Predicting the Political Alignment of Twitter Users

Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini and Filippo Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington

Abstract—**The widespread adoption of social media for political communication creates unprecedented opportunities to monitor the opinions of large numbers of politically active individuals in real time. However, without a way to distinguish between users of opposing political alignments, conflicting signals at the individual level may, in the aggregate, obscure partisan differences in opinion that are important to political strategy. In this article we describe several methods for predicting the political alignment of Twitter users based on the content and structure of their political communication in the run-up to the 2010 U.S. midterm elections. Using a data set of 1,000 manually-annotated individuals, we find that a support vector machine (SVM) trained on hashtag metadata outperforms an SVM trained on the full text of users' tweets, yielding predictions of political affiliations with 91% accuracy. Applying latent semantic analysis to the content of users' tweets we identify hidden structure in the data strongly associated with political affiliation, but do not find that topic detection improves prediction performance. All of these content-based methods are outperformed by a classifier based on the segregated community structure of political information diffusion networks (95% accuracy). We conclude with a practical application of this machinery to web-based political advertising, and outline several approaches to public opinion monitoring based on the techniques developed herein.**

I. Introduction

Political advertising expenditures are steadily increasing [1], and are estimated to have reached four billion US dollars during the 2010 U.S. congressional midterm elections [2]. The recent 'Citizens United' Supreme Court ruling, which removed restrictions on corporate spending in political campaigns, is likely to accelerate this trend. As a result, political campaigns are placing more emphasis on social media tools as a low-cost platform for connecting with voters and promoting engagement among users in their political base.

This trend is also fueled in part by the fact that voters are increasingly engaging with the political process online. According to the Pew Internet and American Life Project, fully 73% of adult internet users went online to get news or information about politics in 2010, with more than one in five adults (22%) using Twitter or social networking sites for political purposes [3].

A popular microblogging platform with almost 200 million users [4], Twitter is an outlet for up-to-the-minute status updates, allowing campaigns, candidates and citizens to respond in real-time to news and political events. From the perspective of political mobilization, Twitter creates opportunities for viral marketing efforts that can be leveraged to reach audiences whose size is disproportionately large relative to the initial investment.

Of particular interest to political campaigns is how the scale of the Twitter platform creates the potential to monitor political opinions in real time. For example, imagine a campaign interested in tracking voter opinion relating to a specific piece of legislation. One could easily envision applying sentiment analysis tools to the set of tweets containing keyword relating to the bill. However, without the ability to distinguish between users with different political affiliations, aggregation over conflicting partisan signals would likely obscure the nuances most relevant to political strategy.

Here we explore several different approaches to the problem of discriminating between users with left- and right-leaning political alignment using manually annotated training data covering nearly 1,000 Twitter users actively engaged in the discussion of U.S. politics. Considering content based features first, we show that a support vector machine trained on user-generated metadata achieves 91% overall accuracy when tasked with predicting whether a user's tweets express a 'left' or 'right' political alignment. Using latent semantic analysis we identify hidden sources of structural variation in user-generated metadata that are strongly associated with individuals' political alignment.

Taking an interaction based perspective on political communication, we use network clustering algorithms to extract information about the individuals with whom each user communicates, and show that these topological properties can be used to improve classification accuracy even further. Specifically, we find that the community structure of the network of political retweets can be used to predict the political alignment of users with 95% accuracy.

We conclude with a proof of concept application based on these classifications, identifying the websites most frequently tweeted by left- and right-leaning users. We show that domain popularity among politically active Twitter users is not strongly correlated with overall traffic to a site, a finding that could allow campaigns to increase returns on advertising investments by targeting lower-traffic sites that are very popular among politically active social media users.

II. Background

A. The Twitter Platform

Twitter is a popular social networking and microblogging site where users can broadcast short messages called 'tweets' to a global audience. A key feature of this platform is that, by default, each user's stream of real-time posts is public. This fact, combined with its substantial population of users,

TABLE I
HASHTAGS RELATED TO #p2, #tcot, OR BOTH. TWEETS CONTAINING ANY OF THESE HASHTAGS WERE INCLUDED IN OUR SAMPLE.

Just #p2	#casen #dadt #dc10210 #democrats #du1 #fem2 #gotv #kysen #lgf #ofa #onenation #p2b #pledge #rebelleft #truthout #vote #vote2010 #whyimvotingdemocrat #youcut
Both	#cspj #dem #dems #desen #gop #hcr #nvsen #obama #ocra #p2 #p21 #phnm #politics #sgp #tcot #teaparty #tlot #topprog #tpp #twisters #votedem
Just #tcot	#912 #ampat #ftrs #glennbeck #hhrs #iamthemob #ma04 #mapoli #palin #palin12 #spwbt #tsot #tweetcongress #ucot #wethepeople

TABLE II
HASHTAGS EXCLUDED FROM THE ANALYSIS DUE TO AMBIGUOUS OR OVERLY BROAD MEANING.

Excl. from #p2	#economy #gay #glbt #us #wc #lgbt
Excl. from both	#israel #rs
Excl. from #tcot	#news #qsn #politicalhumor

renders Twitter an extremely valuable resource for commercial and political data mining and research applications.

One of Twitter's defining features is that each message is limited to 140 characters. In response to these space constraints, Twitter users have developed metadata annotation schemes which, as we demonstrate, compress substantial amounts of information into a comparatively tiny space. 'Hashtags,' the metadata feature on which we focus in this paper, are short tokens used to indicate the topic or intended audience of a tweet [5]; for example, #dadt for 'Don't Ask Don't Tell' or #jlot for 'Jewish Libertarians on Twitter.' Originally an informal practice, Twitter has integrated hashtags into the core architecture of the service, allowing users to search for these terms explicitly to retrieve a list of recent tweets about a specific topic.

In addition to broadcasting tweets to an audience of followers, Twitter users interact with one another primarily in two public ways: *retweets* and *mentions*. Retweets act as a form of endorsement, allowing individuals to rebroadcast content generated by other users, thus raising the content's visibility [6]. Mentions serve a different function, as they allow someone to address a specific user directly through the public feed, or to refer to an individual in the third person [7]. These two means of communication serve distinct and complementary purposes and together act as the primary mechanisms for explicit, public, user to user interaction on Twitter.

The free-form nature of the platform, combined with its space limitations and resulting annotation vocabulary, have led to a multitude of uses. Some use the service as a forum for personal updates and conversation, others as a platform for receiving and broadcasting real-time news and still others treat it as an outlet for social commentary and critical culture. Of particular interest to this study is the role of Twitter as a platform for political discourse.

B. Data Mining and Twitter

Owing to the fact that Twitter provides a constant stream of real-time updates from around the globe, much research has focused on detecting noteworthy, unexpected events as they rise to prominence in the public feed. Examples of this work include the detection of influenza outbreaks [8], seismic events [9], and the identification of breaking news stories [10]–[12]. These applications are similar in many respects to streaming data mining efforts focused on other media outlets, such as Kleinberg and Leskovec's 'MemeTracker' [13].

Its large scale and streaming nature make Twitter an ideal platform for monitoring events in real time. However, many of the characteristics that have led to Twitter's widespread adoption have also made it a prime target for spammers. The detection of spam accounts and content is an active area of research [14]–[16]. In related work we investigated the purposeful spread of misinformation by politically-motivated parties [17].

Another pertinent line of research in this area relates to the application of sentiment analysis techniques to the Twitter corpus. Work by Bollen *et al.* has shown that indicators derived from measures of 'mood' states on Twitter are temporally correlated with events such as presidential elections [18]. In a highly relevant application, Goorha and Ungar used Twitter data to develop sentiment analysis tools for the Dow Jones Company to detect significant emerging trends relating to specific products and companies [19]. Derivations of these techniques could be paired with the machinery from Section IV to accomplish the kind of real-time public opinion monitoring described in the introduction.

C. Data Mining and Political Speech

Formal political speech and activity have also been a target for data mining applications. The seminal work of Poole and Rosenthal applied multidimensional scaling to congressional voting records to quantify the ideological leanings of members of the first 99 United States Congresses [20]. Similar work by Thomas *et al.* used transcripts of floor debates in the House of Representatives to predict whether a speech segment was provided in support of or opposition to a specific proposal [21].

Related efforts have been undertaken for more informal, web-based political speech, such as that found on blogs and blog comments [22], [23]. While these studies report reasonable performance, the Twitter stream provides several advantages compared to blog data: Twitter provides a centralized data source, updated in real-time, with new sources automatically integrated into the corpus. Moreover, Twitter represents a broad range of individual voices, with tens of thousands of active contributors involved in the political discourse.

III. DATA AND METHODS

A. Political Tweets

The Twitter 'gardenhose' streaming API (dev.twitter.com/pages/streaming_api) provides a sample of about 10% of the entire Twitter feed in a machine-readable JSON format. Each tweet entry is composed of several fields, including a unique identifier, the text of the tweet, the time it was produced, the username of the account that produced the tweet, and in the case of retweets or mentions, the account names of the other users associated with the tweet.

This analysis focuses on six weeks of gardenhose data collected as part of a related study on political polarization [24]. The data cover approximately 355 million tweets produced during the period between September 14th and November 1st, 2010 — the run-up to the November 4th US congressional midterm elections.

Among all tweets, we consider as political communication any tweet that contained at least one politically relevant hashtag. Political hashtags were identified by performing a simple tag co-occurrence discovery procedure. We began by seeding our sample with two widely used left- and right-leaning political hashtags, `#p2` ("Progressives 2.0") and `#tcot` ("Top Conservatives on Twitter"). For each of these, we identified the set of hashtags with which it co-occurred in at least one tweet, and ranked the results using the Jaccard coefficient. For the set of tweets S containing a seed hashtag, and the set of tweets T containing another hashtag, the Jaccard coefficient between S and T is given by

$$\sigma(S,T) = \frac{|S \cap T|}{|S \cup T|}. \tag{1}$$

Thus, when the tweets in which a hashtag and seed both occur make up a large portion of the tweets in which either occurs, the two are similar. Using a similarity threshold of 0.005 we identified 66 unique hashtags, eleven of which were excluded due to overly-broad or ambiguous meaning (see Tables I and II.) The set of all tweets containing any one of these hashtags, 252 thousand in total, is used in all of the following analyses.

It's important to note that politically-motivated individuals often annotate content with hashtags whose primary audience would not likely choose to see such information ahead of time, a phenomenon known as *content injection*. As a result, hashtags in this study are frequently associated with users from both sides of the political spectrum, and therefore this seeding algorithm does not create a trivial classification scenario [24].

B. Communication Networks

From the set of political tweets we also construct two networks: one based on *mention* edges and one based on *retweet* edges. In the mention network, nodes representing users A and B are connected by a weighted, undirected edge if either user mentioned the other during the analysis period. The weight of each edge corresponds to the number of mentions between the two users. The retweet network is constructed in the same manner: an edge between A and B means that A retweeted B or viceversa, with the weight representing the number of retweets between the two. In total, the mention network consists of 10,142 non-singleton nodes, with 7,175 nodes in its largest connected component (and 119 in the next-largest). The retweet network is larger, consisting of 23,766 non-singleton nodes, with 18,470 nodes in its largest connected component (and 102 nodes in the next-largest).

TABLE III
CONTINGENCY TABLE OF INTER-ANNOTATOR AGREEMENT ON MANUAL CLASSIFICATIONS.

	Left	Ambiguous	Right
Left	303	51	23
Ambiguous	19	32	24
Right	22	59	423

TABLE IV
FINAL CLASS ASSIGNMENTS BASED ON RESOLUTION PROCEDURES DESCRIBED IN TEXT.

Left	Ambiguous	Right
373	77	506

C. Labeled Data

Let us now describe the creation of the labeled data used in this study for for training and testing our classifiers. We randomly selected a set of 1,000 users who were present in the largest connected components of both the mention and retweet networks. All users were individually classified by two annotators working independently of one another.

Each annotator assigned users to one of three categories: 'left', 'right', or 'ambiguous', based on the content of his or her tweets produced during the six week study period. The groups primarily associated with a 'left' political alignment were democrats and progressives; those primarily associated with a 'right' political alignment were republicans, conservatives, libertarians, and the Tea Party. Users coded as 'ambiguous' may have been taking part in a political dialogue, but it was difficult to make a clear determination about political alignment from the content of their tweets.

Using this coding scheme each of the annotators labeled 1,000 random users. Forty four accounts producing primarily non-English or spam tweets were considered irrelevant and excluded from this analysis. Table III shows the classifications of each annotator and their agreement.

Inter-annotator agreement is quite high for the 'left' and 'right' categories, but quite marginal for the 'ambiguous' category. This means that there were several users for whom one annotator had the domain knowledge required to infer a political alignment while the other did not. To address this issue we assigned a label to a user when either annotator detected information suggesting a political alignment in the content of a user's tweets. This mechanism was used to resolve ambiguity in 16% of users. Among the 956 relevant users in the sample there were 45 for whom the annotators explicitly disagreed about political alignment ('left' vs. 'right'). These individuals were included in the 'ambiguous' category.

After this resolution procedure, 373 users were labeled by the human annotators as expressing a 'left' political alignment, 506 users were labeled as 'right', and 77 were placed in the 'ambiguous' category, for a total of 956 users (Table IV). Ambiguous classifications are a typical result of scarce data at the individual level, but for completeness we report worst-case bounds on accuracy for the scenario in which all of these users are classified incorrectly.

IV. Classification

One of the central goals of this paper is to establish effective features for discriminating politically left- and right-leaning individuals. To this end we examine several features from two broad categories: user-level features based on content and network-level features based on the relationships between users. Each feature set is represented in terms of a feature-user matrix M, where M_{ij} encodes the value for feature i with respect to user j.

For content-based classifications we use linear support vector machines (SVMs) to discriminate between users in the 'left' and 'right' classes. In the simple case of binary classification, an SVM works by embedding data in a high-dimensional space and attempting to find the hyperplane that best separates the two classes [25]. Support vector machines are widely used for document classification because they are well-suited to classification tasks based on sparse, high-dimensional data, such as those commonly associated with text corpora [26].

To quantify performance for different feature sets we report the confusion matrix for each classifier, as well as *accuracy* scores based on 10-fold cross-validation. For a confusion matrix containing true left (tl), true right (tr), false left (fl) and false right (fr), the accuracy of a classifier is defined by:

$$accuracy = \frac{tl + tr}{tl + tr + fl + fr} \tag{2}$$

where tl is the number of left-leaning users who are correctly classified, and so on.

A. Content Analysis

1) Full-Text: To establish a performance baseline, we train a support vector machine on a feature-user matrix corresponding to the TFIDF-weighted terms (unigrams) contained in each user's tweets [27]. In addition to common stopwords we remove hashtags, mentions, and URLs from the set of terms produced by all users, a step we take to facilitate comparison with other feature sets. Additionally, we exclude terms that occur only once in the entire corpus because they carry no generalizable information and increase memory usage. After these preprocessing steps, the resulting corpus contains 13,080 features, each representing a single term.

To make it clear how we compute vectors for each user and his associated tweets let us define TFIDF in detail. The TFIDF score for term i with respect to user j is defined in terms of two components, term frequency (TF) and inverse document frequency (IDF). TF measures the relative importance of term i in the set of tweets produced by user j, and is defined as:

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}} \tag{3}$$

where n_{ij} is the number of times term i occurs in all tweets produced by user j, and $\sum_k n_{k,j}$ is the total number of terms in all tweets produced by user j. IDF discounts terms with high overall prominence across all users, and is defined as:

$$IDF_i = \log \frac{|U|}{|U_i|} \tag{4}$$

TABLE V
Summary of confusion matrices and accuracy scores for various classification features, with the sections in which they are discussed.

Features	Conf. matrix	Accuracy	Section
Full-Text	$\begin{bmatrix}266 & 107\\ 75 & 431\end{bmatrix}$	79.2%	§ IV-A1
Hashtags	$\begin{bmatrix}331 & 42\\ 41 & 465\end{bmatrix}$	90.8%	§ IV-A2
Clusters	$\begin{bmatrix}367 & 6\\ 38 & 468\end{bmatrix}$	94.9%	§ IV-B
Clusters + Tags	$\begin{bmatrix}366 & 7\\ 38 & 468\end{bmatrix}$	94.9%	§ IV-B

where U is the set of all users, and U_i is the subset of users who produced term i. A term produced by every user has no discriminative power and its IDF is zero. The product $TF_{ij} \cdot IDF_i$ measures the extent to which term i occurs frequently in user j's tweets without occurring in the tweets of too many other users.

The classification accuracy for this representation of the data is 79%, and its confusion matrix is shown in Table V. The lower accuracy bound for this approach, assuming that all ambiguous users are incorrectly classified, is 72.6%.

2) Hashtags: Hashtags emerged organically within the Twitter user community as a way of annotating topics and threads of discussion. Since these tokens are intended to mark the content of discussion, we might expect that they contain substantial information about a user's political leaning.

In this experiment we populate the feature-user matrix with values corresponding to the relative frequency with which user j used a hashtag i. This value is equivalent to the TF measure from Equation 3, but described in terms of hashtags rather than unigrams. We note that weighting by IDF did not improve performance. Eliminating hashtags used by only one user we are left with 4,701 features. For this classification task we report an accuracy of 90.8%; see Table V for the confusion matrix. The lower bound on this approach, assuming that all ambiguous users were misclassified, is 83.5%.

As evidenced by its higher accuracy score, a classifier that uses hashtag metadata outperforms one trained on the unigram baseline data. Analogous findings are observed in biomedical document classification, where classifiers trained on abstracts outperform those trained on the articles' full text [28]. The reasoning underlying this improvement is that abstracts are necessarily brief and information rich. In the same way, Twitter users must condense substantial semantic content into hashtags, reducing noise and simplifying the classification task.

3) Latent Semantic Analysis of Hashtags: Latent semantic analysis (LSA) is a technique used in text mining to discover a set of topics present in the documents of a corpus. Based on the singular value decomposition, LSA is argued to address issues of polysemy, synonym, and lexical noise common in text

TABLE VI

MOST EXTREME HASHTAG COEFFICIENTS FOR SECOND LEFT SINGULAR VECTOR. THIS LINEAR COMBINATION OF HASHTAGS APPEARS TO CAPTURE VARIANCE ASSOCIATED WITH POLITICAL ALIGNMENT.

Hashtag	Coeff.	Hashtag	Coeff.
#tcot	0.380	#p2	-0.914
#sgp	0.030	#dadt	-0.071
#ocra	0.020	#p21	-0.042
#hhrs	0.013	#votedem	-0.039
#twisters	0.012	#lgbt	-0.038
#tlot	0.011	#p2b	-0.032
#whyimvotingdemocrat	0.009	#topprog	-0.027
#rs	0.005	#onenation	-0.025
#ftrs	0.004	#dems	-0.023
#ma04	0.004	#gop	-0.021
#tpp	0.003	#hcr	-0.017

corpora [29]. Given a feature-document matrix, the singular value decomposition $U\Sigma V^t$, produces a factorization in terms of two sets of orthogonal basis vectors, described by U and V^t. The left singular vectors, U, provide a vector basis for terms in the factorized representation, and the right singular vectors, V, provide a basis for the original documents, with the singular values of matrix Σ acting as scaling factors that identify the variance associated with each dimension. In practice, LSA is said to uncover hidden topics present in a corpus, a claim supported by the analytical work of Papadimitriou *et al.* [30].

We apply this technique to the hashtag-user matrix in an attempt to identify latent factors corresponding to political alignment. The coefficients of the linear combination of hashtags most strongly associated with the second left singular vector, shown in Table VI, suggest that one is present in the data. Hashtags with extreme coefficients for this dimension include #dadt for 'Dont Ask Don't Tell', #p2 for Progressives 2.0, #tcot for Top Conservatives on Twitter, and #ocra for 'Organized Conservative Resistance Alliance.' The hashtag #whyimvotingdemocrat originally became a trending topic among left-leaning users, but was subsequently hijacked by right-leaning users to express sarcastic reasons they might vote for a Democratic candidate. A consequence of these coefficients is that users who use many left-leaning hashtags will have negative magnitude with respect to this dimension, and users who use many right-leaning hashtags will have positive magnitude in this dimension. Figure 1 shows clear separation between left- and right-leaning users in terms of the first and second right singular vectors.

A support vector machine trained on features describing users in terms of the first two right singular hashtag vectors does not improve accuracy compared to hashtag TF scores alone. Expanding the feature space to the first three LSA dimensions improves performance by an insignificant amount (about 0.1%), and the addition of subsequent features only degrades performance.

B. Network Analysis

The previous two feature sets are based on the *content* of each user's tweets. We might also choose to ignore this content entirely, focusing instead on the relationships between users.

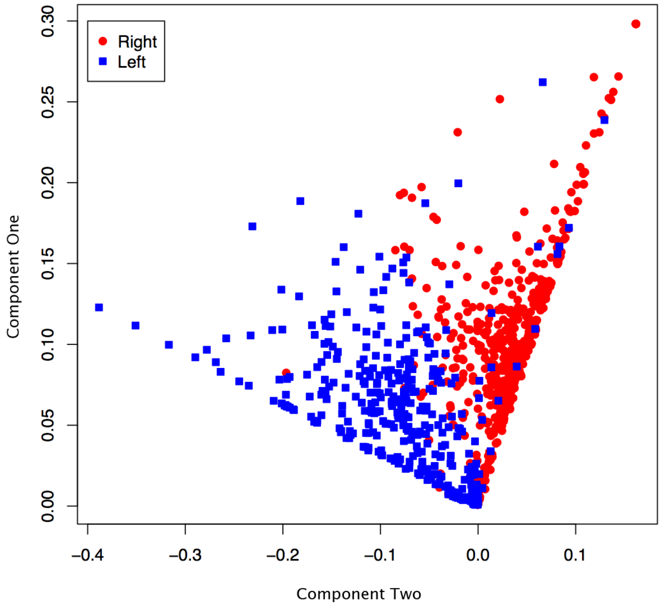

Fig. 1. Users plotted in the latent semantic space of the first and second right singular vectors. Colors correspond to class labels.

Many social networks exhibit homophilic properties — that is, users prefer to connect to those more like themselves — and as a consequence structural information can be leveraged to infer properties about nodes that tend to associate with one another [31], [32]. In the following, we focus on the largest connected component of the retweet network, as previous work suggests that it may tend to segregate ideologically-opposed users [24].

The cluster structure of the retweet network was established by applying a community detection algorithm using the label propagation method of Raghavan *et al.* [33]. Starting with an initial arbitrary label (cluster membership), this greedy method works by iteratively assigning to each node the label that is shared by most of its neighbors. Ties are broken randomly when they occur. Owing to this stochasticity, the label propagation method can return different cluster assignments for the same graph, even with the same initial conditions. Empirical analysis highlighted further instability resulting from random starting conditions: the algorithm easily converges to local optima.

To address this issue we used initial label assignments based on the clusters produced by Newman's leading eigenvector modularity maximization method for two clusters [34], rather than assigning labels at random. To verify that consistent clusters are produced across different runs of the algorithm for the same starting conditions, we repeated the analysis one hundred times and compared the label assignments produced at every run.

The similarity of two label assignments C and C' over a graph with n nodes can be computed by the Adjusted Rand Index (ARI) [35] as follows. Arbitrarily number the two clusters of C as c_1, c_2, and likewise number the clusters of C'

TABLE VII
MINIMUM, MAXIMUM, AND AVERAGE ARI SIMILARITIES BETWEEN 4,950 PAIRS OF CLUSTER ASSIGNMENTS COMPUTED BY LABEL PROPAGATION FOR THE RETWEET NETWORK.

Min	Max	Mean
0.94	0.98	0.96

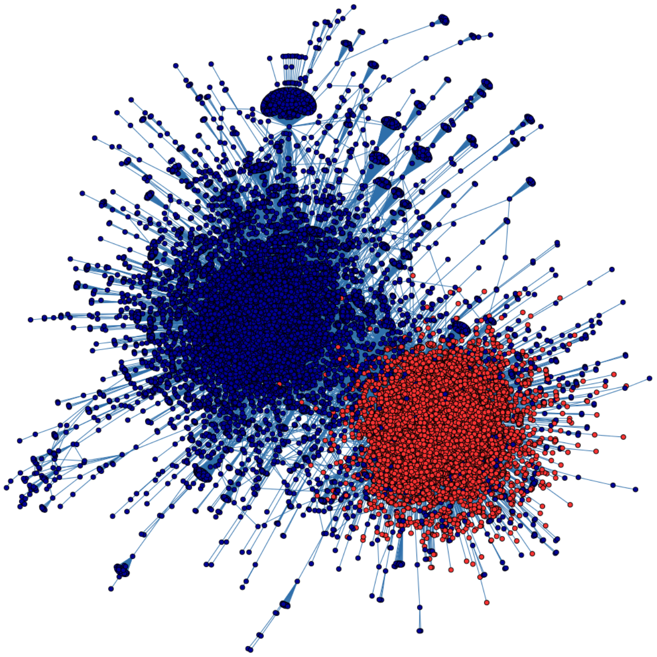

Fig. 2. The political retweet network, laid out using a force-directed algorithm. Node colors reflect cluster assignments (see text).

as c'_1, c'_2. Define then the *contingency matrix* $\mathbf{N}$, where $\mathbf{N}_{i,j}$ is the number of nodes of the graph simultaneously in c_i and c'_j. Define for this matrix the row sum $a_i = \mathbf{N}_{i,1} + \mathbf{N}_{i,2}$, and similarly the column sum $b_j = \mathbf{N}_{1,j} + \mathbf{N}_{2,j}$. The ARI is then defined by:

$$\mathrm{ARI}(C, C') = \frac{\sum_{ij} \binom{\mathbf{N}_{i,j}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \quad (5)$$

This measure varies from -1 when the two cluster assignments disagree for every item, to $+1$ when they are identical. A value of 0 indicates agreement at chance levels. Table VII reports the high average agreement between the 4,950 pairs of cluster assignments resulting from the 100 runs.

This method resulted in a cluster assignment for the retweet network with a modularity of 0.48, a high value indicating well-separated communities [34]. Figure 2 shows the retweet network laid out using a force-directed layout algorithm [36], with colors determined by the community to which each node was assigned. Table VIII reports the proportions of users labeled as 'left', 'right' and 'ambiguous' present in each retweet network cluster, and indicates that each network community exhibits clear partisan composition. Nodes in the left-dominated cluster are colored blue in Figure 2, while nodes in the right-dominated community are colored red. We note here that a similar procedure, applied to the network of mentions, does not reveal a clear partisan divide, and instead indicates that mentions form a communication bridge across which information flows between ideologically-opposed users [24].

TABLE VIII
PARTISAN COMPOSITION (AS DETERMINED BY MANUAL INSPECTION OF A RANDOM SAMPLE OF 1,000 USER PROFILES) AND SIZE OF THE TWO RETWEET NETWORK CLUSTERS.

	Left	Right	Ambiguous	Nodes
Cluster A	1.2%	93.4%	5.4%	7,115
Cluster B	80.1%	8.7%	11.1%	11,355

The strong association between cluster membership and political alignment suggests a simple classifier. This classifier would accept the cluster label of a user (either A or B), and assign that user to the 'left' if she was in cluster B, and the 'right' otherwise. The accuracy of this method is 95%; its confusion matrix is show in Table V. Assuming that all 'ambiguous' users are incorrectly classified yields a lower bound of 87.3% on the accuracy. This finding emphasizes the importance of the structure relative to content with respect to the loci of information about actors engaged in digitally-mediated communication.

We also experimented with combining topological information with the content data introduced earlier, resulting in a feature set comprised of cluster assignments and 19 hashtag features selected by Hall's feature selection algorithm [37]. Using a support vector machine, this method performed no better than using the network cluster label alone (Table V).

V. APPLICATIONS

Social media data has proven valuable for predictions in various domains, including politics [18], [38], [39]. The machine learning apparatus outlined above, paired with sentiment analysis techniques, could supplement traditional phone-based opinion surveys by allowing political campaigns to monitor public opinion regarding specific candidates and issues among users in their voting base [40]. Similarly, burst-detection mechanisms could be employed to detect a rise in prominence of a specific candidate or issue, allowing campaigns to shape marketing and messaging efforts in response to emerging topics and trends [11], [41], [42]. Critically, analyses of this nature depend on the ability to disambiguate users of different political identities, lest conflicting signals from users of opposing ideology cancel one another out at the aggregate level.

As an application of these data mining techniques consider buying decisions for web-based political advertising. Here we produce ranked lists of the domains most frequently tweeted by users of each political alignment, based on the predictions of the network classification method. Many Twitter users rely on URL shortening services to hash hyperlinks into a more

TABLE IX
WEBSITES MOST FREQUENTLY TWEETED BY LEFT- AND RIGHT-LEANING USERS, RANKED BY POPULARITY.

Popular Left	Popular Right
`feedproxy.google.com`	`feedproxy.google.com`
`mediamatters.org`	`hotair.com`
`politicalwind.com`	`gop2112.com`
`youtube.com`	`youtube.com`
`dailykos.com`	`redstate.com`
`truthy-out.org`	`firstthings.com`
`msnbc.msn.com`	`americanthinker.com`
`thinkprogress.org`	`google.com`
`harryreid.com`	`survivalstation.org`
`realclearpolitics.com`	`newsbusters.org`
`www.google.com`	`biggovernment.com`
`twitwall.com`	`realclearpolitics.com`
`thedailybeast.com`	`conservatives4palin.com`
`feeds.dailykos.com`	`newsmax.com`
`crooksandliars.com`	`nationalreview.com`

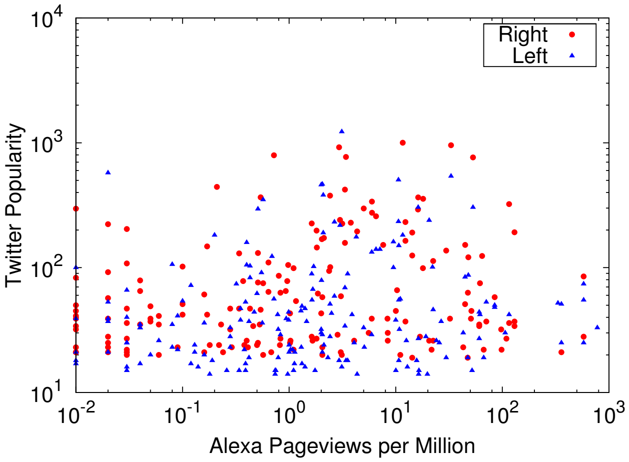

Fig. 3. Scatter plot of popularity of the 200 most frequently tweeted domains for members of left- and right-leaning network clusters versus global traffic among users of the Alexa toolbar.

compact format, and here we focus on links encoded using the popular *bit.ly* platform.

Ordered lists of the most popular domain names among left- and right-leaning users are presented in Table IX, and predictably tend to correspond to left- and right-leaning media. One exception is feedproxy.google.com, which is popular in both communities but is not politically aligned; it is the domain used for RSS feeds by Google Reader. Given these results, we emphasize that the domains most popular among left- and right-leaning Twitter users are not simply those with high traffic volume generally. Using the Alexa Web Information Service (aws.amazon.com/awis/) we obtained traffic statistics for each of the 200 most popular domains among users in each community. Alexa reports popularity in terms of pageviews per million impressions among users who have downloaded the Alexa toolbar plugin. Figure 3 suggests a weak correlation between the popularity among politically active Twitter users and global traffic volume. The Kendall's correlation coefficient between site popularity on Twitter and site popularity as measured by Alexa is $\tau = 0.12$ for sites popular among left-leaning users and $\tau = 0.14$ for right-leaning users.[1] These values confirm that the correlation is weak for both groups. Consequently, marketing efforts targeted at users of a specific alignment (for example, calls for campaign contributions and issue-specific ads targeted at mobilizing a political base) may achieve a higher return on investment by purchasing advertising on sites that are popular among social media users but have lower traffic from the internet population at large.

VI. CONCLUSIONS

In this article we used a rigorously constructed dataset to demonstrate that politically-active Twitters users generate text- and network-based information that can be used to effectively predict the political alignment of large numbers of individuals.

We detailed several approaches to this problem based on both content and network analysis. Techniques based on the statistical analysis of political communication networks provide the highest accuracy, thanks to the strong signal present in the partisan structure of the retweet network. However, we have shown that information-rich hashtag features are almost as effective at capturing political alignment, and have the benefit of generalizing without the need to recluster the network to accommodate new users.

Finally, as a proof of concept we illustrate the utility of this prediction capability by identifying the websites most popular among Twitter users from the political left and right, respectively. This approach reveals novel information about the popularity of different media outlets that can be leveraged to improve web-based advertising purchasing decisions.

We see much potential in the techniques described herein, which together represent a critical component in the real-time analysis of public opinion at the scale of tens of thousands of individual political actors. Looking forward, interesting open questions remain with respect to the generalizability of these approaches to international political discourse, multi-party systems, and the overall representativeness of communication on social media platforms. It is our hope that with continued refinement, techniques like these may yield insights into dynamics of political consensus and conflict throughout the world, perhaps even illuminating the role social media play in the process of political revolution.

Acknowledgments: We are grateful to Andrew Ratkiewicz and Matthew Francisco for their generous assistance in the annotation process, and to all the members of the Truthy project (truthy.indiana.edu). We thank Predrag Radivojac for his useful comments and analytical insights. We acknowledge support from the Lilly Foundation (Data to Insight Center Research Grant), the National Science Foundation (award CCF-1101743), the Pervasive Technology Institute, the Center for Complex Networks and Systems Research, and the School of Informatics and Computing at Indiana University, Bloomington.

[1] Since the data are broadly distributed, the assumption of normality required for computing the Pearson correlation coefficient does not hold. Therefore, we turn to a non-parametric test of dependence and use Kendall's τ to measure *rank* correlation.

REFERENCES

[1] L. Hau. (2007, December) Political ad spending set to climb sharply. Forbes.com. [Online]. Available: http://www.forbes.com/2007/12/07/politics-campaign-spending-biz-media-cx_lh_1207bizpolitics.html

[2] K. Wheaton. (2010, November) Political spending for midterm elections could top 4 billion. Advertising Age. http://adage.com/article?article_id=146818. [Online]. Available: http://adage.com/article?article_id=146818

[3] Pew Internet and American Life Project, "The internet and campaign 2010," Pew Research Center, Tech. Rep., 2011, http://pewresearch.org/pubs/1931/online-political-use-2010-over-half-us-adults. [Online]. Available: http://pewresearch.org/pubs/1931/online-political-use-2010-over-half-us-adults

[4] E. Schonfeld. (2010, June) Costolo: Twitter now has 190 million users tweeting 65 million times a day. TechChruch.com. [Online]. Available: http://techcrunch.com/2010/06/08/twitter-190-million-users/

[5] S. Yardi and d. boyd, "Dynamic debates: An analysis of group polarization over time on Twitter," *Bulletin of Science, Technology and Society*, vol. 20, pp. S1–S8, 2010.

[6] boyd, d., S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. of the Hawaii International Conference on Systems Sciences*, 2008, pp. 1–10.

[7] C. Honeycutt and S. C. Herring, "Beyond microblogging: Conversation and collaboration via Twitter," in *Proc. 42nd Hawaii International Conference on System Sciences*, 2008.

[8] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proc. of the 1st Workshop on Social Media Analytics*. ACM, 2010, pp. 115–122.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proc. of the 19th International Conference on the World Wide Web*. ACM, 2010, pp. 851–860.

[10] O. Phelan, K. McCarthy, and B. Smyth, "Using Twitter to recommend real-time topical news," in *Proc. of the Third ACM Conference on Recommender Systems*. ACM, 2009, pp. 385–388.

[11] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in *Proc. of NAACL*. Citeseer, 2010.

[12] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2009, pp. 42–51.

[13] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 497–506.

[14] S. Yardi, D. Romero, G. Schoenebeck, and boyd, d., "Detecting spam in a Twitter network," *First Monday*, vol. 15, no. 1, pp. 1–13, 2009.

[15] A. Wang, "Don't follow me: Spam detection in twitter," in *Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010)*, 2010.

[16] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. of the 19th International Conference on World Wide Web*, 2010, pp. 591–600.

[17] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Detecting and tracking the spread of astroturf memes in microblog streams," in *Proc. of 20th International World Wide Web Conference*, 2011.

[18] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," CoRR, Tech. Rep. arXiv:0911.1583, 2009.

[19] S. Goorha and L. Ungar, "Discovery of significant emerging trends," in *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 57–64.

[20] K. Poole and H. Rosenthal, "D-NOMINATE after 10 years: A comparative update to Congress: A political-economic history of roll-call voting," *Legislative Studies Quarterly*, vol. 26, no. 1, pp. 5–29, 2001.

[21] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 327–335.

[22] M. Efron, "Using cocitation information to estimate political orientation in web documents," *Knowledge and Information Systems*, vol. 9, no. 4, pp. 492–511, 2006.

[23] K. Durant and M. Smith, "Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection," in *Proc. of the 8th International Workshop on Knowledge Discovery on the Web (WebKDD)*. Springer, 2007, pp. 187–206.

[24] M. D. Conover, B. Gonçalves, J. Ratkiewicz, M. Francisco, A. Flammini, and F. Menczer, "Political polarization on twitter," in *Proceedings of the 5th InternationalConference on Weblogs and Social Media*, 2011.

[25] H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview," *Journal of Computational Biology*, vol. 10, no. 6, pp. 821–855, 2003.

[26] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. of the 10th European Conference on Machine Learning (ECML)*, 1998, pp. 137–142.

[27] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[28] A. Lourenço, M. D. Conover, A. Wong, F. Pan, A. Abi-Haidar, A. Nematzadeh, H. Shatkay, and L. Rocha, "Testing extensive use of NER tools in article classification and a statistical approach for method interaction extraction in the protein-protein interaction literature," *BMC Bioinformatics*, vol. in press, 2010.

[29] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[30] C. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proc. of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, 1998, pp. 159–168.

[31] L. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proc. of the 3rd International Workshop on Link Discovery (LinkKDD)*, 2005, pp. 36–43.

[32] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, pp. 415–444, 2001.

[33] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.

[34] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.

[35] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[36] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[37] M. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," Ph.D. dissertation, University of Waikato, August 2008.

[38] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[39] S. Asur and B. Huberman, "Predicting the future with social media," in *Proc of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2010, pp. 492–499.

[40] N. Diakopoulos and D. Shamma, "Characterizing debate performance via aggregated Twitter sentiment," in *Proc. of the 28th International Conference on Human Factors in Computing Systems (CHI)*. ACM, 2010, pp. 1195–1198.

[41] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[42] M. Mathioudakis and N. Koudas, "TwitterMonitor: trend detection over the Twitter stream," in *Proc. of the 2010 International Conference on Management of Data (SIGMOD)*. ACM, 2010, pp. 1155–1158.