staff members to act as social media coordinators. The speed by which a candidate can now access voters has led to extreme sophistication in the use of these systems. Twitter, with its 190 million registered users, is a particularly popular tool, allowing for rapid micro-blogged *tweets* (status updates) to be fired off to any *follower*.

Recent successful use of social media as part of political campaigns, particularly in the 2008 U.S. Presidential campaign of Barack Obama, had drawn both popular and academic attention. Obama's renowned tweet "We just made history…" which was published shortly after his victory, reflected the popularity of Twitter in political messaging. Today it seems as if every self-respecting campaign must have an online presence and the formula for a successful online campaign is highly sought after. Campaigners look for viral channels to garner supporters. Notably, understanding how political social networks form and communicate has broad implications not only within the political sphere but in the study of any network of competing agents in which information is transferred. In 2010, 22% of online adults used social networks or Twitter to engage with the election (Smith 2011).

In this work we investigate how the U.S. 2010 election campaigns were expressed on Twitter. We specifically analyze over three years' worth of tweets (over 460k) from 687 candidates running for national House, Senate, or state governor seats. As tweets are limited in size (140 characters) we augment our data by crawling nearly 233k outgoing links referred to by candidate tweets.

In addition to observing the behavior of Republicans and Democrats, the two major political parties, we also pay particular attention to Tea Party members. Although not an official party, self-identified members of the conservative Tea Party have been the subject of significant analysis and discussion. By separating Tea Party candidates in analysis from their official party position we are able to analyze the behaviors of this "virtual" party.

Our methods of analysis include both text and graph mining techniques. We suggest a novel use of language modeling for estimating the coherency of each group and the extremism of single candidates. We use graph analysis to compare the density of each group as well as to compute various graph properties of individual candidates. Finally, we combine the results in order to build a model that predicts whether a candidate is likely to be elected.

Our contributions include a detailed analysis of the social media behaviors of candidates in the 2010 midterm elections. We demonstrate a method for content-based, structural, and combined analysis of these candidates relative to each other and their parties as a whole. Using these techniques we characterize the attributes of the different parties, demonstrating high levels of structural and content coherence for conservative (Republican and Tea Party) members. We further analyze how centrality in structure and content correlate with election outcomes (positively) by employing a prediction model.

2. Related Work

Twitter Networks

The growing number of Twitter users, and the ease of access to their tweets, makes Twitter a popular subject for research in various research communities (Java et al. 2007). Though most are about the general population of users, a number are relevant to political structures (e.g., influence, viral marketing, computer-mediated communication, etc.). For example, Romero et al. (2010) portrayed influential users, refuting the hypothesis that users with many followers necessarily have bigger impact on the community. Honeycutt and Herring (2009) showed that Twitter often serves as a framework for discussions rather than for one-way communication. Another direction of study focused on commercial usage of Twitter (e.g., viral marketing). Jansen et al. (2009) performed sentiment analysis of tweets in that context (specifically targeting products and brands). Our work here is informed by previous work on Twitter content and structure.

Social Media and Politics

While initially focused on blogs (Adamic and Glance, 2005) and Facebook (Williams and Gulati, 2008), the analysis of social media in political contexts has since transitioned to include Twitter. Broadly, work in the area has focused on the analysis of the content and structure of elected political figures (e.g., members of Congress) or the use of Twitter as a social sensor to predict elections.

A number of studies (Golbeck et al. 2010; Glassman et al. 2010; Senak 2010), identified specific patterns of tweeted communication between members of Congress and their constituents in terms of quantity and content type (e.g., informational, fundraising, etc.). Sparks (2010) further analyzed partisan structure to identify groups with ideological leanings. Though we note similar structural features in our findings (e.g., increased messaging and density among conservatives), we concentrate our attention on candidates. By manually classifying tweets of candidates one week before the 2010 election, Amman (2010) found that most messaging by Senate candidates was informational and does appear to have a relationship to voter turnout.

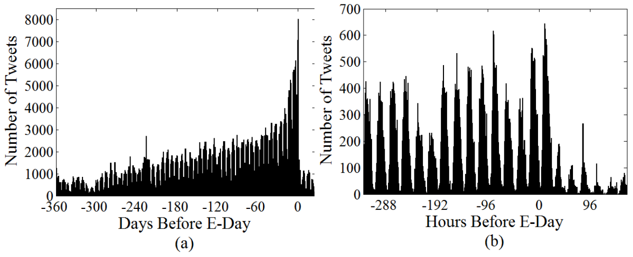

Figure 1. Daily (a) and hourly (b) volume of tweets

The use of Twitter as a "social sensor" for election prediction has been applied in a number of recent studies. Tumasjan et al. (2010) used chatter on Twitter to predict the German federal election, finding the number of tweets mentioning a political party to be almost as accurate as traditional polls in predicting election outcomes. Diakopoulos and Shamma (2010) showed that tweets can be used to track real-time sentiment about a candidate's performance during a televised debate. However, these previous analyses of political activity on Twitter did not specifically examine the candidates themselves, or the structure of their networks.

Language Models and Graphs

To model content we employ statistical Language Models (LM). Language models are statistical models in which probability is assigned to a sequence of words, thus representing a language as a probability distribution over terms. It was first used in speech recognition (Jelinek 1997) and machine translation (Brown et al. 1990). Ponte and Croft (1998) were the first to apply LM to the task of document ranking. Metzler et al. (2004) improved LM accuracy and (Song and Croft 1999) used smoothing to tackle text sparseness.

The construction of user profiles can lead to better results in information retrieval tasks such as web-search (Sugiyama et al. 2004) and recommendation systems (Zhang & Koren 2007). Xue et al. 2009 used LM for constructing user profiles to enhance search results. Similarly, Shmueli-Scheuer et al. (2010) described a distributed framework using Hadoop to construct LM-based user profiles (a technique we employ below).

3. Data

The system described in this paper makes use of data crawled from Twitter. In order to build a fairly complete list of candidate Twitter accounts we semi-automatically generated this collection. For each candidate, we executed a query on Google using their name and the keyword "twitter" and retrieved the top 3 results from the twitter.com domain. Each result was manually inspected and filtered (e.g., fake accounts mocking the candidates were removed), leaving only accounts that were operated by the candidates or their staff. Our data spans 687 users—339 Democrats and 348 Republicans. Of the 348 Republican candidates, 95 were further identified as Tea

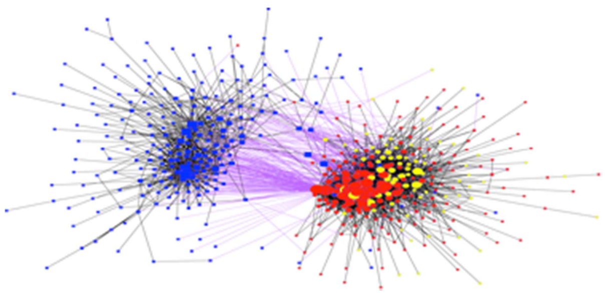

Figure 2. Plot of the candidate network (force-directed graph embedding layout modified to emphasize separation, nodes size proportional to indegree)

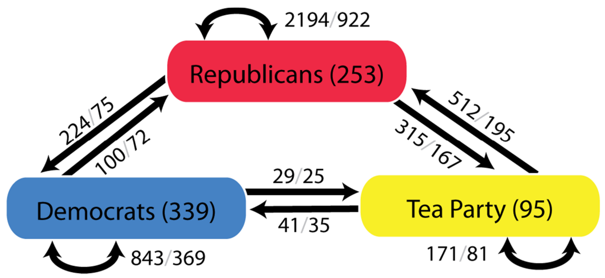

Figure 3. Number of explicit follower edges and unique @mention edges (follower / mention)

Party candidates[1]. Note that notationally we exclude Tea Party candidates from the Republican set. When it is interesting to analyze the inclusion or exclusion of Tea Party candidates we employ the notation Rep+TP and Rep-TP respectively.

Using Twitter's API, we downloaded 460,038 tweets for candidate accounts dating back to March 25, 2007. Figure 1 shows the number of tweets in the days (a) and hours (b) surrounding the Election Day. We see temporal patterns, as less activity is observed during weekends and nights. As expected, the volume of tweets increases towards November the 2nd, abruptly decreasing afterward.

The data include 84, 81 and 522 candidates from the Senate elections, the gubernatorial elections and the Congressional elections respectively, covering about 50% of the number of candidates in each of the races. We crawled all the edges connecting users in our dataset. To identify social structures we consider a *"follower → followed"* relation as a directed edge going from the follower to the followed user (identifying 4,429 such edges between candidates in our pool).

To enrich the dataset we crawled the homepage of candidates who maintained one and each of the valid URLs that appeared in the tweets and considered them as additional documents. Out of 351,926 URLs (186,000 distinct) 233,296 were valid pages (132,376 distinct), which eventually contributed 96% of the content to the dataset (182,523,302 terms out of 190,290,041). We filtered out stop words and extracted both unigram and bi-gram terms. We found no significant difference when n-grams of higher order were considered.

4. Methodology

In this work, we analyze two aspects of the data – the content produced by the users and the structure of the network formed by the follow-up edges. We start by providing some theoretical background to our content analysis methods.

User Profile Model

Notations

Our system consists of a set of candidates U where each candidate has a set of documents D_u associated with her. The entire corpus is denoted by $D = \bigcup_{u\in U} D_u$. Documents are represented using the *Bags of Words* model where each term $t \in d$ is associated with its number of occurrences in the document $tf(t,d)$. The vocabulary of the corpus is denoted by V. Our model is based on the $tf \times idf$ model; therefore we make use of the document frequency of a term $df(t)$ and the inverse document frequency $idf(t) = \log(1 + |D|/df(t))$. We denote the document frequency of a term in the set of user u's documents by $df(t, D_u)$. We also make use of $udf(t, D_u) = df(t, D_u)/D_u$, the maximum likelihood estimation of the probability to find term t in D_u.

Term Weighting

We set the initial weight of a term in a user LM to be

$$w(t,u) = \overline{tf}(t,D_u)udf(t,D_u)idf(t,D)$$

where $\overline{tf}(t,D_u) = \sum_{d\in D_u} tf(t,d)/D_u$ stands for the average frequency of term t in the collection D_u. In addition, we calculate the marginal probability of $t \in V$ in the language model of the entire corpus as

$$P(t|D) = \overline{tf}(t,D)udf(t,D)$$

These values are then normalized in order to obtain a probability distribution over the terms.

$$P^N(t|D) = \frac{P(t|D)}{(\sum_{t\in V} P(t|D))}; w^N(t,u) = \frac{w(t,u)}{(\sum_{t\in V} w(t,u))}$$

We then smooth the weights using the LM of the corpus,

$$P(t|u) = (1-\lambda)w^N(t,u) + \lambda P^N(t|D)$$

using a normalization factor of $\lambda = 0.001$. Finally, we divide these values by their sum to normalize them.

$$P^N(t|u) = \frac{P(t|u)}{(\sum_{t\in V} P(t|u))}$$

In a similar manner we constructed a LM-based profile for the Democrat and Republican parties, as well as to the group of Tea Party members. In order to compute the LM-based profile of a group G we applied the same process

[1] The Tea Party classification was obtained from The New York Times feature "Where Tea Party Candidates are Running," October 14, 2010 (nytimes.com/interactive/2010/10/15/us/politics/tea-party-graphic.html).

described above with the exception that the set of users' documents D_u is replaced with $D_G = \bigcup_{u \in G} D_u$, the union of the documents of the users in the group.

Content Analysis

We consider the content produced by a user to be the tweets that were produced by the user as well as the content of the URLs that appear in his tweets. We assume that in the majority of the cases these cited pages represent a candidate's opinion. In the discussion section we propose a more delicate interpretation using sentiment analysis.

In order to perform large scale analysis of the content we constructed a LM-based profile for each user, as described in the previous subsection. We apply the symmetric version of the Kullback-Leibler (KL) divergence on two LM profiles to estimate the difference between the content of the two corresponding users. For two distributions $P_1(t)$ and $P_2(t)$ over the terms in the vocabulary $t \in V$, the symmetric KL divergence is defined as:

$$D_{KL}^{S}(P_1||P_2) = \sum_{t \in V} P_1(t)\frac{\log P_1(t)}{\log P_2(t)} + P_2(t)\frac{\log P_2(t)}{\log P_1(t)}$$

We also used the (non-symmetric) KL divergence in order to measure the contribution of single terms to the difference of one profile from another.

5. Results

Basic Structure Analysis

The network structure of the candidate graph is visualized in Figure 2. Unsurprisingly, the Tea Party members are fairly intertwined within the Republican subgraph. We also note the relative densities (higher for Republicans) of the party substructures.

This is further confirmed through an analysis of subgraph density of edges within the same group. For a subgraph with *N* nodes and *E* edges, we utilize the density definition of *E/(N*^{2}*-N)*, or the ratio between the number of actual edges and the number of possible edges. Since density is sensitive to the size of the graph we considered in-degree as well.

Table 1. Subgraph Density by Group

	Democrat	Rep-TP	Rep+TP	Tea Party
Density	0.007	0.032	0.025	0.020
In-degree	2.55	8.37	8.97	1.82

Table 1 shows the calculated subgraph densities and mean in-degree. We note that the Democratic network is sparser than the networks of Republicans and Tea Party members, consistent with prior studies (Adamic & Glance 2005). This difference in density holds even when we consider the group of Republicans and Tea Party members (Rep+TP) which has more candidates than the group of Democrats, and so has more possible edges. Figure 3 represents the number of cross-party edges, for example we see 512 instances of a Republican being followed by a Tea Party member. Consistent with Figure 2, the Republicans and Tea Party members interact with one another more frequently than either do with Democrats.

Basic Content Analysis

Table 2 shows some statistics of the content produced by candidates in each party. Each value is the mean over the users in that group.

Table 2. Mean Usage Patterns by Group

	Democrat	Rep-TP	Tea Party
tweets	551	723	901
tweets per day	2.66	2.97	5.21
retweets	40	52.3	82.6
replies	172.6	260.5	472.7
hashtags	196	404	753
hashtags per tweet	0.37	0.54	0.68

Of note are the high levels of *tweets* and *tweets per day* for Tea Party candidates and relatively higher levels of Republicans over Democrats. We find the same relationship (Tea Party > Rep-TP > Democrat) for *retweets* (the rebroadcast of someone else's message) and *replies* (a response to someone's tweet). These results indicate that not only are conservative candidates more likely to "broadcast", they are more likely to communicate with each other. Finally, we note conservative candidates use more hashtags, potentially to provide additional unity. Hashtags—keywords/topics indicated with a "#"—are frequently used by communities for grouping tweets to create a *Trending Topic* to be highlighted by Twitter.

Hashtag Use

We took a closer look at the usage of hashtags by each of the groups. Table 3 presents the top 5 hashtags used by each group along with their number of occurrences and the number of unique users in the group that used this hashtag. The first part of the table shows the hashtags that were used by the greatest number of unique users, while the second part shows the hashtags with most occurrences.

It is somewhat surprising to find a conservative-related hashtag (*tcot*) as one of the top Democratic topics. However, a closer inspection of these tweets reveals negative information intentionally attached to this topic. Such behavior is consistent with previous observations on the number of mentions of opposing entities in political networks (Adamic and Glance 2005). Interestingly, we find the health care reform (*hcr*), a topic under much debate, to be almost equally brought up by both Republicans and Democrats. A number of hashtags—*ar02* and *alaska*—were utilized by a small number of extremely active candidates to refer to specific elections (rather than specific topics). Finally, we note the high levels of use of the Facebook (*fb*) tag produced automatically by programs cross-posting to the candidates' Facebook pages.

Table 3. Top Hashtags (# times used, unique users). p2 (Progressives 2.0); tcot (Top Conservatives on Twitter); nvsen (Nevada Senator); fb (Facebook); hcr (Health Care Reform); gop (Grand Old Party); nrcc (National Republican Congressional Committee); ar02 (Arkansas District #2); ff (FollowFriday); sgp (Smart Girls Politics).

Sorted by # unique candidates:		
Democrat	Rep-TP	Tea Party
p2, 4564, 96	tcot, 13347, 169	tcot, 11482, 70
hcr, 1176, 82	gop, 3929, 125	gop, 2262, 60
ff, 639, 80	hcr, 1772, 110	teaparty, 4419, 52
jobs, 427, 52	teaparty, 1706, 93	sgp, 1149, 38
oilspill, 708, 45	ff, 1160, 81	ff, 1188, 32
Sorted by mentions:		
p2, 4564, 96	tcot, 13347, 169	tcot, 11482, 70
tcot, 3403, 38	gop, 3929, 125	teaparty, 4419, 52
nvsen, 2471, 3	fb, 3882, 45	ar02, 3762, 2
fb, 1232, 32	nrcc, 2091, 29	alaska,2372, 1
hcr, 1176, 82	hcr, 1772, 110	gop, 2262, 60

Profiles Review

Extending beyond simple content features, we employ the language model (LM) based profiles described above. Table 4 provides a glimpse of some of the top terms in each party's profile (calculated as the marginal KL divergence of the term compared to the LM of the corpus). Note that the higher the marginal KL divergence of a term compared to the LM of the corpus, the more it contributes to differentiating a profile from the rest of the corpus. In other words, these terms serve best as features for identifying content produced by each party.

We found Tea Party members frequently mentioning Democratic political figures such as Nancy Pelosi, Barney Frank, and Ellen Tauscher (generally in a negative context). The Republican profile consists mostly of terms relating to the economy, such as *spending*, *bills*, *budget*, *tax cuts*, and the *deficit*, as well as various references to the *Wall Street Journal*. From a qualitative observation of keywords, the Democratic profile seems to cover the widest range of topics such as energy (*clean energy*, *solar*, *renewable energy*); education (*education*, *school*, *teachers*); the oil spill (*BP*, *oil spill*); military (*Afghanistan*, *Iraq*, *military*) and economics (e.g., *jobs*, *health care reform*, *recovery act*, and *social security*).

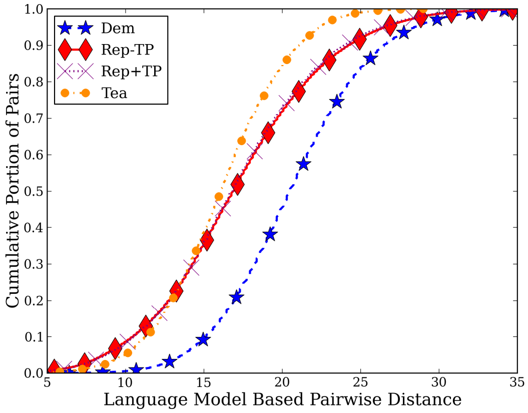

Figure 4. Pairwise KL divergence

Table 4. Top Terms

Democrat	Rep-TP	Tea Party
education	spending	barney_frank
jobs	bills	conservative
oil_spill	budget	tea_party
clean_energy	wsj (wall street journal)	clinton
afghanistan	bush	nancy_pelosi
reform	deficit	obamacare

Content Cohesiveness

To understand the cohesiveness of content amongst the different parties we apply we calculated the KL divergence between every pair of candidates from the same party (i.e., determining how similar party members were to each other). Figure 4 demonstrates the cumulative distribution of the pairwise distances. Intuitively, the more quickly the cumulative distribution reaches 1, the more similar the profiles of users from this group are.

It can be seen that the content of the Tea Party members is more homogenous compared to the rest of the Republicans while the Democrats lag behind, indicating they produce heterogeneous content. This finding correlates with a qualitative inspection of topics generated through topic modeling (Blei et al. 2003) where we found the profile of the Democratic Party covers a wider range of topics than the conservative groups. In addition, we see Tea Party members having a negligible effect on the LM of the Republican group as a whole. This can be explained by the relatively small number of Tea Party members and the similarity in the content attributed to these two groups.

Content Distance versus Structural Distance

We hypothesize that the closer two users are in graph distance, the more similar their content would be. This can, in part, be explained through models of homophily and social influence.

To test this idea, we looked at every pair of candidates, calculating the shortest path in the network as well as the KL divergence in their language models. The results are

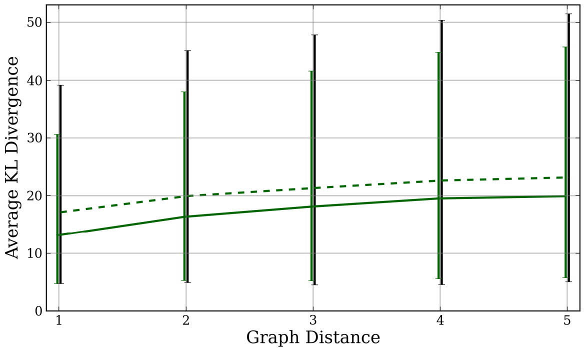

Figure 5. Mean pairwise KL divergence vs. pairwise distance considering retweets (solid line) and ignoring retweets (dashed line). The left (green) error margins describe the 10% and 90% percentiles of the data with retweets, while the right (black) error margins stands for the data without retweets.

depicted by the solid line in Figure 5, along with error bars at the 10% and 90% percentiles. Note the significant increase in the KL divergence as the distance increases from one to three hops. The effect diminishes for distances greater than 3 steps. We found this phenomenon to be consistent for each of the political groups as well as for pairs of candidates from different parties. As we discuss with further detail in Section 6, this could indicate the boundaries of micro-communities surrounding a minor issue or reflect a “radius of influence”–the distance to which the content of a user is still influential.

Arguably, connected individuals are more likely to retweet each other, increasing the similarity by copying content. To ensure that this was not a primary driver of measured similarity, we repeated the analysis while removing retweets and the corresponding webpages. The results, represented by dashed line in Figure 5, show slightly higher KL divergence, consistent with retweets contributing to a small portion of the observed correlation between network and content proximity.

Predicting Elections Results

In order to test the importance of content and structure to election outcome we devised a “win” model for all candidates in our dataset. However, we note that for this experiment we filtered out tweets that were created during and after Election Day and that the network was crawled during the hours prior to the beginning of the elections.

We built different logistic regression models where the dependent variable is the binary result of a race, indicating whether a candidate won or not. The independent variables[2] we used are described below:

- *Closeness-{in,out,all}* (Freeman 1979) – measuring the centrality of a candidate in the graph. Calculated as $n/\sum_{t\in V} d(v,t)$ where V is the set of all nodes reachable from v and $n = |V|$. $d(v,t)$ denotes the distance between v and t. In/out/all stands for incoming, outgoing or all paths.
- *HITS’ Authority score* (Kleinberg et al. 1999) and *PageRank* (Page et al. 1998) – measuring the relative importance of a node in the graph.
- *In/Out-degree* – number of edges to/from the node.
- *Incumbency* – Boolean variable indicating whether the candidate was incumbent or a challenger.
- *KL-party/corpus* – the KL-divergence between the LM of a user and the LM of his party/the entire corpus.
- *Party* – indicating the political group a user belongs to (Democrat, Tea-Party or Republican).
- *Same-party* – indicating whether the party of the candidate is the same as the party that last held the seat.
- *Tweets, hashtags, replies and retweets* – basic statistics of a candidate’s Twitter activity, as described above.

For all the graph properties we considered the whole graph consisting of all the candidates (experiments using only edges from the candidates own party yielded diminished accuracy). We start by examining each variable independently. Table 5 summarizes this set of experiments, showing each variable along with its coefficient, statistical significance and the accuracy of the model. We measured accuracy using a 10-fold cross-validation evaluation.

Table 5. Logistic Regression Model with Single Variables

Variable	Estimate	Prob(>|z|)	Accuracy
same_party	2.67	<0.0001	78.9%
incumbent	3.163	<0.0001	76.9%
indegree	0.252	<0.0001	74.6%
closeness_all	486.7	<0.0001	73.5%
kl-corpus	-0.281	<0.0001	66.7%
pagerank	486.7	<0.0001	66.4%
closeness_in	1017.2	<0.0001	64.7%
authority	0.442	<0.001	63.8%
republican	0.976	<0.0001	61.0%
teaparty	-0.277	0.38	61.0%
retweets	-0.00113	0.15	58.4%
hashtags	-0.00016	0.11	58.1%
tweets	-0.00022	0.08	57.8%
replies	-0.00026	0.08	57.5%
closeness_out	-20.9682	0.1	57.5%
outdegree	0.023	<0.01	57.5%
kl-party	-0.047	<0.05	55.9%

The first variable, “same-party”, indicates that guessing that a party will retain a seat correctly predicts 78.9% of the races. Incumbency is known to be a major factor in

[2] There are, of course, more sophisticated models for election prediction (e.g., Kastellec et al. 2008). However, our interest is specifically in understanding the importance of structural and content “centrality.”

winning elections, as is well reflected in the results. Closeness-all and in-degree are also predictive as opposed to closeness-out and out-degree, confirming that having followers is more important than following others.

An interesting finding is that KL-corpus is significantly more predictive than KL-party. The negative coefficient of these variables suggests that the more *similar* the LM of a user to the LM of the party/corpus, the more likely she is to be elected. We interpret this as meaning that focusing on centrist issues correlates more highly with winning than merely conforming to the agenda of one's own party (though both matter). Unsurprisingly, given Republicans' success in gaining seats in 2010, the Republican variable is predictive of winning. Finally, we see that simple usage statistics such as the number of tweets are uninformative. This result suggests that merely spamming Twitter is not a useful strategy.

In the last experiment we constructed a set of logistic regression models combining subsets of the variables described above. Table 6 presents the accuracy achieved by each model in 10-fold CV evaluation (with automated model selection applied). The results show that information hidden in graph structure and content can significantly improve the accuracy of election prediction (88% accuracy over 81% accuracy omitting Twitter-derived variables). Finally, we verified that the model performed similarly on Republicans as well as on Democrats.

Table 6. Logistic Regression Models

Name	Variables	Accuracy
All	tweets, kl-corpus, incumbent, party, closeness_all, closeness_out, same_party	88.0%
All but kl-corpus	tweets, corpus, incumbent, same_party party, closeness_all, closeness_out	85.5%
No content	incumbent, party, same_party, closeness_all, closeness_out	84.0%
No graph	tweets, kl-corpus, incumbent, party, same_party	83.8%
No graph & content	incumbent, party, same_party	81.5%

6. Discussion and Future Work

The model described above determines if any given candidate would win. Thus, in any given *race*, the model might find that neither or both candidates won. To test for the ability to predict race outcome we apply a simple scheme in which the most probable candidate is chosen as victor. As we do not have information for every candidate, only 63 races were used in this analysis. Applying this technique, we correctly predict 49 out of 63 (77.7%) of the races. Note that this is precisely .88 × .88, or the probability of picking one winner and one loser correctly. This result could likely be improved using better models or machine learning schemes such as joint inference.

Our findings suggest that the Republican Party, which made gains in the 2010 midterm election, succeeded in running a strong social media campaign on Twitter. This is consistent with the observations of Chittal (2010) and Stewart (2010). This is indicated by several metrics. First, the Republicans formed a denser graph of followers, and mentioned one another more often. Their tweets were also more topically similar, judging by the similarity of their language models. The top terms in the language models related to economic issues. In contrast, the network of Democratic candidate Twitter accounts was sparser, and their tweets were scattered over many topics, failing to convey a single coherent message.

Within the Republican Twitter network, the presence of the Tea Party members was boisterous. From their frequent use of hashtags and coherent language model, Tea Party members appeared to be running an organized Twitter campaign. This is somewhat surprising given the grassroots nature of this movement. However, a qualitative inspection of Tea Party messages and LM profile indicates a possible reason: members had joined forces on Twitter in attacking key Democrats.

Beyond allowing us to quantify political activity on Twitter, network and content variables are also predictive of election outcomes. Candidates whose tweets resembled that of many others in the corpus, that is, they were centrist in their topic selection rather than extremist, were more correlated with victory. Interestingly, based on the higher predictiveness of *KL-corpus* over *KL-party*, candidates are judged based on their position on the broad political spectrum rather than on intra-party positioning.

We also mention here one metric that was not predictive of election outcomes: the relation between the KL divergence of two opponents' LMs and the percentage of votes each candidate garnered. This suggests that perhaps it is more important how a candidate addresses more broadly discussed issues, than how much they mimic or try to differentiate themselves from their opponent. It is important to keep in mind that the KL divergence suggests an estimation of directionless distance. It would be interesting to repeat these experiments with a distance measure that also contains a notion of direction, to position candidates on the political spectrum.

Our content analysis is further limited in the sense that we relied on the Bag-of-Words model, ignoring the word meaning and the expressed sentiment. It is possible that sometimes users quote other users in order to mock them. In future work we plan to assign positive and negative weights to edges using sentiment analysis in order to improve the accuracy of our model. Additionally, we found that (in part) due to tweet length, an initial attempt to apply Latent Dirichlet Analysis (LDA) to the corpus failed to produce topics of high enough quality. We are pursuing other mechanisms for generating high quality topics.

Finally, the correspondence between network and content proximity suggests that homophily and social influence shape political candidates' activity on Twitter. By tracing the time-evolution of mentions and content, we might be able to approximate the range of individuals' influence within the network.

7. Conclusions

In this paper we studied the usage patterns of Twitter by candidates in the 2010 U.S. midterm elections. Our study addresses House, Senate and gubernatorial races as well as the virtual Tea Party. We incorporated structural and content analysis, and demonstrated the utility of using language modeling to estimate group cohesiveness as well as divergence of individuals. Our results indicate strong cohesiveness among conservatives, even for the largely unstructured Tea Party. We also find significant relationships between content, graph structure and election results by building a model that predicts whether a candidate will win or lose with accuracy of 88.0%. While we do not claim the use of Twitter determined the results, we do think a broader analysis over several campaigns could provide insight into what kinds of Twitter-based campaign activities are more effective.

8. Acknowledgments

We would like to thank Abe Gong for his helpful insights. This work was supported in part by NSF IIS-0746646.

References

Adamic, L. A. & Glance., N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *3rd Int. Workshop on Link Discovery*, 36–43.

Ammann, S. L., 2010, A Political Campaign Message in 140 Characters or Less: The Use of Twitter by U.S. Senate Candidates in 2010, http://ssrn.com/abstract=1725477

Blei, D. M., Ng. A.Y., and Jordan, M.I. 2003. Latent dirichlet allocation, *J. of Machine Learning Research*, 3:993-1022.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* vol. 16, no. 2, pp. 79–85.

Chittal, N. 2010. Twitter Reality: The Republicans are Crushing the Democrats When it Comes to Tweeting, AlterNet, August 13, 2010, http://www.alternet.org/story/147822/?page=1.

Diakopoulos, N. and Shamma, D. A. 2010. Characterizing Debate Performance via Aggregated Twitter Sentiment. CHI'10, 1195-1198.

Freeman, L.C. 1979. Centrality in social networks conceptual clarification. *Social Networks* 1(3):215-239

Golbeck, J., Grimes, J., and Rogers, A. 2010. Twitter use by the U.S. Congress, JASIST 61(8):1612-1621.

Glassman, M. E., Straus, J.R., and Shogan, C.J. 2010. Social Networking and Constituent Communications: Member Use of Twitter During a Two-Month Period in the 111th Congress, Congressional Research Service.

Honeycutt, C., and Herring, S. C. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. HICSS'09

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. 2009. Twitter Power: Tweets as Electronic Word of Mouth. *JASIST*, 60: 1–20.

Java, A., Song, X., Finin, T. and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. WebKDD and SNA-KDD, 56–65.

Jelinek, F. 1997. *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Kastellec, J.P., Gelman, A., and Chandler, J.P., 2008. Predicting and Dissecting the Seats-Votes Curve in the 2005 U.S. House Election, *Political Science*, 41:139-145.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632.

Metzler, D., Lavrenko, V. and Croft, W.B. 2004. Formal Multiple-Bernoulli Models for Language Modeling. SIGIR'04, 540–541.

Milgram, S. 1967. The Small World Problem. *Psychology Today*, 2:60-67.

Page, L., Brin, S., Motwani, R. and Winograd, T. 1998. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford University, Stanford, CA.

Ponte, J. & Croft, W. B. 1998. A Language Modeling Approach to Information Retrieval. SIGIR'98, 275–281.

Romero, D. M., Galuba, W., Asur, S. and Huberman, B. A. 2010. Influence and Passivity in Social Media. Technical Report. arXiv:1008.1253, CoRR, http://arxiv.org/abs/1008.1253.

Senak, M. 2010. Twongress: The Power of Twitter in Congress. White Paper, eyeonfda.com.

Shmueli-Scheuer, M., Roitman, H., Carmel, D., Mass, Y., & Konopnicki, D. 2010. Extracting User Profiles from Large Scale Data. Workshop on Massive Data Analytics on the Cloud, 1-6.

Smith, A. 2011. 22% of online Americans used social networking or Twitter for politics in 2010 campaign, Report of the Pew Internet Research Center.

Song, F. & Croft, W. B.1999. A General Language Model for Information Retrieval. CIKM'99, 316-321.

Sparks, D. B., Birds of a Feather Tweet Together: Partisan Structure in Online Social Networks, Presented at the 2010 meeting of the Midwest Political Science Association.

Stewart, M. 2010. House Republicans compete in new media challenge, CNN politicalticker, April 20, 2010.

Sugiyama, K., Hatano, K. and Yoshikawa, M. 2004. Adaptive Web Search Based on User Profile Constructed Without any Effort from Users. WWW'04, 675–684.

Tumasjan, A., Sprenger, T. O. , Sandner, P. G. and Welpe. I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM'10.

Williams, C., and Gulati, G. 2008. What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries. In Annual Meeting of the American Political Science Association, 1-17.Xue, G., Han, J., Yu, Y. and Yang, Q. 2009. User Language Model for Collaborative Personalized Search. *ACM Transactions on Information Systems* 27(2): 1-28.

Zhang, Y. and Koren, J. 2007. Efficient Bayesian Hierarchical User Modeling for Recommendation System. SIGIR'07, 47–54.