Voices of Victory: A Computational Focus Group Framework for Tracking Opinion Shift in Real Time

Yu-Ru Lin[1,2] Drew Margolin[2] Brian Keegan[1] David Lazer[1,2]
[1]College of Social Sciences and Humanities, Northeastern University, Boston, MA 02115, USA
[2]College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA
{yu-ru.lin,d.margolin,b.keegan,d.lazer}@neu.edu

ABSTRACT

Social media have been employed to assess public opinions on events, markets, and policies. Most current work focuses on either developing aggregated measures or opinion extraction methods like sentiment analysis. These approaches suffer from unpredictable turnover in the participants and the information they react to, making it difficult to distinguish meaningful shifts from those that follow from known information. We propose a novel approach to tame these sources of uncertainty through the introduction of "computational focus groups" to track opinion shifts in social media streams. Our approach uses prior user behaviors to detect users' biases, then groups users with similar biases together. We track the behavior streams from these like-minded subgroups and present time-dependent collective measures of their opinions. These measures control for the response rate and base attitudes of the users, making shifts in opinion both easier to detect and easier to interpret. We test the effectiveness of our system by tracking groups' Twitter responses to a common stimulus set: the 2012 U.S. presidential election debates. While our groups' behavior is consistent with their biases, there are numerous moments and topics on which they behave "out of character," suggesting precise targets for follow-up inquiry. We also demonstrate that tracking elite users with well-established biases does not yield such insights, as they are insensitive to the stimulus and simply reproduce expected patterns. The effectiveness of our system suggests a new direction both for researchers and data-driven journalists interested in identifying opinion shifting processes in real-time.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications-Data mining; H.3.5 [**Information Storage and Retrieval**]: Online Information Services; H.5.4 [**Information Interfaces and Representation**]: Hypertext/Hypermedia

General Terms

Measurement; Experimentation; Human Factors

Keywords

real time system; social meter; public opinion; data-driven journalism; process inference; computational social science

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

1. INTRODUCTION

Tweets provide real-time representations of millions of individuals' thoughts and feelings. This data is an unprecedented resource for social scientists, politicians, marketers, and journalists to understand the behavior and propensities of large groups of people. Tweet streams integrate the representative scope of polls and surveys with the free-form responses of focus groups and interviews. An additional benefit of these streams is that Twitter users are embedded within their usual social contexts rather than artificial contexts created by polls, focus groups, and other survey methods. The scope and sensitivity of Twitter has thus become an attractive means of measuring and assessing the responses of the public to events and information [5, 17].

The methods for studying tweet streams have only begun to scratch the surface of this potential. For example, there is often broad interest in how a new event, be it the release of a new product, a report of violent conflict in a foreign land, or changes in the nation's economic climate will affect the views and behavior of the public [3]. There is also increasing interest amongst those in the field of "data-driven journalism" to harness social media like Twitter to enhance the understanding of these public responses [7, 16, 17]. Yet analysis in these arenas tends to draw on only a limited portion of the information Twitter has to offer.

The most common method for analyzing tweet streams is to naïvely combine large volumes of tweets into aggregate measures such as sentiment or counts [3, 29, 24]. Though this method has the benefit of simplicity in implementation, it creates difficulties when it comes to interpretation. Since tweets represent a biased subset of a larger population, additional real-time data, such as user adoption rates and tweet rates, must be gathered to support inferences from tweet streams to larger populations of interest [12]. At the same time, expressed attitudes have complex relationships with beliefs and behaviors [26, 9, 15]. By combining all responses that fit a simple criterion, without reference to the history or context of the individuals that produced them, these aggregations are ambiguous and—in some cases—misleading.

In this paper we describe and implement a system for computing aggregates for which interpretation is much more clear. Specifically, we argue that typical aggregations of social media streams run into difficulties because they attempt to make *population inferences*—measures of the distribution of states within a population—while the optimal use of social media is to make *process inferences* [14, 11]—detection of the mechanisms that transform these states in response to stimuli. Further, while there are already many good meth-

ods for inferring a population's opinions or attitudes that do not rely on social media, many processes that produce real-time shifts in opinion are only now observable through social media streams.

For example, when Apple releases a new iPhone, a consumer survey is likely to do a better job estimating the population segments that view the phone favorably or intend to purchase it than an aggregation of tweets. However, our approach would suggest that with appropriate aggregation, Twitter streams could do a better job of revealing consumers' reasons for their purchases or non-purchases. Such reasons would not only provide information about how to redesign or improve the product, they may also provide more information for predicting the long term shifts in the adoption rate of the population [13]. Process inferences are also useful for gauging how the public reacts to breaking news and events. Surveys and focus groups take time to develop, and thus can have difficulty capturing individuals' reactions to news in the moment, before they have been exposed to the interpretations of the media or other authorities [18].

The key to making these kind of process inferences is the aggregation of responses within well-defined sub-populations for which there are clear *a priori* expectations about the likely responses of these sub-populations. We propose a novel framework called *computational focus groups* to identify meaningful shifts against expected baseline behaviors. We demonstrate our approach in the context of tracking real-time responses to the 2012 U.S. presidential debates. As shown in Figure 1, our framework is designed to transform the social media streams into a meaningful process measures through tracking the responses of a reference sub-population for whom partisan biases have already been identified.

The key contributions of this paper include:

- We propose a novel real-time system that tracks user responses to novel content and highlights shifts in their opinions by accounting for their underlying biases.
- We present a comprehensive analysis of user responses during four presidential debates, and demonstrate that the behavior of computational focus groups lead to different conclusions than would be drawn from simple aggregations of Twitter data.
- We show that systematic behavior in computational focus groups leads to plausible yet surprising inferences about the topical and rhetorical sources of opinion shift and political agreements and disagreements.
- Comparisons of our focus groups with groups of media elites provide quantitative evidence for a widely discussed but difficult to measure phenomenon: political polarization and intransigence amongst media elites.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the rationale behind our approach. Section 4 describes the implementation details and the procedure for constructing computational focus groups. Section 5 presents our experimental results, including a comprehensive analysis of group responses to the 2012 U.S. presidential debates. Finally we provide discussion and conclusion in Sections 6 and 7.

2. RELATED WORK

Social networking sites like Twitter and Facebook show profound increases in traffic and information sharing during media and news events suggesting the value of using social media data for assessing public opinions during major events [27]. Twitter data have been mined in real time for temporal cues during sporting events [21] and television shows [6]. These back-channel sources of communication reflect individual and collective attitudes, opinions, preferences and beliefs of large numbers of people around a focused set of topics. In particular, the massive outpouring of political communication on social media sites permits analysis of sentiment and topics during political events such as elections [3, 29, 24] and debates [8, 19].

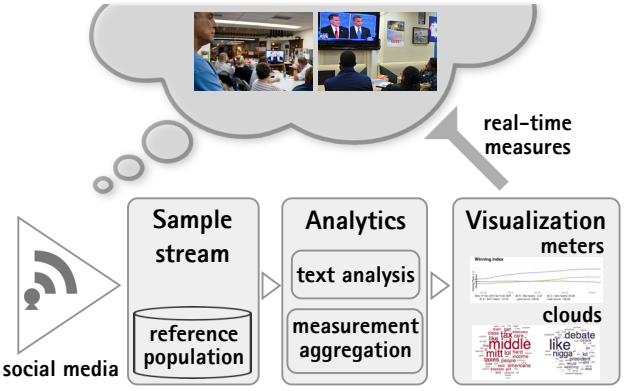

Figure 1: System overview. Our framework is designed to capture data from Twitter's Streaming API, analyze the content of these tweets, aggregate and normalize this output, and visualize it on an interactive web page.

Prior analysis of sentiment on Twitter has found correlations between the extracted emotive trends and major events such as 2008 U.S. presidential election [3], consumer confidence, and political opinion [24]. The mere number of tweets mentioning a party reflects the results of the German federal election [29]. While interesting events such as elections can be detected by anomalies in the pulse of the sentiment signal and controversial topics can be identified by correlated sentiment responses, there are also difficulties in using the responses of Twitter users to infer the responses of the voting population as a whole [8].

One way to address this limitation is to track responses for different categories of users, much as polls track opinions for different categories of respondents. Research efforts have been made on automatically classifying different types of users on Twitter [4, 25]. Machine learning algorithms can identify users with different characteristics including political leaning, ethnicity, and affinity for a particular business [25] as well as users' political leaning is based on number of tweets referring to a particular political party [4]. Although it would be useful to classify users with various interests, the best classification for process inferences are those that are most likely to reveal changes or disturbances brought on by the event or stimuli of interest, rather than the most typical examples of an established category.

The immediacy of social media users responding to an event has also increased interest in harnessing social media for data-driven journalism [7, 16, 17]. Methods for filtering and detecting potentially interesting information sources [7] as well as visualizing sentiment around an event and identifying sub-events [17] are complementary to our proposed framework as most of this work focuses efficiently on extracting credible information that journalists might otherwise discover through painstaking effort. Our approach thus

emphasizes information that would be difficult to discover through alternative means.

As we discuss in the next section, the key distinction in this work is the focus of detecting *processes* rather than detecting the population's *states*. This work employs a lexicon-based sentiment analysis [22], although other approaches such as WordNet Affect [28] and SentiWordNet [10] are also applicable.

3. RESEARCH DESIGN: COMPUTATIONAL FOCUS GROUP

This section describes the rationale behind the design of our framework.

3.1 Challenge: Population Inference vs. Process Inference

The size of the data in tweet streams suggests that observed patterns are likely to be reliable and generalizable. The soundness of an inference depends, however, on what is being generalized to. Social scientists often distinguish between *population inferences* and *process inferences* [14]. A population inference is one in which data about a sample are used to make inferences about the qualities of a larger population, *e.g.* a political poll in which the opinions of a sample of voters are used to infer the opinions of the voting population at large. In a process inference, a particular hypothesis about how individuals behave under particular conditions is considered and evidence for or against this hypothesis is sought, *e.g.*, in an experiment in which participants are given different treatments and their responses are compared.

Both approaches have their limitations. Population inferences are limited by the researcher's ability to obtain an appropriate sample [14]. A biased sample will lead to biased results in estimating the underlying population. Process inferences are less threatened by problematic sampling, so long as sampling biases do not correlate with the process hypotheses being tested. Reliable knowledge is built around process inferences through repeated studies and replications of the effect in different contexts [11].

3.2 Design Goal: Twitter as Process Inference Tool

We suggest that Twitter is a relatively poor tool for population inference but a potentially useful tool for process inference, for the following reasons. First, in many cases the population of interest includes individuals that do not use Twitter as well as those that do. If use of Twitter is correlated with particular political leanings, results will be skewed by these response biases.

Secondly, even if this general bias can be identified and controlled for, tweet streams at any particular moment only show behavior for users who are actively sending messages at a particular point in time [20]. Observed changes in a typical tweet stream filtered by a simple criterion thus always reflect two processes: changes in the expressions of a set of Twitter users *or* changes in the set of Twitter users who chose to tweet. Treating this distinct processes as the same will produce inconsistent results. For example, increases in tweet volumes mentioning Barack Obama during the 2008 election campaign correlated strongly in time with moves toward Obama in the polls [24]. However, tweets for Obama's opponent, John McCain, correlated with moves away from McCain in these polls. This effect might easily be explained by reference what is known as the *spiral of silence*, which states that people are more likely to express opinions when they believe others share their views [23]. If Twitter users tend to favor Obama in general, and the spiral of silence is in operation, then McCain will be mentioned more as Obama's popularity rises simply because the disproportionately large set of pro-Obama Twitter users will discuss politics when their favored candidate is doing well.

The ambiguity caused by these entries and exits from the tweet stream limits the potential of Twitter as a population inference tool. With a heterogeneous and potentially shifting user base, it is difficult to to say whether a stream of tweets "represents" a population as a whole. However, if a set of users with relevant characteristics can be identified, the information yielded by these entries and exits can be used to make process inferences by pointing to the kinds of things to which individuals respond and how they do so.

To construct such a system, we attempt to aggregate tweets into groups with known, relevant and interpretable baseline attributes. Given these attributes, movements in tweet streams that reflect deviations from what we would expect are then considered meaningful. To make the system broadly applicable, we sought to identify attributes through data that would be available to researchers in other contexts. Thus, our system identifies *computational focus groups* – groups for which expectations of behavior can be established through observation of prior behavior on Twitter.

We demonstrate our approach using tweets related to the 2012 U.S. presidential debates. Our rationale relies on the debates as a common "treatment" applied to a large number of individuals. In most circumstances, individuals may express different political opinions or discuss different facts and information because they are limited in their exposure [2]. During the debates this condition is temporarily reduced, as most people who are interested in politics are observing the same literal content. This means that differences in their responses can be attributed to differences in their biases or social contexts and circumstances.

By obtaining information on these biases, we can see when they are activated or around which topics. Thus, our goal was to identify individuals who are:

- likely to be watching the debates;
- likely to tweet while watching the debates;
- likely to hold a prior political bias (either Democrat or Republican);
- somewhat likely to deviate from their political bias at any particular moment.

We assign individuals who are likely to share biases into groups. Then during the debates we observe *when* and *in reference to what* are associated with changes within and between the groups' response tendencies. These changes likely reflect processes operating within the group in response to exogenous events. We also compare the aggregate responses of our computational focus groups to groups created through more traditional means such as lists of media elites like journalists and pundits with established political orientations.

3.3 Identifying Groups: Selective Exposure

To meet the above criteria we relied on the theory of selective exposure [2], a well-supported finding in social science which posits that individuals seek information sources that corroborate their existing political views. If the theory is

correct, then an individual's prior information consumption activity can be used to detect both their political bias as well as their interest in political events. Individuals who appear to consume political information frequently and with a strong bias toward a particular partisan view are thus good candidates for our groups.

Identifying groups from prior events. The Democratic National Convention (DNC) and Republican National Convention (RNC) are effectively extended advertisements for each respective party and its candidates. Thus they represent events in which political content has a known and reliable bias. We sampled individuals whose Twitter behavior reveals their exposure to this biased content. If an individual tweets about a convention speaker during the speaker's convention speech, we assume that the individual was watching and responding to this speech.

Based on the logic of selective exposure, we then divide these active watchers into three groups. Individuals that tweet heavily during one party's convention but not the other's are classified as biased toward that party. Individuals that tweet heavily during both conventions are classified as neutral.

As a basis for comparison with our computational focus groups, we also extract groups of "Elite" users for whom bias need not be detected computationally. These users include well-known commentators, columnists, or other professionals with very strong partisan biases. These strong biases serve as a basis for comparison for our focus groups. We are also interested in the extent to which they will deviate in their partisan views when exposed to novel political information expressed during the debates.

4. IMPLEMENTATION

4.1 System Overview

We develop a real-time system to support the research goals based on our design rationale. As shown in Figure 1, our framework consists of sample stream, analytic and visualization components that transform the social media streams into a meaningful process measures.

Sample stream component. This gathers social online social media stream data via Twitter's Streaming API[1]. We use the "follow" parameter to specify the list of users to return statuses (tweets) for in the stream. We called the list a "reference population" which is identified based on the selective exposure methodology described previously.

Analytic component. This component first parses and analyzes the content of tweets via parallel asynchronous threads running on a cluster server. The content analysis will be detailed in a later section. A set of aggregators periodically aggregate[2] the analyzed results to create a time series of various measures, including tweet volume, sentiment, and tag clouds, aggregated by each group. Finally, the time series data are sent to the web server.

Visualization component. This component visualizes the time series data via an interactive web interface implemented in d3js, Rickshaw, and d3-cloud.[3] The demo system was live and available during the debate nights. As the debates were over, we have provided an animated debate recap for visitors to review our analysis[4].

4.2 Constructing Focus Groups

Selective exposure theory allowed us to identify sub-populations of Twitter users with likely political biases. We assemble a list of the authors of all tweets which mention a convention speaker (such as "Obama" or "Romney") during the speaker's convention speech or are sent to one of the official convention hashtags ("#DNC2012"; "#RNC2012"). Any user that sent at least three tweets meeting these criteria for the duration of the convention showed a propensity both to watch the convention and to tweet actively during it. We label these users "active watcher" and examine their tweeting activity during the opposing party's convention. If a Twitter user did not send any tweets mentioning a speaker or hashtag from the opposing convention, we infer this individual was unlikely to have watched the opposing convention. This procedure yielded sub-populations of substantially different sizes, which raises concerns about the representativeness of these populations. These sub-populations of both, RNC, and DNC watchers were randomly re-sampled to produce groups of equal sizes containing 2,500 members.

- Active "Both" Watcher ("B_w") – users identified as "active watchers" of both conventions
- Partisan DNC Watcher ("D_w")– users identified as "active watchers" of the DNC with no evidence of watching the RNC
- Partisan RNC Watcher ("R_w") – users identified as "active watchers" of the RNC with no evidence of watching the DNC

To understand these groups, we ask hand coders to independently review a random sample of 6 tweets issued during the debates from 700 randomly selected users in our groups, as well as the user-profile pages of these users. The inter-coder reliability for these judgments was 95%. Figure 2 shows the extent to which the group extraction method mismatched the hand coded interpretations. We report mismatches where the detection method assigned an individual to a group and the hand-coder assigned them to an opposing group. The low inconsistency rate suggest that these groups do possess leaning expected by the selective exposure theory.

4.3 Constructing Elite Groups

Political elites such as politicians, pundits, journalists who have many followers potentially have different motivations and behaviors when tweeting during the debates. We identify these elite users by extracting the most followed users listed in a Twitter user dictionary website *wefollow.com* where users are tagged with their ideological leanings. Similar to the users in the focus group sub-populations described above, we identify three elite groups for Democrats, Republicans, and users interested in both parties like journalists. The Democratic elites are identified based on tags such as "democrat", "liberal", or "progressive" but not any of the media tags. The Republican elites are based on tags such as "republican", "conservative", or "tea party" but not any

[1] `http://dev.twitter.com/docs/streaming-apis/streams/public`

[2] We use a 10-second update frequency for calculating the debate data.

[3] `http://d3js.org/`; `http://code.shutterstock.com/rickshaw/`; `https://github.com/jasondavies/d3-cloud`

[4] `http://goo.gl/yMoee`

Figure 2: Group identification by coders. We compare the users' group assignment (the D_w, R_w or B_w groups) with their partisan leaning as identified by two human coders. The inconsistency is computed as the ratio of users with leaning given by a coder inconsistent with the watching group assignment (*e.g.*, when the coder gave "leaning-D" to a Republican watcher), which is lower than 30%. The inverse of inconsistency between coders reflects annotation agreement, and the inverse of inconsistency between watcher assignment and annotation suggests accuracy of this identification approach.

of the media tags. The media users are identified based on tags such as "news", "journalist", or "reporter".

The real time responses of these elite groups ("B_e"), Democratic elites ("D_e"), and Republican elites ("R_e") will be compared to the B_w, D_w, and R_w groups.

Table 1 lists the set of tags used for extracting the elite groups and the number of users with the tags. We then randomly select users that meet the above-mentioned criteria into three groups with roughly the same size. Table 2 lists the final number of users in our samples.

leaning	tag	number of users
Democrat	democrat	693
	liberal	959
	progressive	1062
	p2	78
Republican	republican	875
	conservative	2712
	tcot	1730
	teaparty	587
Unknown	politics	2873

Table 1: Relevant tags for elite extraction

	Republican	Democratic	Both
Focus group	2481	2481	2561
Elite	944	964	931

Table 2: Size of each group

4.4 Tracking Behavioral Streams

This section describes the detailed implementation in the analytic component. In this component, we generate two time-dependent measures for comparing the group reactions to the candidates' debate performance: the sentiment index and the "winning" index. Table 3 provides a look-up for reading the two indices. As described below, the indices are scored relative to the incumbent ticket, Obama-Biden.

Sentiment index. The sentiment index tracks how the groups express sentiment when talking about the candidates. Scores above zero on the index indicate members of the group tweet positive words when mentioning Obama-Biden, and negative words when mentioning Romney-Ryan. Scores below zero indicate the reverse: that tweets contain positive words for Romney-Ryan and negative words for Obama-Biden.

Winning index. The winning index tracks the victory declaration words from the groups. Scores above zero on the index indicate members of the group tweet "winning" words when mentioning Obama-Biden, such as "winning" or "victory," or losing words when mentioning Romney-Ryan. Scores below zero indicate the reverse: that tweets contain winning words for Romney-Ryan and losing words for Obama-Biden.

The computation of the two indices can be summarized into four steps:

1. **Text pre-processing**: For each incoming tweet, we first tokenize the content text. We do not use any stemming and lemmatization because the dictionaries we employ (described below) contain the variant forms of a word, which in practice reduces the overhead in the real-time computation. This step generates case-insensitive words (tokens) for each tweet.
2. **Candidate and party extraction**: For each tweet, we check whether it contains either the names or the Twitter handles of the four candidates (@barackobama, @joebiden, @mittromney, @paulryanvp). When a candidate is detected, the candidate's party is assigned to the tweet. Tweets that do not contain any of these candidates, or contain candidates from both parties are ignored when calculating the indices.
3. **Score computation**: We compute a sentiment score for a tweet based on a microblog-specific affect dictionary [22]. The affect dictionary contains a list of affect words and their sentiment valence values (where positive values indicate positive sentiments). Let $\mathcal{D}$ be the affect dictionary and v_w is the valence value for a word w, the sentiment score of a tweet with a set of words $\mathcal{T}$ is computed as: $\sum_{w\in\mathcal{D}\cap\mathcal{T}} v_w/|\{w : w \in \mathcal{D}\cap\mathcal{T}\}|^2$. The winning score of a tweet is computed similarly. We prepare a list of winning words $\mathcal{D}_\mathcal{W}$ and a list of losing words $\mathcal{D}_\mathcal{L}$. The binary winning score is given by $sign(|\{w : w \in \mathcal{D}_\mathcal{W}\cap\mathcal{T}\}| - |\{w : w \in \mathcal{D}_\mathcal{L}\cap\mathcal{T}\}|)$.
4. **Time-dependent group indices**: To capture the temporal change of the groups' responses to the two party candidates, we calculate the group level indices as follows. Let s_{kp} be the sentiment score or a binary winning indicator of a tweet k, and $p \in \{D, R\}$ indicates the tweet mentions one of the party candidates. We use $t_k \in i$ as a short notation for $t_k \in [t_i, t_i + \Delta t)$, which denotes that the posting time t_k of a tweet k falls into the time interval i. At each time interval i, the net score $\bar{s}_i$ of the incumbent party (Democrat) is defined as the difference between parties: $\bar{s}_i = \text{mean}\{s_{kp} : k \in i, p = D\} - \text{mean}\{s_{kp} : k \in i, p = R\}$. In order to capture the trends during the debates, the cumulative score at time t is reported, which is given by $S_t = \sum_{i=t_0}^{t} \bar{s}_i$, where t_0 is the starting time usually reset as one hour prior to a debate.

index	value	interpretation
sentiment	> 0	positive toward Obama-Biden or negative toward Romney-Ryan
	< 0	positive toward Romney-Ryan or negative toward Obama-Biden
winning	> 0	winning words for Obama-Biden or losing words for Romney-Ryan
	< 0	winning words for Romney-Ryan or losing words for Obama-Biden

Table 3: Summary of measures

5. RESULTS

This section describes the results yielded by a set of "treatments" that was issued to all six of our groups (D_w, R_w, B_w; D_e, R_e, B_e). These treatments were the 2012 U.S. presidential debates. They included three presidential debates between Barack Obama and Mitt Romney on October 3, 16, and 22 and one vice presidential debate between Joe Biden and Paul Ryan on October 11.

We then use the responses of the computational focus groups we identified for two tasks. First, we validate our detection mechanism for both activity and partisanship. Second, to the extent to which our detections are valid, we use deviations from expected patterns to identify interesting processes in the interpretation of political content. We focus on the identification of unusual moments which are disguised by aggregate shifts as well as issue terms that show partisan agreement and disagreement.

5.1 Descriptive Results and Group Validation

An assumption of our system as a process inference tool is that the treatments trigger activity in the groups that differs at least somewhat from their typical behavior. Figure 3 shows the volume of tweets per minute emitted by our groups for a six-hour window surrounding the first two debates. The six-hour window begins one hour before the debate and so the debate time is denoted by the period 01:00–02:30 (UTC). Each of the plots shows a similar path for each group – tweet volume rises rapidly and substantially as the debate begins and reaches a peak of 4–10 times the previous rate shortly thereafter. This rate is maintained consistently until approximately 02:30 when the debate ends and then drops off precipitously toward its pre-debate level. These consistent trends provide strong evidence evidence that group members are paying attention to the debates and tweeting in response to them.

The second proposition of the group identification process is that the groups will display a bias consistent with the partisan basis on which they were selected. That is, the D_e and D_w groups will favor Obama-Biden and the R_e and R_w will favor Romney-Ryan.

Figure 4 shows the cumulative sentiment leaning for each group across all four debates. The vertical axis represents leaning toward the incumbent ticket (Obama-Biden). Thus, a rise in the curve indicates that a group is leaning toward Obama-Biden, a decline in the curve indicates that they are leaning toward Romney-Ryan. Panels A-D represent the focus groups. The B group is indicated by the green lines, the D group by the blue lines, and the R group by the red lines. The debate begins at 01:00 in each panel and concludes at 02:30.

Consistent with our expectation, the curves tend to conform with the partisan biases revealed from the convention watching tweets. By the end of each debate, the blue curve is above the red curve, indicating that on the whole the D_w group has been more favorable to Obama-Biden and the R_w group less favorable to the incumbents. The B_w group is in between the two in 3 of the 4 debates (Debate 4 is the only exception).

Table 4 shows the percentage of the minutes in each debate during which the D_w tweets leaned more toward Obama-Biden than Romney-Ryan.

We also calculate the ticket-leaning of the winning indices for each group (see Figure 5). Once again in 3 out of 4 debates the winning index for the D_w and D_e groups leans more toward Obama-Biden than the winning index for the R_e and R_w groups. The one exception is the final debate, in which the two scores are virtually identical. This effect is consistent even though the overall scores swing between the tickets across debates. For example, both the D_w and the R_w group agree that Romney won the first presidential debate (Figure 5(a)), while they agree that Obama won the second presidential debate (Figure 5(c)).

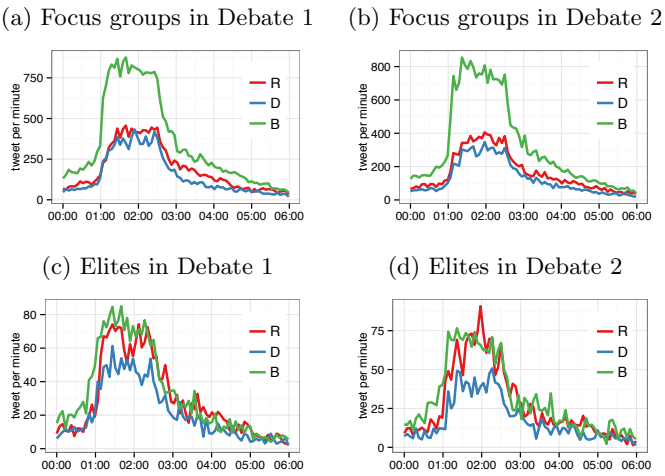

Figure 3: Tweet volume per minute in the first presidential debate and the second (VP) debate. (a,b) Focus groups. (c,d) Elite groups. The x-axis shows the time in UTC, and the y-axis shows the number of tweets per minute. The six groups have similar volume patterns in other presidential debates.

5.2 Comparison to "Elite" Groups

This section compares the results of the focus groups identified using our method (D_w, R_w, B_w) with the elite groups (D_e, R_e, B_e) identified by number of followers. We anticipated that elites may be too rigid in their responses to novel information.

Figure 4(e-h) and Figure 5(e-h) demonstrate this pattern. First, it can be seen that in each panel, the cumulative senti-

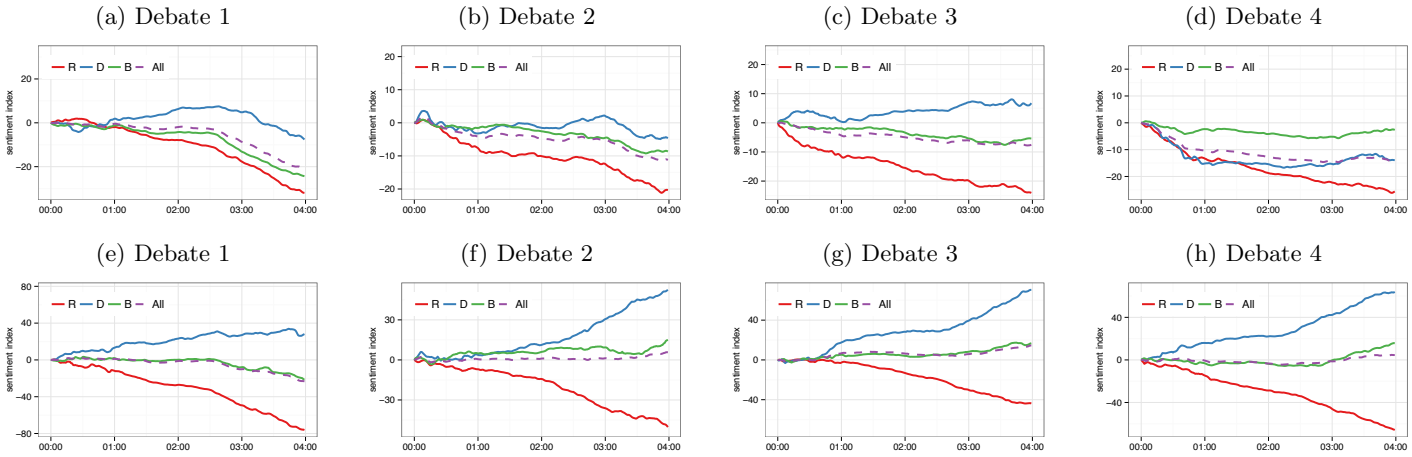

Figure 4: Cumulative sentiment index in the four presidential debates. (a,b,c,d) Focus groups. (e,f,g,h) Elite groups. In each panel, the debate started at 01:00 and concluded at 02:30 UTC. The four debates were held on October 3, 11 (VP), 16, and 22.

	Focus group		Elite	
	prime	moderate	prime	moderate
1	73.3%	96.6%	70.0%	90.0%
2	73.3%	93.3%	66.6%	93.3%
3	70.0%	83.3%	73.3%	90.0%
4	70.0%	83.3%	70.0%	86.6%

Table 4: Group verification via sentiment index: "prime" measures the percentage of time during the debates where D's sentiment index is higher than R's, and "moderate" measures the percentage of time where B's sentiment index is higher than R's or lower than D's.

ment scores and winning indices for D_e and R_e *always* favor the ticket of their preferred candidate. This is the case even when all of the other groups, including the focus groups sharing their biases, appear to favor the opposition candidate. For example, Figure 5 shows there is clear agreement amongst all focus groups that Romney won the first debate. R_e also share this view, as do B_e . Only one group breaks from this pattern – D_e, who tweet as though Obama was the winner. The pattern is the same, with the parties reversed, in the second presidential debate. Figure 5(c) shows clear agreement amongst all focus groups that Obama won the second debate, a view corroborated by the D_e and B_e groups. In this case, only the R_e bucks the trend, tweeting as though Romney was the victor.

A comparison of Figure 5(e) and (g) shows that the elites do respond to the debate content to some degree, as there is clearly a shift toward Obama-Biden across all groups. This shift is much less pronounced than amongst the focus group (Figure 5(a) and (c)), suggesting that though the partisan bias is present, the intractability of the commitment to this bias is weaker, making the watchers in the focus groups for which bias was inferred more sensitive for inferring meaningful processes.

The uniformity, and thus lack of informativeness, in the elite groups' tweets is demonstrated in the sentiment analysis as well. The 4 panels of Figure 4(e-h) showing the elite groups' tweets for each debate are virtually indistinguishable for D_e and R_e groups. That is, in terms of sentiment, these groups may have well simply re-issued their tweets from the previous debate during each new debate, almost completely ignoring the "treatment" of the debates.

5.3 Post-Event Analysis

This section uses the patterns detected by the groups to suggest key moments and topics in the debates worthy of further exploration.

5.3.1 Identifying windows of dominant sentiment

To say that the aggregate sentiment of a population is moving in favor of a candidate does not distinguish whether this is due to supporters becoming louder and more invigorated or opponents softening or perhaps shifting their view. Our method of tracking groups with prior biases but flexible response patterns permits us to detect at which points in time these different processes have likely occurred.

Figure 6 shows the time series of each group's sentiment-leaning fo the first half hour of each debate. As in the other figures, a higher sentiment score indicates sentiment leaning toward Obama-Biden and away from Romney-Ryan. The dashed lines represent each individual group using the same color scheme as in the previous figures. The solid purple line represents the combination of all of the groups in the panel and thus indicates the information that would be conveyed by a typical, aggregate sentiment analysis.

The purple aggregate line indicates that sentiment tended to shift back and forth between the candidates in the early portions of each debate. Analysis of the group responses suggests a somewhat different story, however. The gray shaded regions in each panel indicate time periods for which the purple curve most strongly correlates with the D group's responses. On this dimension, the debates look quite different both within and across debates.

For example, the second 15 minutes, from 01:15 to 01:30 UTC, of Debate 2 (the vice presidential debate) show that the aggregate sentiment was consistently correlated with the responses of the D_w group (most of the region is in gray). This could be either because, during the period, the (presumably neutrual) B_w group tweeted similarly the

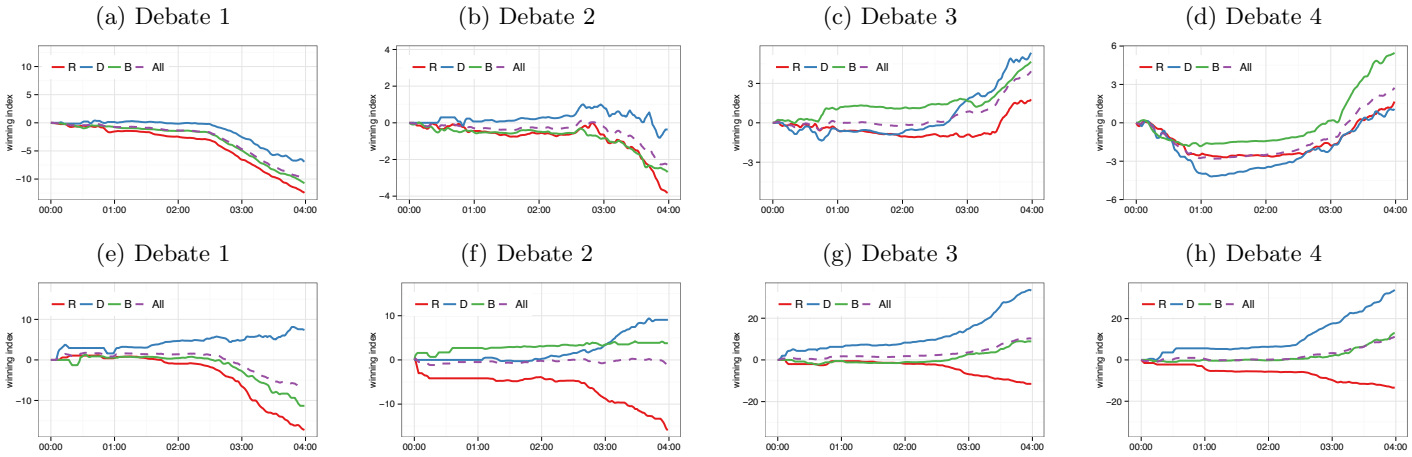

Figure 5: Cumulative winning index in the four presidential debates. (a,b,c,d) Focus groups. (e,f,g,h) Elite groups. In each panel, the debate started at 01:00 and concluded at 02:30 UTC. The four debates were held on October 3, 11 (VP), 16, and 22.

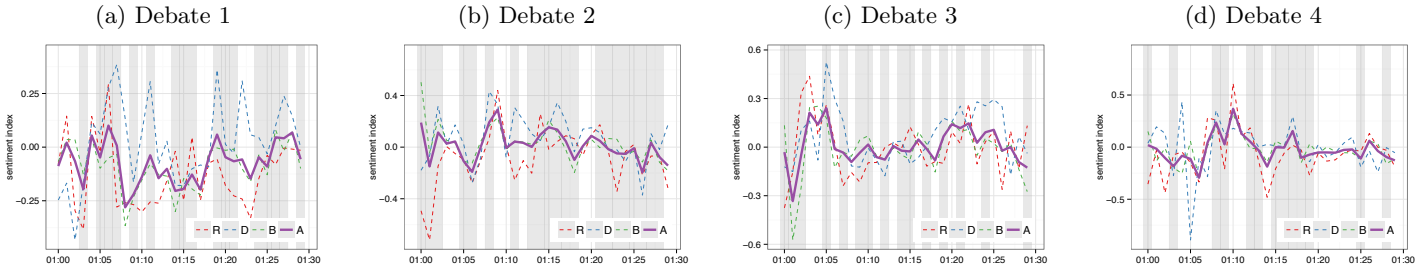

Figure 6: Minute-by-minute sentiment index for focus groups. The purple line ('A') shows the overall sentiment ignoring the group information. The gray areas highlight time ranges where the aggregated sentiment moves with the D_w group's sentiment and not with other groups'. The areas are alternating, indicating that the aggregated sentiment track different groups' sentiment at different times. In each panel, the debate started at 01:00 UTC. The four debates were held on October 3, 11 (VP), 16, and 22.

democrats or because the volume of tweets coming from the D_w group increased. Meanwhile, the aggregate shows a gradual move away from Obama-Biden during this period. Taken together these results indicate that to the extent that Biden was losing favorability or Ryan was gaining it during this period, it was because individuals of negative reactions from the Democratic point of view. This period is conceptually similar to the early stage of Debate 4, in which Obama-Biden gains even as the R_w group leads aggregate sentiment (the region is mostly white). Both of these periods stand out as cases where sentiment shifted due to atypical reactions by partisans.

By contrast, Debate 3 shows the more typical partisan pattern that might be expected. The aggregate achieves peaks (local maxima favoring Obama-Biden) in the gray areas and troughs (local minima favoring Romney-Ryan) in the white areas, indicating that the D_w and R_w groups are driving sentiment in their expected directions.

The groups thus detect meaningful distinctions in the interpretation processes related to debate content. Sometimes:

- Aggregate correlates with D_w as D_w moves toward Obama-Biden
- Aggregate correlates with D_w as D_w moves toward Romney-Ryan
- Aggregate correlates with R_w or B_w as D_w moves toward Obama-Biden
- Aggregate correlates with R_w or B_w as D_w moves toward Romney-Ryan

A comparison of the debate content that corresponds to sustained regions of each kind might yield useful knowledge about both what different groups are responding to and how these bias-consistent and bias-inconsistent ideas are communicated and understood within the groups.

5.3.2 *Identifying controversial issues*

Group analysis also permits the examination of individual topics around which there tends to be divided versus consistent interpretations. Figure 7 shows an issue-based sentiment index for the D_w and the R_w groups relative to the candidates. A score to the right of the center line indicates that the issue received more tweets during minutes when the group was expressing positive sentiments in association with Obama-Biden (or negative in association with Romney-Ryan). A score to the left of the center line indicates that the issue received more tweets when the group was leaning toward Romney-Ryan. Figure 7(a) shows issues for the first presidential debate (which focused on economic policy) and Figure 7(b) shows issues for the last presidential debate (i.e., Debate 4, which focused on foreign policy). Issues are ranked by their divergence, with those where the two groups showed the largest divergence at the top and the smallest divergence at the bottom.

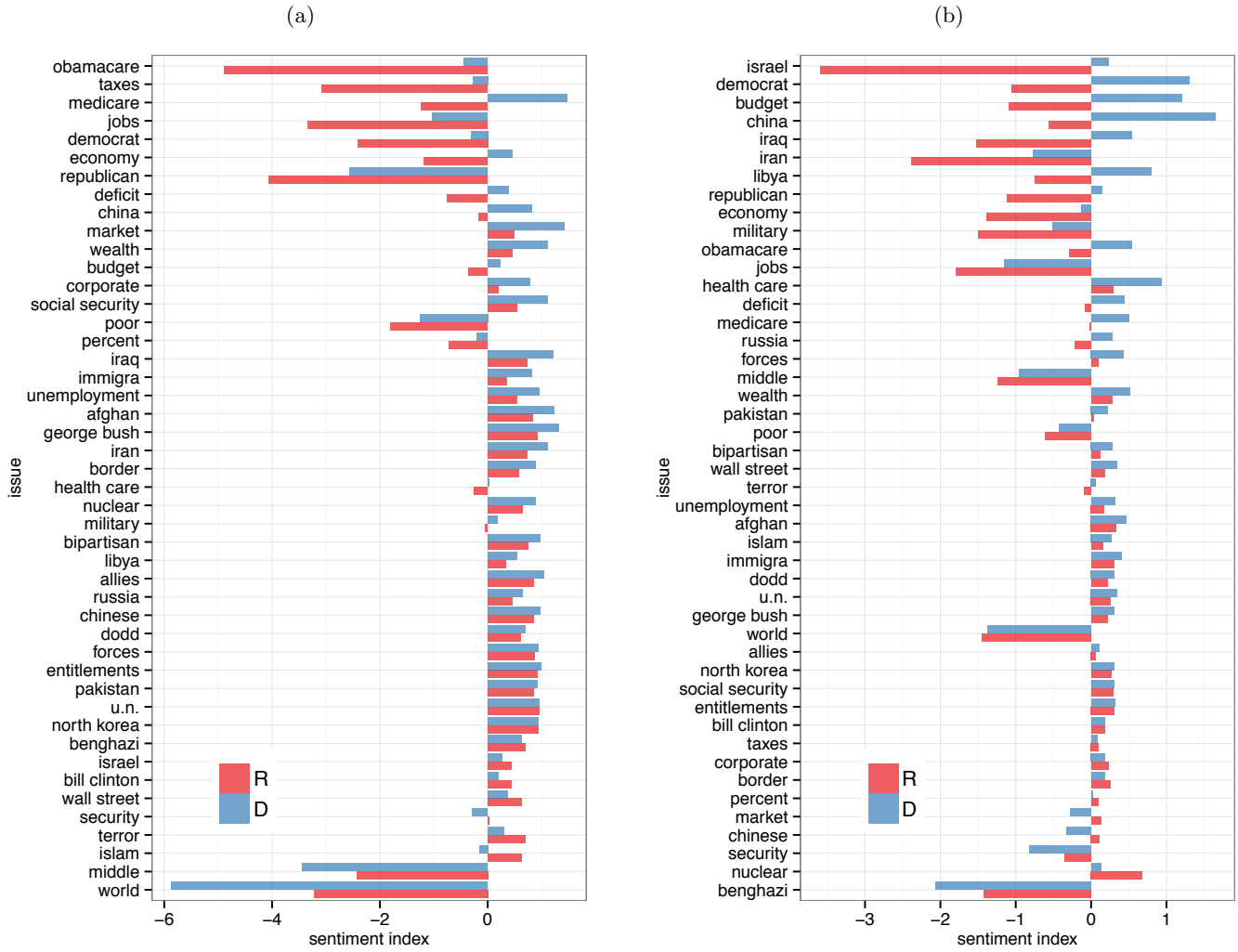

Figure 7: Sentiment index by issues. The issue-based sentiment index for the R_w and D_w groups are shown in (a) the economic debate (Debate 1), and (b) the foreign policy debate (Debate 4). The issues are ordered based on the differences between the two groups' sentiments. Positive sentiment index values indicate the sentiment is supporting the incumbent candidate (Obama) relatively; otherwise it is supporting the challenger (Romney).

The most controversial issues within each debate appear to match conventional expectations. The top five issues during the first presidential debate emphasizing domestic and economic policy were "Obamacare," "taxes," "medicare," "jobs," and "democrat." During the third presidential debate emphasizing foreign policy, the top five issues were "Israel," "democrat," "budget," "China," and "Iraq." Consistent with the group construction, there were only two instances out of 100 where of an issue term on which the D_w group favored Romney-Ryan while the R_w group favored Obama-Biden (these were all toward the bottom of the controversy distribution).

Yet the charts also show several interesting patterns that would likely go undetected in the absence of our method. For example, the groups tended to agree far more than they disagreed in terms of overall valence. In particular during the economic debate, the D_w and R_w group leaning were toward the same ticket in 41 out of 50 issues. During the foreign policy debate (Debate 4) the agreement was less but still the groups expressed sentiment in the same direction on 36 out of 50 issues.

The distribution of the diversity scores in both debates suggest that group divergence is not equally attributable to both groups, however. Of the top five issues, all of them show the R_w group supporting Romney-Ryan with their sentiment. In the economic debate (Debate 1), four of the five issues showed the D_w group leaning to the Romney-Ryan side. The fact that these are the issues with the greatest divergence of opinion is not due to the groups having different valences for these topics, but due to the fact that the R_w group's sentiments were so extreme. Examination of the remainder of the issues also showed that on the less controversial issues (within a debate), both sides seemed to lean toward Obama-Biden. This pattern may be consistent with the idea that Romney, as a challenger, had to attack the President. It also suggests, however, that beyond the key hot button issues there was some underlying satisfaction with the President's performance.

This method also makes it possible to distinguish different kinds of disagreement. On some issues, such as Iran, there is consistency across groups but inconsistency between debates. Both groups lean toward Obama-Biden on this topic

during the economic debate, when the issue was not salient, but shift to Romney-Ryan during the foreign policy debate (Debate 4) when it was a part of the discussion. Other issues show shifts across groups. Israel showed a slight advantage for Obama-Biden for both groups during the first (economic) debate. During the foreign policy debate, however, the D_w group position remained similar while the R_w group position moved sharply toward Romney-Ryan.

6. DISCUSSION

6.1 Implications

We have presented a system for tracking user responses to novel content that accounts for their underlying biases and helps to identify time windows of dominant sentiment and topics of agreement and controversy. Our system extends and promotes research in social media in several ways.

We demonstrate how a theory-based aggregation of tweet streams can be used to create a system for process inferences. We identified tweet attributes (known political biases of Twitter users) that we expected would correspond with particular kinds of responses to the phenomenon of interest (political debates). By distinguishing and then grouping tweets based on these attributes, we reduced the likelihood that our system would report established, expected effects as novel or that it would miss important shifts obscured by several sources of heterogeneity. As a result, our system reports substantively meaningful shifts that are worthy of further investigation, even in cases where the effects do not match "ground truths" that are already known.

This approach should have useful applications in a variety of fields within and beyond politics. Though substantial resources are spent on understanding how political messages influence individual interpretations and voting decisions, we are not aware of any other means of detecting which political statements in a stream of content individuals respond to when in their natural social contexts and settings. More broadly, our system could be applied for any inference where it is useful to know whether an individual is responding to new information or simply re-stating an established bias.

We also demonstrate the usefulness of identifying "average" users rather than "elites" for the purpose of measuring responses. The finding that "elite" political Twitter users do not appear to respond at all to the content of the debates provides further support for the concern about partisan divisions within the media and blogosphere [1]. The results provided by our system demonstrates that this problem may be more manifest through the reporting or repetition of statements from political professionals, however, rather than a general tendency for individuals to refuse to respond to new political content. Further research might examine whether concern is warranted by considering the extent to which these "inflexible elites" influence "average" citizens. Our system's technique for controlling for the influence of individuals' existing biases would be important to such study. Our system also outputs a means of systematically identifying and analyzing controversial issues. This task is difficult for typical approaches to Twitter analysis because of the multiple sources of heterogeneity. Our system distinguishes between topics about which there is general disagreement and topics about which controversy only emerges in the context of an ongoing political contest.

6.2 Limitations

As our system is the first to focus on controlling for expected outcomes in the service of process inferences it contains several limitations that we hope can be addressed through further research.

First, our method for identifying user biases is rudimentary. We considered only a single behavior—tweeting during national political conventions—and the hand-coding of users suggested that the method contained more error than is desirable. Future work should focus on identifying more precisely the quantity of interest such as the extent to which an individual is biased toward a particular point of view but not publicly or professionally committed to adhere to that view. Algorithms might include sampling from users' consumption of other media content as well as expressed sentiment toward particular individuals or products (in cases where the subject of interest is not political).

Second, our method identifies areas of data for further investigation of processes—regions (within the debate) and topics of potential interest—rather than processes themselves. The ultimate goal of a process inference system should be the quantitative corroboration or rejection of an a priori hypothesis regarding the processes that produce a stream of data. In the context of debates these investigations might take the form of analysis of debate content at these particular points in time or on the polarizing or convergent issues. A more fully developed process inference tool would bring much of this analysis online, however, either through the integration of transcripts in real-time or through the principled comparison of controversial/non-controversial topics to other streams of content, such as from elite tweet streams or official partisan statements and releases.

Finally, our system does not make a distinction between the direct effects of the debates themselves and in the indirect effects of the larger social response to the debates. For example, it would be useful to distinguish between sentiment changes due to the construction of many independent messages by group members from sentiment changes due to the mass re-tweeting of a single message (or small set of messages) by group members. Further work might consider principled ways to distinguish "source" effects from "social" effects within these response streams.

7. CONCLUSIONS

We have presented a system using tweet streams to identify meaningful baseline tendencies in user tweeting activity, and then applied those baselines to distinguish and interpret user responses to novel messages and information. Our results show that while partisan bias is alive and well in both media selection and political interpretation, individuals can at least temporarily be persuaded to express atypical points of view, in particular around issues that are not central to the campaign. However, the inflexibility of elites suggests a more pernicious problem of commentators' unyielding partisanship in their comments.

8. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Lazer Lab at Northeastern University, supported in part by MURI grant #504026, DTRA grant #509475, and ARO #504033.

9. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43. ACM, 2005.

[2] W. Bennett and S. Iyengar. A new era of minimal effects? The changing foundations of political communication. *Journal of Communication*, 58(4):707–731, 2008.

[3] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, 2011.

[4] A. Boutet, H. Kim, E. Yoneki, et al. What's in your tweets? I know who you supported in the U.K. 2010 general election. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 411–414, 2012.

[5] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 66–73, 2011.

[6] F. Ciulla, D. Mocanu, A. Baronchelli, B. Goncalves, N. Perra, and A. Vespignani. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1(1):8, 2012.

[7] N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[8] N. Diakopoulos and D. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 2010 ACM Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[9] L. Dinauer and E. Fink. Interattitude structure and attitude dynamics. *Human Communication Research*, 31(1):1–32, 2005.

[10] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, volume 6, pages 417–422, 2006.

[11] R. Frick. Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods*, 30(3):527–535, 1998.

[12] D. Gayo-Avello. I wanted to predict elections with Twitter and all I got was this lousy paper: A balanced survey on election prediction using Twitter data. *arXiv preprint arXiv:1204.6441*, 2012.

[13] S. Green. A rhetorical theory of diffusion. *Academy of Management Review*, 29(4):653–669, 2004.

[14] A. Hayes. *Statistical Methods for Communication Science*. Lawrence Erlbaum, 2005.

[15] C. Hovland, O. Harvey, and M. Sherif. Assimilation and contrast effects in reactions to communication and attitude change. *The Journal of Abnormal and Social Psychology*, 55(2):244–252, 1957.

[16] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32. ACM, 2011.

[17] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 ACM Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[18] S. Meraz. Is there an elite hold? Traditional media to social media agenda setting influence in blog networks. *Journal of Computer-Mediated Communication*, 14(3):682–707, 2009.

[19] P. Metaxas and E. Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

[20] E. Mustafaraj, S. Finn, C. Whitlock, and P. Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. In *Proceedings of the 2011 IEEE International Conference on Social Computing*, pages 103–110. IEEE, 2011.

[21] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using Twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pages 189–198. ACM, 2012.

[22] F. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[23] E. Noelle-Neumann. The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2):43–51, 2006.

[24] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[25] M. Pennacchiotti and A. Popescu. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 430–438. ACM, 2011.

[26] R. Petty and J. Cacioppo. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19(1):123–205, 1986.

[27] D. Shamma, L. Kennedy, and E. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM Workshop on Social media*, pages 3–10. ACM, 2009.

[28] C. Strapparava and A. Valitutti. Wordnet-affect: An affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 4, pages 1083–1086, 2004.

[29] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with tTwitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.