This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

Quantifying Political Leaning from Tweets, Retweets, and Retweeters

Felix Ming Fai Wong, *Member, IEEE*, Chee Wei Tan, *Senior Member, IEEE*, Soumya Sen, *Senior Member, IEEE*, Mung Chiang, *Fellow, IEEE*

Abstract—The widespread use of online social networks (OSNs) to disseminate information and exchange opinions, by the general public, news media and political actors alike, has enabled new avenues of research in computational political science. In this paper, we study the problem of quantifying and inferring the political leaning of Twitter users. We formulate political leaning inference as a convex optimization problem that incorporates two ideas: (a) users are consistent in their actions of tweeting and retweeting about political issues, and (b) similar users tend to be retweeted by similar audience. We then apply our inference technique to 119 million election-related tweets collected in seven months during the 2012 U.S. presidential election campaign. On a set of frequently retweeted sources, our technique achieves 94% accuracy and high rank correlation as compared with manually created labels. By studying the political leaning of 1,000 frequently retweeted sources, 232,000 ordinary users who retweeted them, and the hashtags used by these sources, our quantitative study sheds light on the political demographics of the Twitter population, and the temporal dynamics of political polarization as events unfold.

Index Terms—Twitter, political science, data analytics, inference, convex programming, signal processing

1 Introduction

In recent years, big online social media data have found many applications in the intersection of political and computer science. Examples include answering questions in political and social science (e.g., proving/disproving the existence of media bias [3, 30] and the "echo chamber" effect [1, 5]), using online social media to predict election outcomes [46, 31], and personalizing social media feeds so as to provide a fair and balanced view of people's opinions on controversial issues [36]. A prerequisite for answering the above research questions is the ability to accurately estimate the political leaning of the population involved. If it is not met, either the conclusion will be invalid, the prediction will perform poorly [35, 37] due to a skew towards highly vocal individuals [33], or user experience will suffer.

In the context of Twitter, accurate political leaning estimation poses two key challenges: (a) Is it possible to assign meaningful numerical scores to tweeters of their position in the political spectrum? (b) How can we devise a method that leverages the scale of Twitter data while respecting the rate limits imposed by the Twitter API?

Focusing on "popular" Twitter users who have been retweeted many times, we propose a new approach that incorporates the following two sets of information to infer their political leaning.

Tweets and *retweets*: the target users' temporal patterns of being retweeted, and the tweets published by their retweeters. The insight is that a user's tweet contents should be consistent with who they retweet, e.g., if a user tweets a lot during a political event, she is expected to also retweet a lot at the same time. This is the "time series" aspect of the data.

Retweeters: the identities of the users who retweeted the target users. The insight is similar users get followed and retweeted by similar audience due to the homophily principle. This is the "network" aspect of the data.

Our technical contribution is to frame political leaning inference as a convex optimization problem that jointly maximizes tweet-retweet agreement with an error term, and user similarity agreement with a regularization term which is constructed to also account for heterogeneity in data. Our technique requires only a steady stream of tweets but not the Twitter social network, and the computed scores have a simple interpretation of "averaging," i.e., a score is the average number of positive/negative tweets expressed when retweeting the target user. See Figure 1 for an illustration.

Using a set of 119 million tweets on the U.S. presidental election of 2012 collected over seven months, we extensively evaluate our method to show that it outperforms several standard algorithms and is robust with respect to variations to the algorithm.

The second part of this paper presents a quantitative study on our collected tweets from the 2012 election, by first (a) quantifying the political leaning of 1,000 frequently-retweeted Twitter users, and then (b) using their political leaning, infer the leaning of 232,000 ordinary Twitter users. We make a number of findings:

- Parody Twitter accounts have a higher tendency to

- *Felix M.F. Wong was with the Department of Electrical Engineering, Princeton University. He is now with Yelp, Inc. Email: mwthree@princeton.edu*
- *Chee Wei Tan is with the Department of Computer Science, City University of Hong Kong. Email: cheewtan@cityu.edu.hk*
- *Soumya Sen is with the Department of Information & Decision Sciences, Carlson School of Management, University of Minnesota. Email: ssen@umn.edu*
- *Mung Chiang is with the Department of Electrical Engineering, Princeton University. Email: chiangm@princeton.edu*

Preliminary version in [51]. This version has substantial improvements in algorithm, evaluation and quantitative studies.

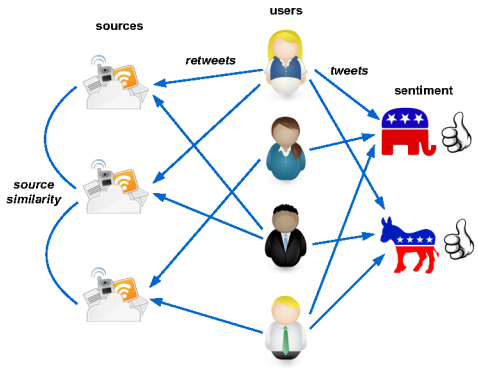


Fig. 1. Incorporating tweets and retweets to quantify political leaning: to estimate the leaning of the "sources," we observe how ordinary users retweet them and match it with what they tweet. The identities of the retweeting users are also used to induce a source similarity measure to be used in the algorithm.

be liberal as compared to other account types. They also tend to be temporally less stable.

- Liberals dominate the population of less vocal Twitter users with less retweet activity, but for highly vocal populations, the liberal-convservative split is balanced. Partisanship also increases with vocalness of the population.
- Hashtag usage patterns change significantly as political events unfold.
- As an event is happening, the influx of Twitter users participating in the discussion makes the active population more liberal and less polarized.

The organization of the rest of this paper is as follows. Section 2 reviews related work in studies of Twitter and quantifying political orientation in traditional and online social media. Section 3 details our inference technique by formulating political leaning inference as an optimization problem. Section 4 describes our dataset collected during the U.S. presidential election of 2012, which we use to derive ground truth for evaluation in Section 5. Then in Section 6 we perform a quantitative study on the same dataset, studying the political leaning of Twitter users and hashtags, and how it changes with time. Section 7 concludes the paper with future work.

2 Related Work

In political science, the ideal point estimation problem [41, 10] aims to estimate the political leaning of legislators from roll call data (and bill text [21, 22]) through statistical inference of their positions in an assumed latent space. The availability of large amounts of political text, e.g., legislative speeches, bill text and party statements, through electronic means has enabled the area of automated content analysis [24, 29] for political leaning estimation. The techniques from these works are not directly applicable to our work because of a disparity in data: political entities' stances on various issues can be directly observed through voting history and other types of data, but we do not have access to comparably detailed data for most Twitter users.

A variety of methods have been proposed to quantify the extent of bias in traditional news media. Indirect methods involve linking media outlets to reference points with known political positions. For example, Lott and Hassett [32] linked the sentiment of newspaper headlines to economic indicators. Groseclose and Milyo [25] linked media outlets to Congress members by co-citation of think tanks, and then assigned political bias scores to media outlets based on the Americans for Democratic Action (ADA) scores of Congress members. Gentzkow and Shapiro [20] performed an automated analysis of text content in newspaper articles, and quantified media slant as the tendency of a newspaper to use phrases more commonly used by Republican or Democrat members of the Congress. In contrast, direct methods quantify media bias by analyzing news content for explicit (dis)approval of political parties and issues. Ho and Quinn [26] analyzed newspaper editorials on Supreme Court cases to infer the political positions of major newspapers. Ansolabehere et al. [4] used 60 years of editorial election endorsements to identify a gradual shift in newspapers' political preferences with time.

Except for [20], the above studies require some form of manual coding and analysis, which is expensive and time-consuming. A more fundamental problem is data scarcity. Because the amount of data available for analysis is limited by how fast the media sources publish, researchers may need to aggregate data created over long periods of time, often years, to perform reliable analysis. Analyzing media sources through their OSN outlets offers many unprecedented opportunities with high volume data from interaction with their audience.

Political polarization has been studied in different types of online social media. Outside of Twitter, Adamic and Glance [1] analyzed link structure to uncover polarization of the political blogosphere. Zhou et al. [53] incorporated user voting data into random walk-based algorithms to classify users and news articles in a social news aggregator. Park et al. [39] inferred the political orientation of news stories by the sentiment of user comments in an online news portal. Weber et al. [49] assigned political leanings to search engine queries by linking them with political blogs. Regarding Twitter, political polarization was studied in [13]. Machine learning techniques have been proposed to classify Twitter users using, e.g., linguistic content, mention/retweet behavior and social network structure [43, 6, 2, 40, 11, 19]. Conover et al. [12] applied label propagation to a retweet graph for user classification, and found the approach to outperform tweet content-based machine learning methods.

Our problem of assigning meaningful political leaning scores to Twitter users is arguably more challenging than the above classification problem. There have already been several works on quantifying political leaning using the Twitter follower network. An et al. [3] and King et al. [27] applied multidimensional scaling on media sources with their pairwise distances computed from their mutual follower sets. Barberá [5] applied Bayesian ideal point estimation using following actions as observations. Golbeck and Hansen [23] proposed a graph-based method to propagate ADA scores of Congress members on Twitter to media sources through their followers. Weber et al. [50] quantified the political leaning of Twitter hashtags.

We argue that using retweet, rather than follower, data has its advantages. First, the huge sizes of most OSNs mean it is difficult for an average researcher to obtain an up-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

to-date snapshot of a network. The Twitter API prevents crawling the network beyond the one-hop neighborhood of a few thousand nodes.[1] On the other hand, our method requires only one connection to the real-time Twitter stream to collect retweets. Second, retweet data is more robust than follower data. Retweeting is often an act of approval or recognition [7],[2] but following has been shown to be of a different nature [9], and follower data obtained through crawling does not capture real-time information flow or contain fine-grained edge creation times.

Besides [12], retweet data have been applied in several recent works. Our retweet-based regularization is related to [16], which built a co-retweeted network for studying political polarization, and [48], which proposed a regularization framework using co-retweeting and co-retweeted information. Volkova et al. [47] built a series of Twitter social graphs to augment neighbors' features (also studied in [2] but not on retweet graphs) to improve performance. Compared to the above, our work (a) does not directly use a co-retweeted network but adds a matrix scaling preprocessing step to account for heterogeneity in Twitter users' popularity, (b) introduces the tweet-retweet consistency condition, and (c) performs a longitudinal study on our dataset collected over seven months.

3 Formulation

3.1 Motivation and Summary

To motivate our approach in using retweets for political leaning inference, we present two examples to highlight the existence of useful signals from retweet information.

From our dataset on the 2012 presidential election (see details in Section 4), we identify the Twitter accounts of two major media sources, one with liberal and the other with conservative leaning. In Figure 2 we plot their retweet popularity (their columns in matrix $\mathbf{A}$, see Section 3.2) during the 12 events in the dataset (see Table 1). We observe negative correlation ($\rho = -0.246$) between the two sources' patterns of being retweeted, especially during events 6 and 7.[3] This can be explained by Democrat/Republican supporters enthusiastically retweeting Romney/Obama-bashing tweets published by the media outlets during the corresponding events.

This example leads us to conjecture that: (a) the number of retweets received by a retweet source during an event can be a signal of her political leaning. In particular, one would expect a politically inclined tweeter to receive more retweets during a favorable event. (b) The action of retweeting carries implicit sentiment of the retweeter. This is true even if the original tweet does not carry any sentiment

1. Currently each authenticated client can make 15 requests in a 15-minute window, with each request returning at most 5,000 followers of a queried user. This is not to say scraping the complete network is impossible, but many tricks are needed [18].

2. This is even more so for our study, because the API ensures the retweets collected are unedited and not self-retweets, meaning the conversational and ego purposes of retweeting from [7] are not applicable.

3. Event 6: a video was leaked with Romney criticizing "47% of Americans [who pay no income tax]" at a private fundraiser. Event 7: the first presidential debate, where Obama was criticized for his poor performance.

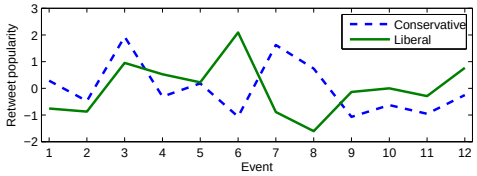


Fig. 2. Negatively correlated retweet patterns of two media Twitter accounts with opposite political leaning.

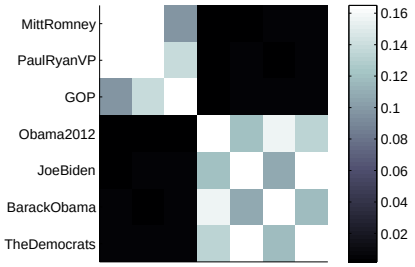


Fig. 3. Similarity of U.S. presidential election candidates by co-retweeter information. Clear separation is observed, with low similarity (dark areas) between two accounts in different parties.

itself. The intuition is that users tend to follow and retweet those who share similar political views, e.g., a user is more likely to retweet a newspaper to which it subscribes than any random newspaper, a manifestation of the homophily principle.

Our second example shows the identities of retweeters are a signal of one's political leaning. In Figure 3, we plot a portion of matrix $\mathbf{S}$, which stores the cosine similarity of any two Twitter accounts based on the overlap of their sets of retweeters (see Section 3.4 for details). By focusing on the election candidates[4] and official political party accounts, we see a clear separation of the two camps: two same-camp accounts have similarity that is at least 14 times of that between two different-camp accounts.

Given retweet and retweeter information are useful for inferring a Twitter account's political leaning, we formulate inference as a graph Laplacian-regularized least squares problem (Section 3.5) which consists of two steps. First, we assume that there is a large Twitter population that tweet and retweet at the same time, and the two forms of expressing political opinions are *consistent*. Then we frame political leaning estimation as a least squares (or linear inverse) problem in Section 3.3. Second, we add a regularization term to the least squares problem to ensure similar Twitter users, i.e., those having similar sets of audience who have retweeted them, have similar political leaning. We remark that naively building the regularization matrix results in poor performance. See Section 3.4 for how we carefully construct the matrix.

3.2 Definitions

Consider two political parties running for an election. During the election campaign there have been E *events* which attracted considerable attention in Twitter. We are interested

4. Obama2012 is Barack Obama's official campaign account, and BarackObama is his personal account. There is no such distinction for Mitt Romney's Twitter account(s).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

in quantifying the *liberal-conservative*[5] political leaning of N prominent retweet *sources*, e.g., media outlets' Twitter accounts and celebrities.

For event i, let U_i be the set of users who tweeted about the event, and T_{iu} be the set of tweets sent by user $u \in U_i$ about the event. Also define each tweet t to carry a score $s_t \in [-1, 1]$, such that it is 1 if the tweet shows full support for one candidate, or -1 if full support is for the other. Then for user u her approval score is

$$\sum_{t \in T_{iu}} \frac{s_t}{|T_{iu}|}.$$

Averaging over all users in U_i, the *average tweet leaning* y_i of event i is[6]

$$y_i = \frac{1}{|U_i|} \sum_{u \in U_i} \sum_{t \in T_{iu}} \frac{s_t}{|T_{iu}|}. \tag{1}$$

For source j, we quantify her political leaning as[7] $x_j \in \mathbb{R}$, interpreted as the *average* approval shown when someone retweets a tweet originating from j.

Now let V_i be the set of users who retweeted any one of the N sources during event i,[8] and $R_{uj}^{(i)}$ be the number of retweets sent by user u with the tweet originating from source j. Then the retweet approval score of user $u \in V_i$ is the average over all sources it has retweeted:

$$\sum_{j=1}^{N} \frac{R_{uj}^{(i)}}{\sum_{k=1}^{N} R_{uk}^{(i)}} x_j \tag{2}$$

and the *average retweet leaning* is the average over all u:

$$\begin{aligned} &\frac{1}{|V_i|} \sum_{u \in V_i} \sum_{j=1}^{N} \frac{R_{uj}^{(i)}}{\sum_{k=1}^{N} R_{uk}^{(i)}} x_j \\ =&\sum_{j=1}^{N} \Big(\frac{1}{|V_i|} \sum_{u \in V_i} \frac{R_{uj}^{(i)}}{\sum_{k=1}^{N} R_{uk}^{(i)}} \Big) x_j \\ =&\sum_{j=1}^{N} A_{ij} x_j, \end{aligned} \tag{3}$$

where A_{ij} is used to denote the inner summation term. The matrix $\mathbf{A}$ with elements A_{ij} can be interpreted as a Retweet matrix that captures the tweet-and-retweet response feature in Twitter.

3.3 An Ill-posed Linear Inverse Problem

The main premise of this paper is that the behavior of tweeting and retweeting is consistent. Mathematically, we require the average tweet and retweet leanings per event to be similar:

$$y_i \approx \sum_{j=1}^{N} A_{ij} x_j, \qquad i = 1, \ldots, E. \tag{4}$$

Our goal is to choose x_j's that minimize the error from the consistency equations Eq. (4), where the error measure is chosen to be the sum of squared differences $\sum_i (\sum_j A_{ij} x_j - y_i)^2$. Writing in matrix form, we are solving the unconstrained least squares problem

$$\underset{\mathbf{x} \in \mathbb{R}^N}{\text{minimize}} \qquad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2. \tag{5}$$

We often have many more tweeters than events ($N = 1000$, $E = 12$ in Sections 5 and 6), then $N > E$ and the system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ is underdetermined, which means there are infinitely many solutions $\mathbf{x}$ that can achieve the minimum possible error of 0 in Problem (5). Then the problem becomes an *ill-posed linear inverse problem* [8]. The challenge of solving ill-posed problems is in selecting a reasonable solution out of the infinite set of feasible solutions. For example, in our initial studies, the least-norm solution to (5) yielded unsatisfactory results.

3.4 Regularization

In statistical inference, solving ill-posed problems requires us to incorporate prior knowledge of the problem to rule out undesirable solutions. One such common approach is regularization, and we can change the objective function in Problem (5), $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$, to $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda f(\mathbf{x})$, where $\lambda > 0$ is a regularization parameter, and $f(\mathbf{x})$ quantifies the "fitness" of a solution such that undesirable solutions have higher $f(\mathbf{x})$ values. For example, Tikhonov regularization for least-squares uses $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ [8]. In this paper, we propose a regularization term that favors political leaning assignments $\mathbf{x}$ with x_i being close to x_j if sources i and j have similar retweet responses.

Let W_{ij} be a regularization weight between sources i and j such that $W_{ij} \geq 0$ and $W_{ij} = W_{ji}$. Furthermore, let $\mathbf{W}$ be the weight matrix whose elements are W_{ij}. Then we set

$$f(\mathbf{x}) = \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} (x_i - x_j)^2, \tag{6}$$

so that if W_{ij} is large (sources i and j are similar), then x_i should be close to x_j to minimize $W_{ij}(x_i - x_j)^2$.

Note that $f(\mathbf{x})$ can be rewritten in terms of a graph Laplacian. Let $\mathbf{D} = [D_{ij}]$ be defined as

$$D_{ij} = \begin{cases} \sum_{k=1}^{N} W_{ik} & i = j, \\ 0 & \text{otherwise}, \end{cases}$$

and $\mathbf{L}$ be the graph Laplacian defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Then it can be shown that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} (x_i - x_j)^2 = 2\mathbf{x}^T \mathbf{L} \mathbf{x}. \tag{7}$$

Our $\mathbf{W}$ is constructed to account for the following.

Similarity based on co-retweeter sets. The first step in constructing $\mathbf{W}$ is to construct a similarity matrix $\mathbf{S} = [S_{ij}]$ that captures the similarity between two sources i and j by

5. In our analysis of the 2012 U.S. presidential election, it is the Republican and Democratic Parties competing, and we assume liberals to support the Democrats/Obama, and conservatives to support the Republicans/Romney.

6. The specific forms of Eqs. (1) and (2) imply a user's contribution is limited in $[-1, 1]$ regardless of the number of tweets/retweets it sends. If we treat all tweets the same, i.e., defining $y_i = \sum_{u \in U_i} \sum_{t \in T_{ui}} s_t / \sum_{u \in U_i} |T_{iu}|$, the performance degrades probably due to the skew from a few highly vocal users.

7. We do not constrain x_j to be bounded in $[-1, 1]$, although x_j and y_i should be on the same scale, and a properly designed algorithm should be able to recover it.

8. In practice, we further restrict U_i and V_i to be the same user population by setting $U_i, V_i \leftarrow U_i \cap V_i$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

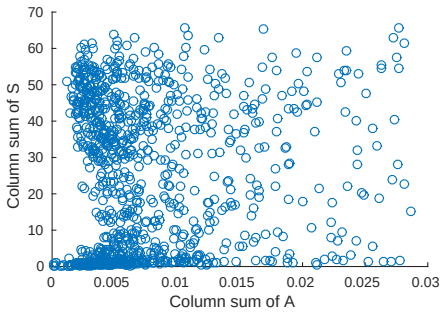


Fig. 4. Relationship between retweet popularity (column sum on $\mathbf{A}$) and regularization strength (column sum on $\mathbf{S}$) of top 1,000 retweet sources in Section 4. A large number of sources are on the bottom left corner, meaning they are both unpopular and insufficiently regularized.

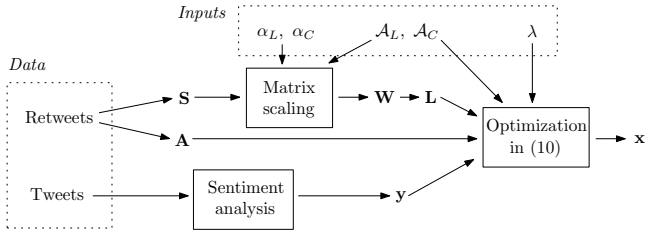


Fig. 5. Flowchart of data processing and inference steps.

who retweeted them. Let $\mathcal{U}_i$ and $\mathcal{U}_j$ be the sets of users who have retweeted sources i and j respectively. We consider two standard similarity measures:

- Cosine similarity: $S_{ij} = \dfrac{|\mathcal{U}_i \cap \mathcal{U}_j|}{\sqrt{|\mathcal{U}_i| \cdot |\mathcal{U}_j|}}$, and
- Jaccard coefficient: $S_{ij} = \dfrac{|\mathcal{U}_i \cap \mathcal{U}_j|}{|\mathcal{U}_i \cup \mathcal{U}_j|}$.

Individualizing regularization strength. Regularization is more important for sources with insufficient information available from $\mathbf{A}$: if source i does not get retweeted often, her corresponding column sum, $\sum_j A_{ji}$, is small, and inferring her score x_i based on the error term $\|\mathbf{Ax} - \mathbf{y}\|_2^2$ suffers from numerical stability issues.

In practice, a source can simultaneously be scarcely retweeted and have low similarity with other sources, i.e., source i has S_{ij} small for all other sources j. If $\mathbf{S}$ is directly used for regularization (i.e., by setting $\mathbf{W} \leftarrow \mathbf{S}$), the source is insufficiently regularized and inference becomes unreliable. See Figure 4 for confirmation from real data.

Our solution to this problem is to apply matrix scaling [44] to $\mathbf{S}$. We compute $\mathbf{W}$ as the matrix closest to $\mathbf{S}$ (under an Kullback-Leibler divergence-like dissimilarity function) such that the row and column sums of $\mathbf{W}$ satisfy equality constraints:

$$\begin{aligned} \underset{\mathbf{W}\geq 0}{\text{minimize}} \quad & \sum_{i,j=1}^{N} W_{ij} \log \frac{W_{ij}}{S_{ij}} && (8)\\ \text{subject to} \quad & W_{ij} = 0 && \text{for } (i,j) \text{ s.t. } S_{ij} = 0 \\ & \textstyle\sum_{j=1}^{N} W_{ij} = u_i && i = 1, \ldots, N \\ & \textstyle\sum_{j=1}^{N} W_{ji} = u_i && i = 1, \ldots, N, \end{aligned}$$

where $\mathbf{u} = \{u_i\}$ is a "regularization strength" parameter vector.

Problem (8) is solved by iterated rescaling of the rows and columns of $\mathbf{S}$ until convegence [44]. If the resultant $\mathbf{W}$ is asymmetric, we take the transformation $\mathbf{W} \leftarrow (\mathbf{W} + \mathbf{W}^T)/2$. It can be shown that the transformed $\mathbf{W}$ also satisfies the constraints in (8).

Incorporating prior knowledge. Prior knowledge can readily be incorporated into our inference technique as constraints of an optimization problem. In this paper we consider two types of prior knowledge:

- Anchors: sources carrying an extreme leaning, e.g., the election candidates themselves, can serve as anchors with fixed political leaning x_i. In the literature this idea has been used frequently [26, 3, 23].
- Score distribution: u_i can be interpreted as the strength of influence that source i exerts on sources similar to herself. Intuitively, anchors should exert higher influence, because of the size of their network and the propensity of their followers to retweet them. Therefore, we set $\mathbf{u}$ as:

$$u_i = \begin{cases} \alpha_L & i \in \mathcal{A}_L \\ \alpha_C & i \in \mathcal{A}_C \\ 1 & \text{otherwise,} \end{cases} \qquad (9)$$

where α_L and α_C are tuning parameters, and $\mathcal{A}_L$ and $\mathcal{A}_C$ are the sets of liberal and conservative anchors respectively. Given there should exist some non-anchor sources with extreme leaning, we tune α_L and α_C as follows: (a) initialize $\alpha_L = \alpha_C = 1$, and then (b) iteratively increase α_L and α_C until the computed most extreme political leaning of a non-anchor source is within 90% of that of an anchor.

3.5 Optimization Problem

Combining Eqs. (5), (6) and knowledge of anchors,[9] our final optimization formulation is:

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^N}{\text{minimize}} \quad & \|\mathbf{Ax} - \mathbf{y}\|_2^2 + \lambda \mathbf{x}^T \mathbf{L} \mathbf{x} && (10)\\ \text{subject to} \quad & x_i = -1 && i \in \mathcal{A}_L \\ & x_i = 1 && i \in \mathcal{A}_C, \end{aligned}$$

where λ is a tuning parameter. Figure 5 summarizes the data processing and inference steps in Section 3.

3.6 Extension: Sparsifying graph Laplacian.

Since most of the sources we consider are popular, most pairs of sources have at least one user who has retweeted both of them, and $\mathbf{S}$ is likely to be dense (in our dataset, 83% of its entries are nonzero). From a computational standpoint it is advantageous to sparsify the matrix, so we also evaluate our algorithm with an extra k-nearest-neighbor step, such that S_{ij} is kept only if j is a nearest neighbor of i, or vice versa. We are able to obtain good performance even when $\mathbf{S}$ is less than 10% sparse. See Section 6.4 for details.

9. We define a liberal (conservative) anchor to have a negative (positive) score to be consistent with the convention that liberals (conservatives) are on the "left" ("right").

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

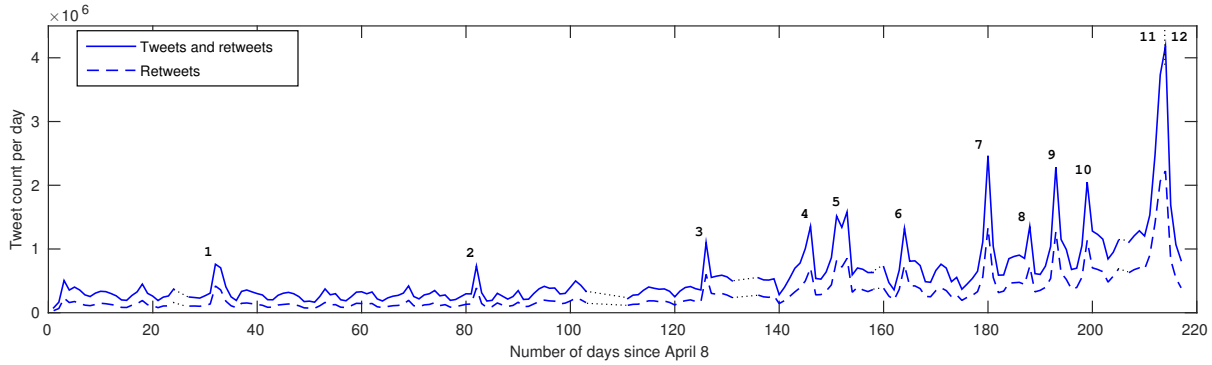


Fig. 6. Number of tweets per day. Numbers on plot indicate events (see Table 1), and dotted lines indicate time periods when significant data were lost due to network outage (five instances).

4 Dataset

In this section we describe the collection and processing of our Twitter dataset of the U.S. presidential election of 2012. Our dataset was collected over a timespan of seven months, covering from the initial phases to the climax of the campaign.

Data Collection. From April 8 to November 10 2012, we used the Twitter streaming API[10] to collect 119 million tweets which contain any one of the following keyword phrases: "obama", "romney", "barack", "mitt", "paul ryan", "joe biden", "presidential", "gop", "dems", "republican" and "democrat" (string matching is case-insensitive).

Event Identification. By inspecting the time series of tweet counts in Figure 6, we manually identified 12 events as listed in Table 1. We defined the dates of an event as follows: the start date was identified based on our knowledge of the event, e.g., the start time of a presidential debate, and the end date was defined as the day when the number of tweets reached a local minimum or dropped below that of the start date. After the events were identified, we extracted all tweets in the specified time interval[11] without additional filtering, assuming all tweets are relevant to the event and those outside are irrelevant.

Extracting Tweet Sentiment. We applied SentiStrength [45], a lexicon-based sentiment analysis package, to extract the sentiment of tweets. We adjusted the provided lexicon by compiling a high-frequency tweet-word list per event, and then removing words[13] that we consider to not carry sentiment in the context of elections. Sentiment analysis was done as a ternary (positive, negative, neutral) classification.

For each tweet t, we set its score $s_t = -1$ if either (a) it mentions solely the Democrat camp (has "obama", "biden" etc. in text) and is classified to have positive sentiment, or (b) it mentions solely the Republican camp ("romney", "ryan" etc.) and has negative sentiment. We set $s_t = 1$ if the opposite criterion is satisfied. If both criteria are not satisfied, we set $s_t = 0$.

10. https://dev.twitter.com/streaming/overview

11. For retweets, we only include those with the original tweet being created within the time interval.

12. A time interval starts at 00:00:00 of start date, and ends at 23:59:59 of end date. Timezone used is UTC.

13. They are "gay" (as in "gay marriage"), "foreign" ("foreign policy"), "repeal" ("repeal obamacare") and "battle*" ("battleship").

5 Evaluation

5.1 Ground truth construction

We compare the political leaning scores learnt by our technique with "ground truth" constructed by human evaluation. First, 100 sources are randomly selected from the 1,000 most popular retweet sources.[14] Then we ask 12 human judges with sufficient knowledge of American politics to classify each of the 100 sources as "L" (Liberal, if she is expected to vote Obama), "C" (Conservative, if expected to vote Romney), or "N" (Neutral, if she cannot decide), supposing each source is one voter who would vote in the presidential election. For each source, a judge is presented with (a) the source's user profile, including screen name, full name, self description, location and etc., and (b) ten random tweets published by the source. Given the set of labels, we compute our ground truth political leaning scores $\{\tilde{x}_i\}$ as follows: for each label of L/N/C, assign a score of $-1/0/+1$, then the score of source i, call it $\tilde{x}_i$, as the average of her labels.

While there are many alternatives to defining and constructing ground truth, our choice is motivated by our implicit assumption of Twitter political leaning being the perceived leaning by a source's retweeters. If source i has $\tilde{x}_i$ with extreme values (-1 or $+1$), then it is unambiguously liberal/conservative, but if $\tilde{x}_i$ takes intermediate values, then some human judges may be confused with the source's leaning, and the general Twitter population is likely to have similar confusion, which suggests that the "correct" x_i should also take intermediate values. Defining $\{\tilde{x}_i\}$ this way also allows us to understand the usefulness of quantifying political leaning with a continuous score. Obviously, if all $\tilde{x}_i$ are either -1 or $+1$, a simple binary classification of the sources is enough, but as we see in Figure 7, $\{\tilde{x}_i\}$ is evenly spread across the range of allowed values $[-1, 1]$, so characterizing sources with simple binary, or even ternary, classification appears too coarse. To further support our claim, we also compute the inter-rater agreement of our manual labels as Fleiss' $\kappa = 0.430$ [17], a moderate level of agreement [28]. This suggests that while the labels are reliable, classifying sources is not trivial and a continuous political leaning score is useful.

14. Those who were retweeted the largest cumulative number of times during the 12 events.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

TABLE 1
Summary of events identified in the dataset.

ID	Dates[12]	Description	# tweets (m)	# non-RT tweets (m)
1	May 9 - 12	Obama supports same-sex marriage	2.10	1.35
2	Jun 28 - 30	Supreme court upholds health care law	1.21	0.78
3	Aug 11 - 12	Paul Ryan selected as Republican VP candidate	1.62	0.96
4	Aug 28 - Sep 1	Republican National Convention	4.32	2.80
5	Sep 4 - 8	Democratic National Convention	5.81	3.61
6	Sep 18 - 22	Romney's 47 percent comment	4.10	2.55
7	Oct 4 - 5	First presidential debate	3.49	2.19
8	Oct 12 - 13	Vice presidential debate	1.92	1.19
9	Oct 17 - 19	Second presidential debate	4.38	2.67
10	Oct 23 - 26	Third presidential debate	5.62	3.35
11	Nov 4 - 6	Elections (before Obama projected to win)	7.50	4.40
12	Nov 7 - 9	Elections (after Obama projected to win)	6.86	4.43
Total			48.90	30.28

Finally, we also manually classify each of the 1,000 most popular sources into four classes:

- Parody: role-playing and joke accounts created for entertainment purposes (example joke tweet: "I cooked Romney noodles Obama self," a pun on "I cooked ramen noodles all by myself")
- Political: candidates of the current election and accounts of political organizations
- Media: outlets for distributing information in an objective manner, setting aside media bias issues
- Others: personal accounts, including those of celebrities, pundits, reporters, bloggers and politicians (excluding election candidates).

5.2 Performance metrics

The quality of political leaning scores is measured under two criteria.

Classification. One should be able to directly infer the liberal/conservative stance of a source i from her sign of x_i, i.e., it is liberal if $x_i < 0$, or conservative if $x_i > 0$. Taking $\{\tilde{x}_i\}$ as ground truth, we say source i is correctly classified if the signs of x_i and $\tilde{x}_i$ agree.[15] Classification performance is measured using the standard metrics of accuracy, precision, recall and F1 score.

Rank correlation. The set of scores $\{x_i\}$ induce a ranking of the sources by their political leaning. This ranking should be close to that induced by the ground truth scores $\{\tilde{x}_i\}$. We measure this aspect of performance using Kendall's τ, which varies from -1 (perfect disagreement) to 1 (perfect agreement).

5.3 Results

We solve Problem (10) with $\mathcal{A}_L = \{\text{Obama2012}\}$ and $\mathcal{A}_C = \{\text{MittRomney}\}$ and compare the results with those from a number of algorithms:

• PCA: we run Principal Components Analysis on $\mathbf{A}$ with each column being the feature vector of a source, with or without the columns being standardized, and take the first component as $\{x_i\}$. This is the baseline when we use only the information from $\mathbf{A}$ (retweet counts).

15. In the unlikely case of $\tilde{x}_i = 0$ (2 out of 100 test cases), we require $x_i = 0$ for correct classification.

• Eigenvector: we compute the second smallest eigenvector of $\mathbf{L}$, with $\mathbf{L}$ becoming computed from $\mathbf{S}$ being either the cosine or Jaccard matrix. This is a technique commonly seen in spectral graph partitioning [15], and is the standard approach when only the information from $\mathbf{S}$ (retweeters) is available. Note that the $\mathbf{x}$ computed this way is equivalent to solving the optimization problem: minimize $\mathbf{x}^T\mathbf{L}\mathbf{x}$, subject to $\|\mathbf{x}\|_2 = 1,\ \mathbf{x}^T\mathbf{1} = 0$.

• Sentiment analysis: we take x_i as the average sentiment of the tweets published by source i, using the same methdology in computing $\mathbf{y}$ [45]. This is the baseline when only tweets are used.

• SVM on hashtags: following [12], for each source we compute its feature vector as the term frequencies of the 23,794 hashtags used by the top 1,000 sources. We then train an SVM classifier (linear kernel, standardized features) using the 900 of the top 1,000 sources that are not labeled by 12 human judges (see Section 5.1) as training data.[16] Hashtags have been suggested to contain more information than raw tweet text and a better source of features [12].

• Retweet network analysis: also following [12], we construct an undirected graph of the 25,000 Twitter users with highest retweet activity. The edge between two users is weighted as the number of times either one has retweeted the other. Then we apply majority voting-based label propagation [42] with initial conditions (label assignments) from the leading eigenvector method for modularity maximization [38]. Given the modularity maximization method used here is analogous to the above eigenvector baseline, we treat its output as political leaning scores and report its performance. We also experimented with synchronous soft label propagation [54], similar to the algorithm in [19], but it did not produce better results.

Table 2 reports the evaluation results. Our algorithm, in combining information from $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{y}$, performs significantly better than all other algorithms in terms of Kendall's τ, F1 score and accuracy. We also observe that if no matrix scaling is applied in constructing $\mathbf{W}$, the algorithm tends to assign all $\{x_i\}$ (except those of anchors) to the same sign, resulting in poor classification performance. In the remaining of this paper, we focus on the political leaning scores computed using cosine similarity.

16. These sources have at least one label from one of the 12 judges. To reduce the risk of noise, we include in training only 712 sources that have unambiguous (L or C) labels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

TABLE 2
Performance of our algorithm compared to others. Best two results (almost always due to our method) are highlighted in bold.

Algorithm	Kendall's τ	Precision, L	Recall, L	Precision, C	Recall, C	F1 score, L	F1 score, C	Accuracy
Ours, cosine matrix	**0.652**	**0.942**	**0.970**	**0.935**	**0.935**	**0.955**	**0.935**	**0.94**
Ours, Jaccard matrix	**0.654**	**0.940**	0.940	**0.879**	**0.935**	**0.940**	**0.906**	**0.92**
Ours, cosine w/o scaling	0.649	0.670	**1**	0	0	0.802	0	0.67
Ours, Jaccard w/o scaling	0.641	0	0	0.31	**1**	0	0.473	0.31
PCA	0.002	0.663	0.791	0.300	0.194	0.721	0.235	0.59
PCA, standardized columns	0.011	0.750	0.224	0.325	0.839	0.345	0.468	0.41
Eigenvector, cosine matrix	0.297	0.667	0.985	0	0	0.795	0	0.66
Eigenvector, Jaccard matrix	0.308	0.663	0.970	0	0	0.787	0	0.65
Sentiment analysis	0.511	0.926	0.746	0.700	0.903	0.826	0.789	0.78
SVM on hashtags	0.436	0.863	0.851	0.840	0.677	0.857	0.750	0.78
Modularity maximization	0.510	0.934	0.851	0.763	0.935	0.891	0.841	0.86
Label propagation + modularity	—	**0.940**	0.940	**0.935**	**0.935**	**0.940**	**0.935**	**0.92**

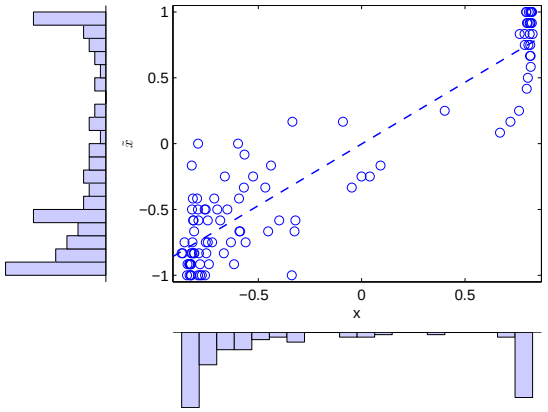


Fig. 7. Relationship of ground truth $\{\tilde{x}_i\}$ and our computed scores $\{x_i\}$ on the 100 sources with manual labels, together with their marginal distributions. Our method is able to recover both the correct classifications (datapoints in bottom-left and upper-right quadrants) and rankings for most sources.

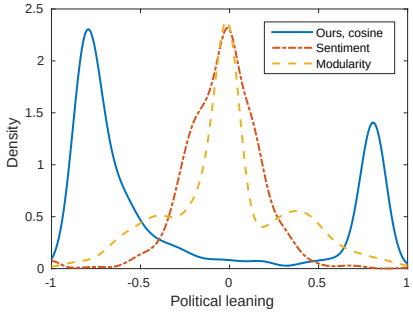


Fig. 8. Kernel density estimates of political leaning of top 1,000 retweet sources produced by different algorithms.

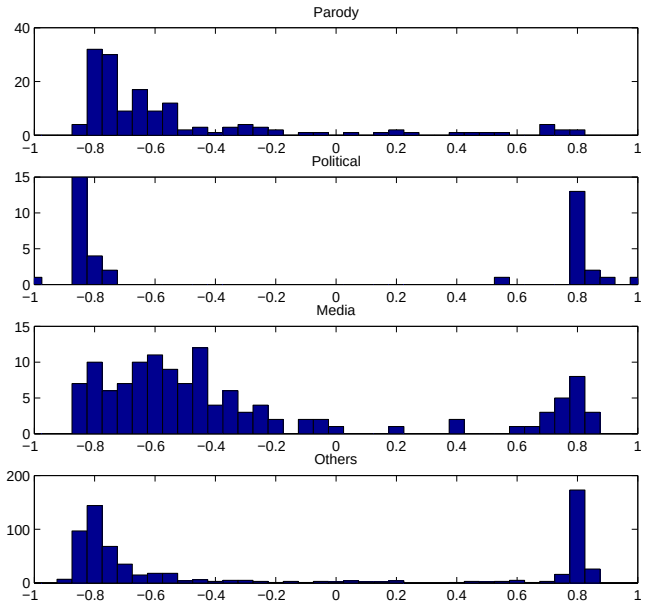


Fig. 10. Distribution of political leaning scores grouped by Twitter account type. Parody/comedy accounts are skewed towards the liberal side. Political accounts are more polarized (no scores close to zero) than other accounts.

Among the other algorithms, sentiment analysis, a content-based method, performs the best in terms of Kendall's τ, i.e., preserving the relative ordering between sources. This is somewhat surprising considering the findings from related work that apply linguistic features [34, 12, 40]. In terms of classification, we find the retweet network-based methods to perform best. In particular, a combination of modularity maximization and label propagation produces precision and recall values close to those due to our method. However, due to the nature of majority voting, the algorithm outputs only binary labels and cannot be used to compare sources of the same class.

Figure 7 shows the correspondence between $\{x_i\}$ and $\{\tilde{x}_i\}$. The two sets of scores exhibit similar rankings of sources and a bimodal score distribution. The scores due to our algorithm are slightly more polarized. We note that our algorithm is the only one in Table 2 that produces a bimodal score distribution. Figure 8 compares our score distribution with those produced by sentiment analysis and modularity maximization, both with high Kendall's τ, showing the two methods produce distributions with a peak at around 0. Other algorithms tend to compress almost all scores into a small range at around 0, and leave one or few scores with very large values.

6 Quantitative Study

6.1 Quantifying Prominent Retweet Sources

We study the properties of the political leaning of the 1,000 most popular retweet sources. Similar to that in Figure 7, the score histogram on the full set (Figure 9) has a bimodal distribution. We note that by incorporating retweeter infor-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

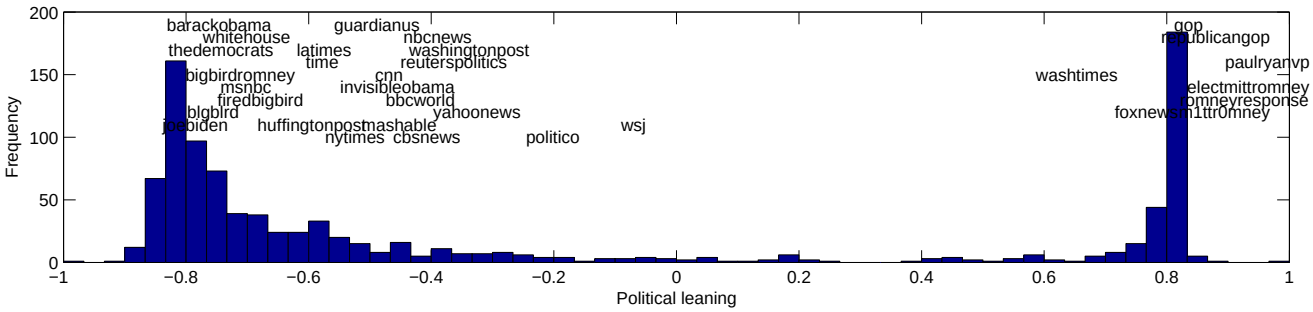


Fig. 9. Distribution of political leaning scores of top 1,000 retweet sources with positions of some example sources.

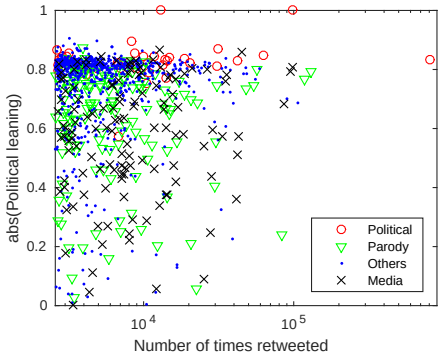


Fig. 11. Relationship between retweet popularity and score polarity.

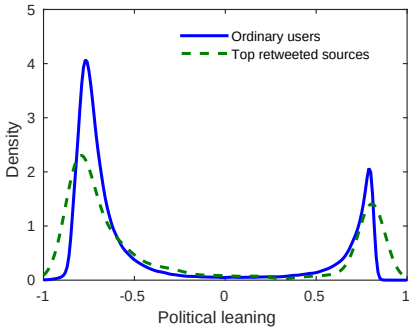


Fig. 12. Kernel density estimates of political leaning of top 1,000 retweet sources and 232,000 ordinary Twitter users.

mation, our algorithm is able to correctly position "difficult" sources that were highly retweeted during events unfavorable to the candidate they support, e.g., JoeBiden, CBSNews and all accounts related to Big Bird, an improvement over the preliminary version of this paper [51]. We also find that WSJ is assigned a slightly liberal score. This is consistent with findings reported in prior studies [25, 32] explained by the separation between WSJ's news and editorial sections.

We also study the score distributions of sources grouped by account type. Figure 10 shows noticeable differences among the different groups. Parody sources are skewed towards the liberal side. Political sources are strongly polarized with no sources having neutral (close to zero) scores. Media sources are less polarized with a more even spread of scores. Sources in the "Others" class have a score distribution close to that of political sources, but also includes a few neutral scores, which can be attributed to celebrities with no clear political stance. To check the differences are not due to artifacts of our algorithm, e.g., a possibility of biasing sources with more available data towards the ends of the spectrum, we plot the relationship between the number of times a source is retweeted and its score polarity in Figure 11, and find no significant correlation between the variables ($\rho = 0.0153$, p-value$= 0.628$).

6.2 Quantifying Ordinary Twitter Users

Given the political leaning of 1,000 retweet sources, we can use them to infer the political leaning of ordinary Twitter users who have retweeted the sources. We consider the set of users seen in our dataset who have retweeted the sources at least ten times, including retweets made during non-event time periods. In total there are 232,000 such users. We caution this set of users is not necessarily representative of the general Twitter population, or even the full population of our dataset (9.92 million users in total), but we believe it is possible to "propagate" score estimates from these 232,000 users to everyone else, which remains as future work.

For user u, we infer her political leaning x_u as

$$x_u = \frac{\sum_{i=1}^{N} R_{ui} x_i}{\sum_{i=1}^{N} R_{ui}}, \tag{11}$$

where R_{ui} is the number of times user u retweeted source i and x_i is source i's political leaning.

Figure 12 shows the kernel density estimate of the political leaning scores of the 232,000 ordinary users, compared with that of the set of 1,000 sources. These users have a slightly less polarized distribution (density function is closer to zero), but are more skewed towards the liberal side (72.5% have $x_u < 0$, compared to 69.5% for retweet sources).

Measuring polarization. Using the learnt political leaning scores, we aim to quantify political polarization of a population, but for this to be possible, we first need a polarization measure. Let us consider one user u. A natural measure P_u for her polarization can be defined as how far her political leaning is away from neutral: $P_u = |x_u - 0| = |x_u|$. Then for a population $\mathcal{U}$, its polarization measure can be taken as the average of all P_u for $u \in \mathcal{U}$. However, such a definition does not account for class imbalance (liberals outnumbering conservatives in our case), so we take a class-balanced definition instead:

$$P_{\mathcal{U}} = \frac{1}{|\mathcal{U}_+|} \sum_{u \in \mathcal{U}_+} x_u + \frac{1}{|\mathcal{U}_-|} \sum_{u \in \mathcal{U}_-} |x_u|, \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

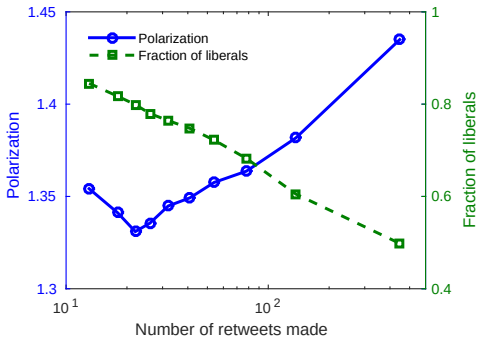


Fig. 13. Skew measures of ordinary users binned by 10%-tiles of rewteet activity. Polarization and liberal-conservative balance increase for more active populations.

where $\mathcal{U}_+ = \{u \mid u \in \mathcal{U},\ x_u > 0\}$ and $\mathcal{U}_- = \{u \mid u \in \mathcal{U},\ x_u < 0\}$.

Now we put the 232,000 ordinary users into ten percentile bins, such that the first bin contains the lowest 10% of users according to their retweet activity (number of retweets made, including retweets of non-top 1,000 sources), then the next bin contains the next lowest 10%, and so on. Figure 13 shows the plot of two skew measures: the polarization measure as defined in Eq. (12), and the fraction of liberals $|\mathcal{U}_+|/(|\mathcal{U}_+ + \mathcal{U}_-|)$. We observe that liberals dominate (>80%) the population of low activity users, but as retweet activity increases, the liberal-conservative split becomes more balanced (roughly 50% in the last bin). This suggests that most Twitter users (80% of them make less than 100 retweets on the election) tend to be liberals, but the most vocal population (those making 100 to 33,000 retweets) consists of users who are more politically opiniated such that they spend more effort to promote their causes in social media. This is supported by the plot of the polarization measure, which increases with retweet activity.

6.3 Temporal Dynamics

With the large amount of data available from Twitter one can perform fine-grained temporal analysis using data from a relatively short timespan. In this section, we study how many events are necessary to obtain sufficiently good performance, and then present two examples in applying our methodology to social media monitoring.

Stability. Here we quantify the political leaning of the 1,000 sources studied in Section 6.1 but with varying amounts of information. We start by running our inference technique using data from only the first event, then we use events 1 to 2, and so on. Then each source has a sequence of 12 political leaning scores, and we evaluate the quality of these scores compared to ground truth. Figure 14(a) shows that three events, or 10% of the data, are enough for achieving performance close to that from using the full dataset. From the description of the events in Table 1, the first two events are skewed towards the liberal side, and it is not clear how conservative users react to them (either object strongly or remain silent). With the third event added, we have a more balanced set of data for reliable inference.

Next we look at the stability of scores per source across the 12 events. Figure 14(b) plots the mean deviation of a source's score per event from her final score. Using the classification from Section 5, we see stability varies with the type of a source. Political sources are the most stable with their score deviations having decreased quickly after three events, consistent with our intuition that they are the easy cases with the least ambiguity in political leaning. On the other hand, parody accounts are the least stable. Their patterns of being retweeted are the least stable with large fluctuations in retweet counts across different events, resulting in less reliable inference. Also, users do not retweet parody sources by how agreeable, but rather by how funny they are, so even retweeter information on these sources is less stable across different events.

Trending hashtags. The political leaning of hashtags [50] can be quantified by how they are being used by Twitter users. Here we consider a simple way to estimate it using the political leaning scores of the top retweet sources. For each hashtag, we compute its political leaning as the average of all sources (within the top 1000 retweet sources) that have used it at least once in their published tweets. Moreover, we perform this computation per event (excluding previous events) to obtain a sequence of hashtag political leaning scores, so as to track trending hashtags as events unfold.

Tables 3 and 4 are the lists of the most liberal or conservative hashtags that have been used by at least ten sources in each of the events. Besides the static hashtags indicating users' political affiliation such as #TCOT (top conservatives on twitter) and #p2b (progressive 2.0 bloggers), we discover hashtags being created in response to events (#ssm (same-sex marriage) in event 1, #MyFirstTime in event 10 for a suggestive Obama ad). Moreover, there are hashtags showing opposite opinions on the same issue (#aca (affordable care act) vs #ObamaTax in event 2, #WrongAgainRyan vs #BidenUnhinged in event 8), and many instances of sarcasm (#StopIt and #YouPeople in event 6, #notoptimal in event 9 for Obama saying the government's response was "not optimal" during the Benghazi attack) and accusations (#MSM (main stream media) for liberal media bias).

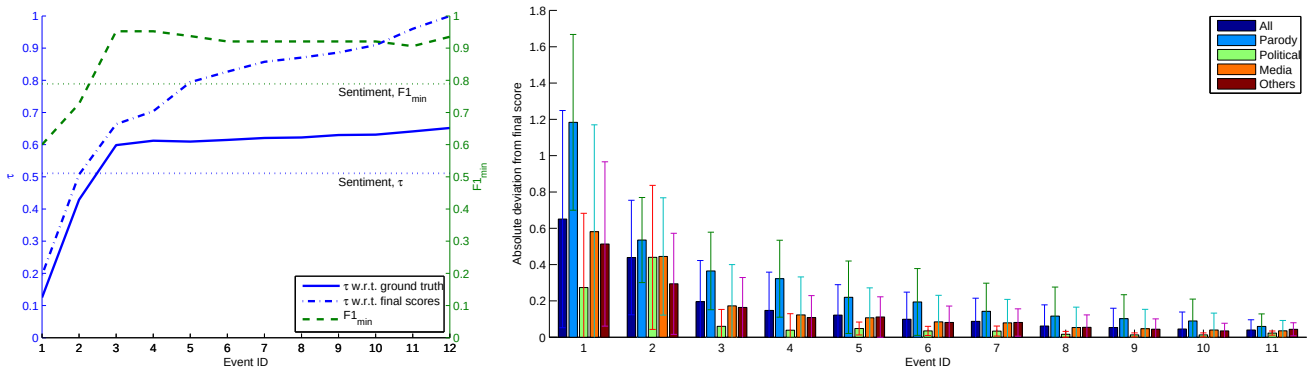
In the latter half of the dataset, liberals tend to focus on accusing the other side of lying during the debate (#LyinRyan, #SketchyDeal, #MittLies, #AdmitItMitt), while conservatives tend to focus on the Benghazi attack (#BenghaziGate, #notoptimal, #7HoursOfHell, #Benghazi). Finally, we note a change in hashtag usage before and after the election outcome came out (from #WhyImNotVotingForRomney to #FourMoreYears).

Tracking temporal variation in polarization. We begin with this question: is the Twitter population more (or less) polarized during an event?

Without analyzing the data, the answer is not clear because it is influenced by two factors with opposite effects. On one hand, an event draws attention from less vocal users who are likely to have weak political leaning, and join the discussion because everyone talks about it. On the other hand, the fact that a usually silent user joining the discussion may indicate she is strongly opinionated about the topic.

To answer the question, we compute the polarization measure in Eq. (12) for each day in our dataset, with population $\mathcal{U}$ taken as the set of users (out of the 232,000 users from Section 6.2) who have tweeted or retweeted on that day. For comparison purposes we also compute the fraction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering



(a) Performance measures. $F1_{\min}$ is the minimum of F1 scores of the C and L classes.

(b) Deviation from final result, with grouping by account type. Error bars indicate one standard deviation from the group average.

Fig. 14. Stability of results with varying number of events: good performance is achieved with only the first three events, and parody accounts are the least stable.

TABLE 3
Hashtags with highest liberal political leaning per event.

Event	Top five liberal hashtags				
1	#MarriageEquality	#equality	#marriageequality	#edshow	#ssm
2	#aca	#p2b	#ACA	#Obama2012	#edshow
3	#p2b	#Medicare	#topprog	#edshow	#Obama2012
4	#p2b	#YouPeople	#Current2012	#msnbc2012	#uppers
5	#ACA	#PaulRyan	#LyinRyan	#nerdland	#DavidGregorysToughQuestions
6	#LyinRyan	#ObamaBiden2012	#StopIt	#p2b	#YouPeople
7	#47Percent	#47percent	#Forward	#TeamObama	#topprog
8	#WrongAgainRyan	#p2b	#Bain	#Sensata	#LyinRyan
9	#SketchyDeal	#Bain	#MittLies	#p2b	#BinderFullofWomen
10	#StrongerWithObama	#RomneyWrong	#ObamaBiden2012	#RomneyNotReady	#AdmitItMitt
11	#p2b	#uppers	#msnbc2012	#ObamaBiden2012	#WhyImNotVotingForRomney
12	#OBAMA	#obama2012	#msnbc2012	#Boehner	#FourMoreYears

TABLE 4
Hashtags with highest conservative political leaning per event.

Event	Top five conservative hashtags				
1	#LNYHBT	#lnyhbt	#twisters	#sgp	#ocra
2	#LNYHBT	#ObamaTax	#Obamatax	#ocra	#lnyhbt
3	#LNYHBT	#sgp	#TCOT	#Mitt2012	#ocra
4	#LNYHBT	#twisters	#ocra	#sgp	#TCOT
5	#Mitt2012	#LNYHBT	#ocra	#twisters	#military
6	#ObamaIsntWorking	#WAR	#Resist44	#2016	#MSM
7	#EmptyChair	#ForwardNotBarack	#sgp	#resist44	#TCOT
8	#BenghaziGate	#sgp	#CNBC2012	#BidenUnhinged	#lnyhbt
9	#CantAfford4more	#BenghaziGate	#notoptimal	#sgp	#Missouri
10	#BenghaziGate	#Benghazigate	#sgp	#caring	#MyFirstTime
11	#military	#7HoursOfHell	#Twibbon	#LNYHBT	#BENGHAZI
12	#sgp	#ocra	#Benghazi	#WAR	#lnyhbt

of liberals per day.

The results are shown in Figure 15. First, liberals outnumber conservatives in every single day, regardless of whether it is an event day or not. Second, there is a slight increasing trend in polarization over the course of events, which corresponds to discussions become more heated as the election campaign progresses. Third, the fraction of liberals is significantly higher at the onset (first day) of an event. This is observed in 10 out of the 12 events. In contrast, polarization drops and reaches a local minimum during an event (10/12 events), and the level is lower than nearby non-event days. It appears the influx of users during an event drives polarization of the Twitter population down, because these extra users tend to have weaker political leaning.

We also contrast the changes in liberal skew and polarization right before and after the election outcome came out: at the climax of the election, the liberal-conservative share in active users is relatively balanced because both sides want to promote their candidate of support; at the same time the population is more polarized. After Obama is projected to win, there is a jump in liberal fraction with conservatives leaving the discussion, and polarization plummets, probably with the departure of strong-leaning conservatives.

6.4 Sensitivity to Parameter Variation

Our algorithm is robust to variation in input parameters λ, α_C and α_L. Starting from the chosen $(\lambda, \alpha_C, \alpha_L)$ tuple from the previous section, we fix two parameters and vary the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering

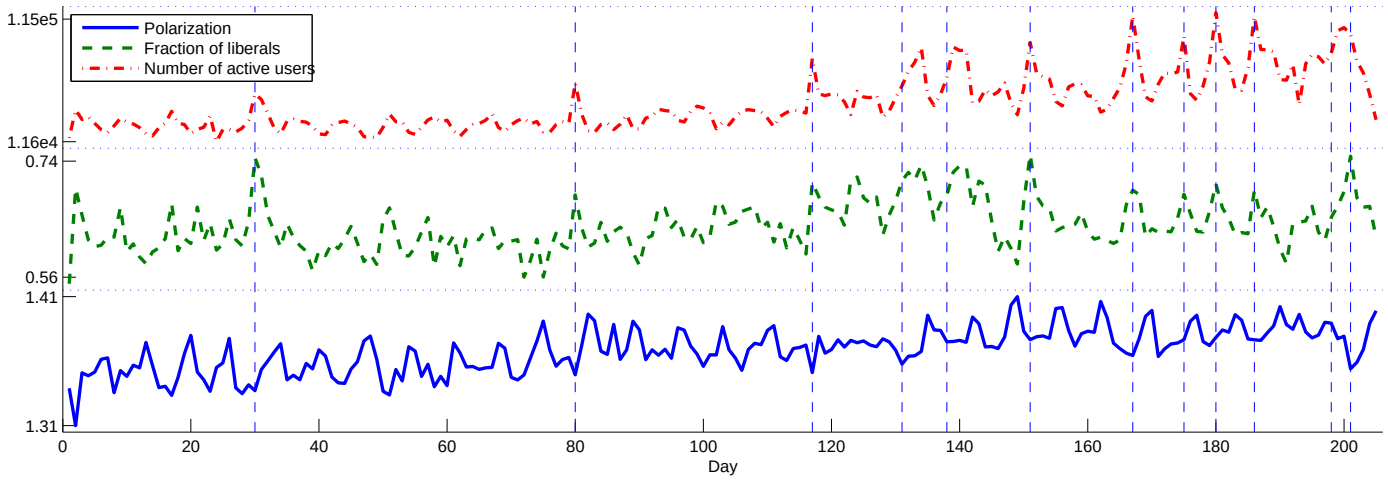


Fig. 15. Skew measures of ordinary users across time. Days with significant data loss during data collection are omitted. Liberal fraction and polarization are negatively correlated during events.

remaining one. Figures 16(a) and 16(c) show the resultant performance does not vary significantly over a wide range of parameter values.

We also consider a simple approach to sparsify the graph Laplacian. Given $\mathbf{S}$, we preprocess it with a k-nearest-neighbor step before passing it to matrix scaling: for each source i we find the k other sources $\{j\}$ with highest S_{ij}, then we keep only the S_{ij} entries (not set to zero) for j being a neighbor of i or i being a neighbor of j. We vary the value of k from 50 to 1000 to vary the sparsity of $\mathbf{S}$ (and $\mathbf{L}$). Figure 16(b) shows even when k is small ($k = 50$ results in a sparsity of 8.7%), the performance is still very good.

6.5 Scalability

The runtime of our method depends on the following steps: (a) computing $\mathbf{A}$ and $\mathbf{S}$ from raw tweets, (b) computing $\mathbf{W}$ from $\mathbf{S}$ through matrix scaling, and (c) solving (10) given $\mathbf{A}$ and $\mathbf{W}$. Note that steps (a) and (b) are readily scalable by parallelization (see, e.g., [52]), so our focus is on step (c). Problem (10) is a convex optimization problem and can be solved efficiently. Even when no structure is exploited, its solution can be computed by direct methods with $O(N^3)$ (assuming $N > E$) flops. In general, there are more efficient iterative methods to solve (10), e.g., interior-point algorithms [8]. For memory requirements, our method requires $O(N^2)$ memory as the matrices in problem (10) have size at most $N \times N$.

Using a desktop computer with modest hardware (one i7-4790K processor, 32GB memory), step (a) requires less than 4 hours to run on our full dataset of size 405GB, and steps (b) and (c) require 5 seconds to complete for $N = 1000$ with a naive implementation in Matlab and CVX [14]. Figure 17 shows our implementaton is able to scale to reasonably large problem sizes at rate lower than the theoretical $O(N^3)$.

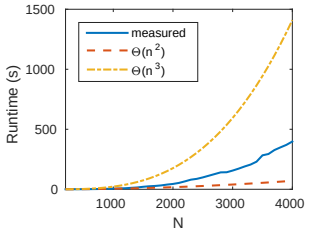


Fig. 17. Runtime of our algorithm with increasing N.

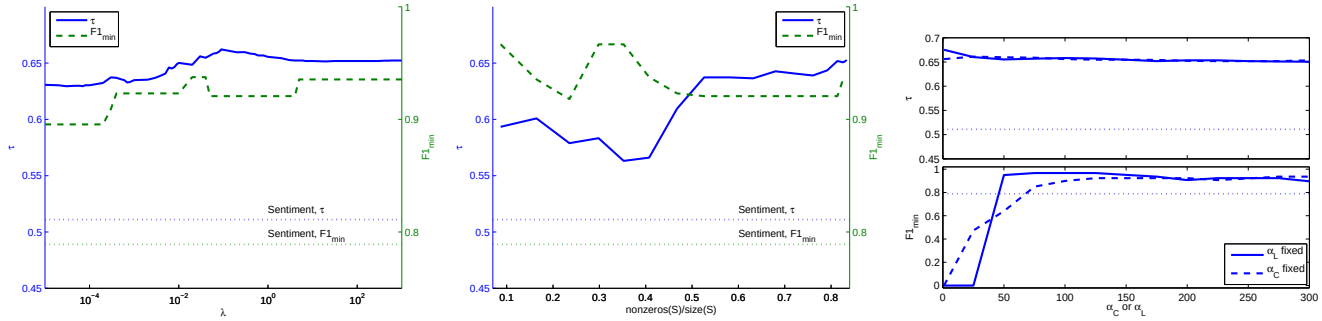
7 Conclusions and Future Work

Scoring individuals by their political leaning is a fundamental research question in computational political science. From roll calls to newspapers, and then to blogs and microblogs, researchers have been exploring ways to use bigger and bigger data for political leaning inference. But new challenges arise in how one can exploit the structure of the data, because bigger often means noisier and sparser.

In this paper, we assume: (a) Twitter users tend to tweet and retweet consistently, and (b) similar Twitter users tend to be retweeted by similar sets of audience, to develop a convex optimization-based political leaning inference technique that is simple, efficient and intuitive. Our method is evaluated on a large dataset of 119 million U.S. election-related tweets collected over seven months, and using manually constructed ground truth labels, we found it to outperform many baseline algorithms. With its reliability validated, we applied it to quantify a set of prominent retweet sources, and then propagated their political leaning to a larger set of ordinary Twitter users and hashtags. The temporal dynamics of political leaning and polarization were also studied.

We believe this is the first systematic step in this type of approaches in quantifying Twitter users' behavior. The Retweet matrix and retweet average scores can be used to develop new models and algorithms to analyze more complex tweet-and-retweet features. Our optimization framework can readily be adapted to incorporate other types of information. The $\mathbf{y}$ vector does not need to be computed from sentiment analysis of tweets, but can be built from exogenous information (e.g., poll results) to match the opinions of the retweet population. Similarly, the $\mathbf{A}$ matrix, currently built with each row corresponding to one event, can be made to correspond to other groupings of tweets, such as by economic or diplomatic issues. The $\mathbf{W}$ matrix can be constructed from other types of network data or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2553667, IEEE Transactions on Knowledge and Data Engineering



(a) Performance w.r.t. overall regularization strength.

(b) Performance w.r.t. sparsity of graph Laplacian.

(c) Performance w.r.t. anchors' regularization strength.

Fig. 16. Our inference technique is robust with respect to parameter variations.

similarity measures. Our methodology is also applicable to other OSNs with retweet-like endorsement mechanisms, such as Facebook and YouTube with "like" functionality.

8 Acknowledgments

This work was in part supported by ARO W911NF-11-1-0036, NSF NetSE CNS-0905086, and the Research Grants Council of Hong Kong RGC 11207615 and M-CityU107/13. We also thank Prof. John C.S. Lui for fruitful discussions.

References

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proc. LinkKDD*, 2005.

[2] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors," in *Proc. ICWSM*, 2012.

[3] J. An, M. Cha, K. P. Gummadi, J. Crowcroft, and D. Quercia, "Visualizing media bias through Twitter," in *Proc. ICWSM SocMedNews Workshop*, 2012.

[4] S. Ansolabehere, R. Lessem, and J. M. Snyder, "The orientation of newspaper endorsements in U.S. elections," *Quarterly Journal of Political Science*, vol. 1, no. 4, pp. 393–404, 2006.

[5] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," *Political Analysis*, 2014.

[6] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? I know who you supported in the UK 2010 general election," in *Proc. ICWSM*, 2012.

[7] d. boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. HICSS*, 2010.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. ICWSM*, 2010.

[10] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 2, pp. 355–370, 2004.

[11] R. Cohen and D. Ruths, "Classifying political orientation on Twitter: It's not easy!" in *Proc. ICWSM*, 2013.

[12] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in *Proc. IEEE SocialCom*, 2011.

[13] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, "Political polarization on Twitter," in *Proc. ICWSM*, 2011.

[14] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2012.

[15] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Czechoslovak Mathematical Journal*, vol. 25, no. 4, pp. 619–633, 1975.

[16] S. Finn, E. Mustafaraj, and P. T. Metaxas, "The co-retweeted network and its applications for measuring the perceived political polarization," in *Proc. WEBIST*, 2014.

[17] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[18] M. Gabielkov, A. Rao, and A. Legout, "Studying social networks at scale: Macroscopic anatomy of the Twitter social graph," in *Proc. SIGMETRICS*, 2014.

[19] D. Gayo-Avello, "All liaisons are dangerous when all your friends are known to us," in *Proc. HT*, 2011.

[20] M. Gentzkow and J. M. Shapiro, "What drives media slant? Evidence from U.S. daily newspapers," *Econometrica*, vol. 78, no. 1, pp. 35–71, January 2010.

[21] S. Gerrish and D. Blei, "How the vote: Issue-adjusted models of legislative behavior," in *Proc. NIPS*, 2012.

[22] ——, "Predicting legislative roll calls from text," in *Proc. ICML*, 2011.

[23] J. Golbeck and D. Hansen, "A method for computing political preference among Twitter followers," *Social Networks*, vol. 36, pp. 177–184, 2014.

[24] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, 2013.

[25] T. Groseclose and J. Milyo, "A measure of media bias," *The Quarterly Journal of Economics*, vol. 120, no. 4, pp. 1191–1237, November 2005.

[26] D. E. Ho and K. M. Quinn, "Measuring explicit political positions of media," *Quarterly Journal of Political Science*, vol. 3, no. 4, pp. 353–377, 2008.

[27] A. S. King, F. J. Orlando, and D. B. Sparks, "Ideological extremity and primary success: A social network approach," in *Proc. MPSA Conference*, 2011.

[28] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

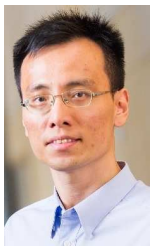
[29] M. Laver, K. Benoit, and J. Garry, "Extracting policy positions from political texts using words as data," *American Political Science Review*, vol. 97, no. 2, 2003.

[30] Y.-R. Lin, J. P. Bagrow, and D. Lazer, "More voices than ever? quantifying media bias in networks," in *Proc. ICWSM*, 2011.

[31] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The party is over here: Structure and content in the 2010

election," in *Proc. ICWSM*, 2011.
[32] J. R. Lott and K. A. Hassett, "Is newspaper coverage of economic events politically biased?" Online, October 2004, http://dx.doi.org/10.2139/ssrn.588453.
[33] C. Lumezanu, N. Feamster, and H. Klein, "#bias: Measuring the tweeting behavior of propagandists," in *Proc. ICWSM*, 2012.
[34] Y. Mejova, P. Srinivasan, and B. Boynton, "GOP primary season on twitter: "Popular" political sentiment in social media," in *Proc. WSDM*, 2013.
[35] P. T. Metaxas and E. Mustafaraj, "Social media and the elections," *Science*, vol. 338, pp. 472–473, 2012.
[36] S. A. Munson, S. Y. Lee, and P. Resnick, "Encouraging reading of diverse political viewpoints with a browser widget," in *Proc. ICWSM*, 2013.
[37] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opinions of the long tail," in *Proc. SocialCom/PASSAT*, 2011.
[38] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, 2006.
[39] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song, "The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns," in *Proc. CSCW*, 2011.
[40] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to Twitter user classification," in *Proc. ICWSM*, 2011.
[41] K. T. Poole and H. Rosenthal, "A spatial model for legislative roll call analysis," *American Journal of Political Science*, vol. 29, no. 2, pp. 357–384, 1985.
[42] 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, 2007.
[43] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *Proc. SMUC*, 2010.
[44] E. Seneta, *Non-negative Matrices and Markov Chains*. Springer, 2007.
[45] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and K. A., "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.
[46] 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. ICWSM*, 2010.
[47] S. Volkova, G. Coppersmith, and B. Van Durme, "Inferring user political preferences from streaming communications," in *Proc. ACL*, 2014.
[48] J. Wang, W. X. Zhao, Y. He, and X. Li, "Infer user interests via link structure regularization," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 2, 2014.
[49] I. Weber, V. R. K. Garimella, and E. Borra, "Mining web query logs to analyze political issues," in *Proc. WebSci*, 2012.
[50] I. Weber, V. R. K. Garimella, and A. Teka, "Political hashtag trends," in *Proc. ECIR*, 2013.
[51] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang, "Quantifying political leaning from tweets and retweets," in *Proc. ICWSM*, 2013.
[52] R. B. Zadeh and A. Goel, "Dimension independent similarity computation," *Journal of Machine Learning Research*, vol. 14, no. 1, 2013.
[53] D. X. Zhou, P. Resnick, and Q. Mei, "Classifying the political leaning of news articles and users from user votes," in *Proc. ICWSM*, 2011.
[54] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnigie Mellon University, Tech. Rep., 2002.

Felix Ming Fai Wong (S'10-M'15) received the B.Eng. in computer engineering from the Chinese University of Hong Kong, Hong Kong, in 2007, the M.Sc. in computer science from the University of Toronto, Toronto, ON, Canada, in 2009, and the Ph.D. in electrical engineering from Princeton University, Princeton, NJ, USA, in 2015. He is currently a Software Engineer in search and data mining with Yelp, Inc., San Francisco, CA, USA. Dr. Wong was a winner of the Yelp Dataset Challenge, and a recipient of the Best Paper Award Runner-up at IEEE INFOCOM 2015.



Chee Wei Tan (M'08-SM'12) received the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 2006 and 2008, respectively. He is an Associate Professor with the City University of Hong Kong. He was a Postdoctoral Scholar at the California Institute of Technology (Caltech), Pasadena, CA. He was a Visiting Faculty with Qualcomm R&D, San Diego, CA, USA, in 2011. His research interests are in networks, inference in online large data analytics, and optimization theory and its applications. Dr. Tan was the recipient of the 2008 Princeton University Wu Prize for Excellence. He was the Chair of the IEEE Information Theory Society Hong Kong Chapter in 2014 and 2015. He was twice selected to participate at the U.S. National Academy of Engineering China-America Frontiers of Engineering Symposium in 2013 and 2015. He currently serves as an Editor for the IEEE Transactions on Communications and the IEEE/ACM Transactions on Networking.

Soumya Sen is an Assistant Professor of Information & Decision Sciences at the Carlson School of Management of the University of Minnesota. He received his MS and PhD from the University of Pennsylvania in 2008 and 2011 respectively, and did his postdoctoral research at the Princeton University during 2011-2013. His research takes an interdisciplinary approach involving computer networks, economics, and human-computer interaction. His research interests include network economics, resource allocation, incentive mechanisms and e-commerce. Dr. Sen was awarded the Best Paper Award at IEEE INFOCOM in 2012 and INFORMS ISS Design Science Award in 2014. He has co-edited a book titled "Smart Data Pricing", published by Wiley in August 2014.

Mung Chiang (S'00, M'03, SM'08, F'12) is the Arthur LeGrand Doty Professor of Electrical Engineering at Princeton University and the recipient of the 2013 Alan T. Waterman Award. He created the Princeton EDGE Lab in 2009 to bridge the theory-practice divide in networking by spanning from proofs to prototypes, resulting in a few technology transfers to industry, several startup companies and the 2012 IEEE Kiyo Tomiyasu Award. He serves as the inaugural Chairman of Princeton Entrepreneurship Council and the Director of Keller Center for Innovation in Engineering Education at Princeton. His Massive Open Online Courses on networking reached over 250,000 students since 2012 and the textbook received the 2013 Terman Award from American Society of Engineering Education. He was named a Guggenheim Fellow in 2014.