Online and social media data as a flawed continuous panel survey[*]

Fernando Diaz, Michael Gamon, Jake Hofman, Emre Kıcıman, and David Rothschild[†]

Microsoft Research

Abstract

There is a large body of research on utilizing online activity to predict various real world outcomes, ranging from outbreaks of influenza to outcomes of elections. There is considerably less work, however, on using this data to understand topic-specific interest and opinion amongst the general population and specific demographic subgroups, as currently measured by relatively expensive surveys. Here we investigate this possibility by studying a full census of all Twitter activity during the 2012 election cycle along with comprehensive search history of a large panel of internet users during the same period, highlighting the challenges in interpreting online and social media activity as the results of a survey. As noted in existing work, the online population is a non-representative sample of the offline world (e.g., the U.S. voting population). We extend this work to show how demographic skew and user participation is non-stationary and unpredictable over time. In addition, the nature of user contributions varies wildly around important events. Finally, we note subtle problems in mapping what people are sharing or consuming online to specific sentiment or opinion measures around a particular topic. These issues must be addressed before meaningful insight about public interest and opinion can be reliably extracted from online and social media data.

I. Introduction

Online and social media provide an increasingly popular forum for public discussion of a large number of topics, including political conversations. Digital records of these discussions complement traditional approaches to opinion polling and offer the opportunity to better understand societal opinions at large (Golder & Macy, 2012). These sources have the advantage of *scale*. With over a billion active participants globally, online and social media potentially each capture the revealed actions and stated thoughts of a large segment of the population. These sources also have the advantage of *low latency*. Real-time online and social media data allow for continuous analysis as events unfold, and time granular post-event analysis critical to isolating the impact of key sub-events.

In recent years, much research has focused on understanding the expression of political opinion online, and exploring its use as an alternative data collection modality for surveys and fundamental election data as a way to predict elections, identify support, and related tasks. The results have been mixed. For example, while many papers have reported positive correlations between measures of social media and political outcomes (O'Connor et al., 2010; DiGrazia et al., 2013; Tumasjan, 2010), others have criticized their methodologies and reported contradictory, negative results (Gayo-Avello, 2011; Gayo-Avello, 2012; Lui et al., 2011). There are two key reasons for potential problems in this research. First, it frequently focuses on election outcomes where standard data and methods are extremely accurate already. Second, it does not account for the uniquely non-stationary nature of the data as technology and its users evolve and, thus, how the data correlates with outcomes such as support, likeliness to vote, and donations. Given these contradictory results without significant, fundamental progress in resolving the underlying questions of the usefulness of online and social media data, we believe a new approach is needed.

In this paper, rather than attempt to reproduce a traditional task, such as predicting election results from online and social media data, we instead ask a seemingly simpler question:

> *If we were to assume that online and social media data is the output of some hypothetical pseudo-survey methodology, how would this methodology differ from conventional survey techniques?*

We believe that a rigorous answer to this question–highlighting the fundamental differences between online and social media data and gold standard survey methodologies–can provide a new roadmap toward the use of online and social media data. To this end, we analyze what, to our knowledge, is the largest such corpora to be used for this purpose: the comprehensive search activity of a large panel of web users, and the complete set of tweets during the period of the 2012 election cycle. We focus our analysis on two fundamental characteristics of a (pseudo-) survey methodology: (1) the dynamics of demographics and participation in the pseudo-survey; and (2) the dynamics of topics addressed in the pseudo-survey.

The results of our analysis confirm what past research has suggested: both participation and topical coverage are more dynamic and less predictable than conventional survey methodologies. Our study shows that not only are the participants in online and social media platforms unrepresentative of the offline population, as characterized by key demographics such as gender and geography, participation also shifts dramatically on a daily and even hourly basis, especially around key events. In other words, at just the times when online and social media data's real-time nature holds the most promise for providing timely insights, the participant base is most aggressively in flux.

Online and social media activity function as an opt-in panel where different users engage at different amounts during different times. Most research counts each of these engagements independently, ignoring user identity information. If this were a survey, it would be the equivalent of allowing users to respond as many times as they want. Further, we explore the value of treating the online population as a panel rather than a cross-section, as there is a small fraction—but large number—of users who repeatedly discuss the same topics.

Studying the topics addressed by participants in our hypothetical pseudo-survey, we find that the population of people searching and tweeting contribute different types of information at different times. Specifically, the topics they discuss shift during major events, as activity moves from content sharing to active commentary. Treated as a survey, this corresponds to users answering systematically different questions at different times.

In short, if online and social media data are to be treated as surveys, they must be treated as very flawed surveys indeed. Conventional surveys follow a rigorous procedure, asking the

3

same question—Gallup has asked the same presidential approval question to its respondents since the late 1930's—to repeated cross sections of a random sample of a representative group of people. A search and social "survey," however, is essentially polling a varying, non-random sample of voluntary participants who selectively respond to questions of their choice.

While online and social media data does not fit the mold of a traditional survey because of the complexities of population and selection biases and the difficulties of interpreting open-ended answers to unknown questions, our research provides hope for several different methods of interpreting the data. First, search and social media data can provide insights into the levels of interest and engagement among different online populations. Who is responding and when they are doing so, as well as the topics and sentiments expressed, are important. This applies at the individual level as well, potentially allowing for micro-targeting by interest or support. Second, researchers can re-weight the data to either a fixed or representative target population. This provides a more consistent measure to correlate with outcomes such as voting or donations. Furthermore, online and social media data can be interpreted as a panel response rather than a cross-section, where insights can be gained from studying the shift in individual opinions and behaviors over time. Our detailed data shows individual-level actions and repeat actions. Using a panel can provide unique insight into how conversations shift over the course of an event.

These directions for future work complement one another as well as traditional surveys. They represent a step towards exploiting the strengths of this data to provide low latency, highly quantifiable answers to relatively unexplored questions of interest from varying populations, rather than the standard desire to answer traditional questions, such as election outcomes, that are still most efficiently answered by traditional methods.

The scale and real-time nature of online and social media promises to expand our range of insights into new questions and domains, and to provide answers in real-time. This paper uses a uniquely comprehensive dataset of search queries and Twitter messages to outline a new method of considering online and social media data as a flawed panel survey.

II.   Past Research

Much research has been done on the promise of online and social media, from the potential for online and social media to be used for grass roots political mobilization (Bond et al.

4

2012) and astroturfing (Ratkiewicz et al. 2010) to predicting outbreaks of influenza (Ginsberg et al., 2009), reproducing economic indices (Choi and Varian, 2012) and measuring people's happiness and mood (Dodds et al., 2011; Golder and Macy, 2011). Many studies have documented the existence of a robust and vibrant political discussion on social media across many countries and elections over the past several years (Hanna et al., 2011; Moe and Larsson, 2012; Mustafaraj and Metaxas 2011, Vallina-Rodriguez, 2012; Boutet et al., 2012; Ceron et al., 2012). As participation in online and social media discussions continues to grow more prevalent, monitoring and analysis of these information sources promises to provide new insights into political events and policies. Already, several commercial services track the popularity of political candidates and political issues in online and social media data.[1] Others, such as Conover et al. (2013) and Huberman et al. (2008) outline further applications of social media monitoring and user classification, including for political advertising.

Many researchers have reported positive correlations between the volume of social media mentions of political candidates and the results of various elections, from presidential elections around the world to US Congressional and UK Parliamentary races (Soler et al., 2012; Skoric et al., 2012; DiGrazia et al., 2013). Similar research has found positive correlations between the volume-weighted sentiments expressed in social media and electoral results and public opinion polls (Bermingham and Smeaton, 2011; Lee et al., 2013; O'Connor et al., 2010; Monti et al., 2013). Notably, prior to the 2012 US presidential election, Choy et al. (2012) predicted state-level election results, based on a demographically re-weighted sentiment analysis of tweets. Arguing the necessity of incorporating information beyond volume and sentiment, Shi et al. (2012) present a more sophisticated model that includes behavioral attributes of users, including follower counts and engagement rates, to predict election results. Here we focus less on predicting event outcomes and more on understanding public opinion.

Despite such reports of success, serious questions have been raised about the usage of online and social media in the current literature (Gayo-Avello, 2011; Gayo-Avello, 2012; Lui, 2011), including potential methodological issues in data normalization and mistakes in causal modeling. Several attempts to reproduce published techniques in the context of different

---

[1] Prominent examples include Google Trends (http://www.google.com/trends/) for search and Topsy (http://topsy.com/) for social media.

elections have found weak or insignificant correlations or predictions with electoral outcomes (Gayo-Avello et al., 2011; He et al, 2012; Johnson, Shukla and Shukla; Jungherr et al., 2012). Complementing social media analyses, Lui et al. (2011) examines search query volume, and find no strong correlation between search popularity of candidates and election outcomes. Others point to the challenges of predicting electoral outcomes based on data that is easily manipulated through astroturfing and "Google-bombing" campaigns (Choy et al, 2011; Metaxas and Mustafaraj, 2012).

Recent research has begun to study more deeply the processes that generate online and social media data, including demographic and behavioral attributes of participants. For example, Hampton et al. (2011) provide a host of self-reported online behavior with a traditional cross-sectional poll of Americans, while Goel et al. (2012) examine variation in revealed preferences across demographic groups using web browsing logs from a large online panel. Mislove et al. (2011) document the demographic skew on Twitter along with other studies that focus on inferring unknown demographics and attributes of social media users (Chang et al., 2010; Pennacchiotti and Popescu, 2011; Zamal et al., 2012; Kosinski et al., 2013), but do little in joining this information with activity and content. On the other hand, papers studying the implications of the online or social media data for politics have mostly ignored demographics (Tumasjan et al., 2010; O'Connor et al., 2010; DiGrazia et al., 2013), which shift heavily from election to election (Hampton et al., 2011). Borondo et al. (2013) characterize and model social media participation by politicians and the electorate during the 2011 Spanish presidential election. Highlighting the importance of understanding the implications of participation behavior, Chen et al. (2012) find significant variation in the relative predictive power of subgroups of Twitter users with different behaviors. In the context of web search, both task-based (Broder, 2002) and intent-based (Li et al., 2008) classification schemes have been explored. Java, et al. (2007) develop a similar taxonomy for Twitter. Shifting and non-stationary intents have been explored (Diaz, 2009), but only in the context of optimizing search engines.

Eliminating selection bias and non-representativeness is new work for surveys and we borrow from that budding research for polling samples of convenience. Wang et al. (2013) show how even extremely non-representative opt-in samples can be used as accurate representation of

the opinions of a target population with the proper translation of the raw data into an appropriate indicator. Further, we also exploit the panel nature of our data.

Data collection is difficult and expensive, thus most papers rely on samples of either search or Twitter data to represent online and social media discussions. Papers using both query log data and Twitter firehose data have been confined to the information retrieval literature (Kairam et al., 2013; Alonso et al., 2013; Yom-Tov and Diaz, 2013), where the focus has been on optimizing search engine effectiveness. Outside of this community, there are limited examples of papers utilizing the full Twitter stream (Wu et al., 2011), but those are mainly focused on examining the network, not the content. Morstatter et al. (2013) show that there are meaningful sampling issues between the full firehose and any limited streams from Twitter. Several papers have been written by academics within Google, Microsoft, or Yahoo! which utilize the full census of search over a period of time (Goel et al., 2010; Choi and Varian, 2012), but, none of them, to our knowledge, examine the same methods or outcomes as our paper. There is limited work comparing online and social media data to traditional media sources (Zhao and Jiang, 2011).

III.   Data

Our study focuses on two large-scale datasets representing online and social media data, one dataset based on search queries at major search engines, and another representing a complete selection of tweets during the period of the 2012 election cycle. On the surface, these two platforms share several similarities. Both record short snippets of online, user-generated text; both are indicative of people's real-time interests; both can be tied to individuals, presenting a longitudinal record of a person's interests. However, there is also a key difference in people's intentions while using these platforms. Social media is intended as a public or semi-public communication with other people, whereas search queries are typically intended for private retrieval and consumption of information. Twitter content is a little longer as well, limited to 140 characters, but represents a complete thought, while search queries generally contain just 2-3 words or fewer.

We use Comscore's records of searches that contain either the keywords "Obama" and/or "Romney" during the 2012 election cycle. Comscore follows a subset of the population, so this

data is not a census, but it includes comprehensive coverage for these users across major search engines. Comscore pays these 267,518 panel members in collecting this data so that they are aware of the logging, but they use a very unobtrusive method of data collection, intended to limit the bias of being tracked. We are able to capture whenever a user in their panel searched for either Obama or Romney from July 1, 2012 through November 6, 2012. While anonymized, the Comscore data includes detailed demographics for each user, including gender, geographical location, income, and age.

For our social dataset, we use the Twitter firehose—the complete stream of Twitter messages—to capture all of the English language tweets made between August 1, 2012 and November 6, 2012, inclusively, that mention the words "Obama" and/or "Romney". For each tweet, we capture the text of the tweet, shared URLs, the date and time the message was written, and profile information about the author, including name and self-identified location. As Twitter profiles do not explicitly report demographic information, including gender and location, we infer these attributes; this process is detailed in the Appendix. Shortened URLs embedded in a tweet are resolved to their full URLs at the time the tweet is collected.

This results in a dataset with similar information from two of the most meaningful online and social media sources: search and Twitter. For any given search or tweet we have the following: timestamp, a user id, a user name, gender, geography, discussion, and whether it includes Obama, Romney, or both candidates. We also record the top co-occurring words in the text for each day. For search we have age and income information for the users. We store the data by mentions of Obama, Romney, and both for every hour in the dataset, along with counts for overall activity at each time. Table 1 shows some summary statistics aggregating over the entire period of our dataset from August 1, 2012 through November 6, 2012.

<Table 1. Summary statistics of demographics associated with search and Twitter>

There are some key trends that permeate both platforms. First, users discuss Obama much more often than Romney, likely because Obama is the sitting president and Romney is the challenger. That said, Obama is also the progressive candidate who captured 62% of the 18-29 two-party vote-share while Romney captured 56% of the 65+ two-party vote-share; so this could also be driven by the younger demographics of online and social media. Second, while this varies by platform, election-related online and social media activity is dominated by males, while

the voting population is 54% female. Third, the northeastern states, including the Mid-Atlantic, which includes New York, (with DC, DE, and MD) and New England, are overrepresented compared to the average voter. Fourth, the demographics of our search data show politically active users are younger than the average voter with a different income distribution.

The two mediums do differ slightly in their representativeness. First, in raw counts the engagement by candidate is much closer to the voter percentages in Twitter than in search. Second, on gender, search is closer to a representative split than Twitter.

The demographics of social media users are not representative of the voting population. While this is not a surprising finding, it is not yet well documented for political discussion. It is worth noting that these observed demographics of politically active users on Twitter differ from previous studies of demographics of general Twitter users indicating that demographics must be topically conditioned, an important finding for social media surveys in general. For example, 64% of Twitter users are female compared with 43% of individuals who made at least one political tweet during our time period being female and 36% of tweets about election 2014 being sent by females (Hampton, 2011).

Further, we also know that demographics on social media shift over time; we would be wrong to assume that any relationship between social media users and voting populations would be the same in 2016 or 2020 as adaptation of social media shifts within demographic groups much faster and differently than trends among voters. For example, from 2008 to 2010 the population of users between 18 and 22 on social media decreased from 28% to 16%, while the percent between 50 and 65 more than doubled from 9% to 20% (Hampton, 2011). Thus, with age being heavily correlated with political positions, the relationship between raw counts and actual outcomes will likely vary over time.[2]

IV.   Online and Social Media Data as a Survey

How significant and fundamental are the structural discrepancies between how survey data and online search and social media data are collected? We answer this question in three subsections. First, we show that online user demographics are not only biased, but shift dramatically around

---

[2] See the 2012 exit polls: http://elections.nytimes.com/2012/results/president/exit-polls where Obama won 60% of 18-29 year olds and 44% of 65+ year olds.

major events, just when we likely care most about the data users generate. Second, we show that individual-level engagement varies dramatically around major events. Third, we show that the nature of user activity in online and social media discussions also shifts around major events.

Unpredictably Shifting Demographics

Our first major contribution is to show that engagement by demographics is not only unrepresentative overall, shifting between election cycles as previously noted in critiques such as Hampton (2011), but also shifting dramatically within cycles. All of the statistics in Table 1 are averaged over the last three months of the election cycle; they are not necessarily representative of any point in the election cycle. There are two types of shifts we consider: periodic shifts, following diurnal and septan cycles common in online data, and event-driven shifts, where the demographics and engagement on online and social media dramatically change during a major event such as a presidential debate or breaking news event.

It should not be surprising that activity increases as Election Day approaches. There are also typical daily patterns of discussion. This is ultimately not critical to the analysis as long as it is accounted for. In addition, however, the percentage of online and social media discussion that engages in political discussions shifts dramatically during major events; this movement is much less predictable and thus difficult to correct for ex-ante. Figure 1 shows the percentage of search queries and social media messages that include either a mention of Obama or Romney (or both) during our timeframe (henceforth referred to as *discussion about the presidential candidates*). Notice that the massive jumps during two conventions and the three debates range from a few multiples to 3 orders of magnitude over the normal level of activity depending on the platform and timeframe. For example, in Twitter the day over day differences move around 1 order of magnitude for all three debates: 10x, 8x, and 9x respectively, but the search day over day differences decline down to just 2x for the last debate after holding steady in the first two debates. The differences are both substantial and unpredictable.

*<**Figure 1.** Percent of overall search and Twitter discussion about the presidential candidates>*

The overall male to female ratio does not resemble the voting population and it shifts dramatically away from being male-dominated at key moments. Figure 2 charts the percentage of online and social media discussion about politics that males generate. For search we focus on the

two major party conventions. Not shown on the chart, men actually search more, but women search for politics more often when they search. This leads to women having a similar turnout, despite lower search totals. This holds steady during the Republican convention, but in the final day of the Democratic convention only 33% of searchers were males, while the average day had a ratio of 55% males. After that dramatic shift, there was no significant difference in gender representation during the debates. For Twitter we focus on the three debates. For Twitter, which has a much greater regular imbalance of men to women, the increase in female participation during major events leads to a more representative gender balance. During the three debates the day over day difference was 8 percentage points, 5 percentage points, then 3 percentage points respectively. Again the differences in both platforms were both significant and unpredictable. For completeness, the full plot from August 1 to November 6 is given in the Online Appendix as Figure OA1.

<Figure 2. Percent of search and Twitter discussion about the presidential candidates conducted by males>

Similar to gender shifts, there are shifts in the geographical divides during major events as well. Figure 3 charts the percentage of discussion from each geographical division during our timeframe for each of the two platforms. Here the imbalance actually gets worse during major events. The proportion of tweets from the Mid-Atlantic (15% of voters) jumps during the conventions and debates, just as they drop for the Mountain (7% of voters) region. The Mountain region's drop is sizeable; it averages 6.6% of discussion between August 15 and September 15, 2012, but oscillates between 4.7% and 5.4% during the two conventions. The Mid-Atlantic averages 20.2% during that timeframe and jumps above 23% for the entire Republican convention, but only peaks at 21.3% for the Democratic convention. Again, these are large leaps, but the inconsistency makes it hard to predict and account for in the data. For completeness, we have the full plot from August 1 to November 6 in the Online Appendix as Figure OA2.

<Figure 3. Percent of Twitter discussion about the presidential candidates conducted by geographical region>

What all of these differences and movements in demographics and responses demonstrate is that any index that simply shows the unadjusted response of the population can be misleading. A consumer of an index that does not adjust for changes in the population is not capable of disentangling the meaning of daily or hourly changes. A change in interest levels or sentiment is

more likely to reflect shifts in social media engagement (sometimes predictable, sometimes unpredictable), rather than any real shifts in interest or sentiment within a consistent group.

Answering Multiple Times

Conventional surveys limit respondents to a single response. If we wish to consider online and social media discussion as a survey, we must take into account the fact that, unlike in a conventional survey, search and Twitter allow individuals to "respond" multiple times. Here, we move beyond aggregated results and analyze the individual-level responses.

Figure 4 shows the distribution of daily messages per user. Most discussion is dominated by just a few users: 70% of tweets come from the top 10% of users, with 40% of the discussion from the top 1% of users. Further, this persists over time, with 10% of all tweets generated by the 0.04% of users who tweeted on more than 90% of the days in our sample. While the typical number of posts per individual rise during the debates, the variance in individual activity drops on those same days.

*<**Figure 4.** Average number of tweets about the presidential candidates per user>*

Since user activity varies substantially across the population, we get a very different view of engagement if we restrict our analysis to one "response" per user per day. Figure 5 repeats Figure 2 and shows the difference in results when we count all tweets compared to counting each user's contributions only once a day in examining the fraction of election-related activity on Twitter generated by males. We see both an overall shift in the amount of election-related activity from males counting all tweets (an average of 68%) and counting users just once (an average of 63%) as well as different day-to-day movements in these measures. For instance, the drop between the day before the debate and the first debate was 8 percentage points counting all tweets and 9 percentage points counting users just once, and the drop around the second debate was 5 percentage points counting all tweets and 7 percentage points counting users just once. The deviation around the debates appears sharper when users are counted just once.

*<**Figure 5.** Percent of Twitter discussion about the presidential candidates conducted by males>*

More surprisingly, we find that if a user engages more than once a day, it is likely not about the same candidate. If a user tweets twice with either a mention of Obama or Romney in the tweet (but not both at the same time) 34% of the time she mentions Obama once and Romney

once. If a user tweets three times in a day with either Obama or Romney in the tweet (but not both at the same time) 55% of the time the mix is 2 for one candidate and 1 for another candidate. People who tweet about politics are not likely to be ambivalent voters, thus this provides further evidence that simple counts of mentions are not good indicators of individual-level support.

<Table 2. Percent of tweets on any given day by the same user that reference Obama>
Shifting Questions

The nature of user interaction shifts around major events. For example, an election-related tweet containing a URL is likely to be informational in nature, linking to media around a candidate, whereas tweets expressing opinions and support may contain fewer links to additional content. Figure 6 charts the percentage of election-related tweets with a URL, showing dramatic daily shifts. Overall, 62% of election-related tweets the day before the first debate contained a URL and only 12% of tweets the day of the first debate did. The shift is even more pronounced for discussions involving Obama. The day before the first debate Obama or Romney tweets were about equally likely to contain a URL, but the day of the debate a tweet mentioning Obama was 140% more likely to contain a URL than one mentioning Romney.

<Figure 6. Percent of Twitter discussion about Obama, Romney, and both that contain a URL>

These changes are not uniform across different platforms, which answer different questions about interest at different times. Figure 1 shows how the discussion levels for Obama, Romney, and both move during the election cycle. We compare this engagement with the mainstream media and break it down by each of the three topics. Figure 7 shows the progression of the discussion about Obama, Romney, or both in search, Twitter, and mainstream media. As a proxy for the mainstream media we use the number of indexed stories by Lexis-Nexis, which covers an array of general news sources. First, notice that Obama is always searched for the most; people have an interest in learning about the sitting president, while Romney tops Obama at points in what people are interested in discussing. Second, the mainstream media slowly builds up coverage around major events and then slowly winds them down, while online and social media hit the event with a more sudden burst and decline. For example, while there are sharp differences in the discussion level about Obama, Romney, and both on the three debate days in

13

search and Twitter, the discussion is statically flat in the major newspapers. Online and social media data provide a different interest index than that of the mainstream media.

*<**Figure 7.** Percent of overall search, Twitter, and print media discussion about the Obama, Romney, and both>*

We also examine the terms that co-occur with candidate mentions to gain further insight into the nature of political discussion via online and social media data. Figure 8 shows a selection of the top terms mentioned alongside Obama on Twitter over time, each day, where darker cells indicate higher interest (corresponding figures for search and Obama and both search and Twitter for Romney are in the Appendix as A1-A3). First, this provides further verification of how the discussion changes between general interest terms to trending terms around specific events. Second, important for the campaigns, charts like Figure 8 can provide a quick and clear indicator of how their messaging is received by the different populations. We can spot the evolution of terms on both platforms that the campaigns are either promoting or avoiding. For instance, this illustrates how the attacks on the U.S. embassy in Benghazi, Libya shift from being "attack" and "Libya" into "Benghazi". We can focus on major events and see what people are discussing. During the Democratic Convention, searches were dominated by queries about Obama's family members: Michelle, Sasha, and Malia, but tweets focused more on policy terms and Bill Clinton. We can compare across candidates and see how while Hurricane Sandy, and related terms, dominated the discussion around Obama in the days before the election, Romney was battered with searches regarding false claims in his advertisement about Jeep. Domain experts know Jeep as a negative term for Romney and Benghazi as a negative term for Obama; this type of chart is a powerful sentiment indicator for domain experts.

*<**Figure 8.** Top terms co-occurring with "Obama" on Twitter>*

V.   Discussion

This paper utilizes large corpora of search and Twitter data during the 2012 election cycle to provide a multi-platform view of online and social media data. We find that if search and Twitter data are to be treated as a survey, they would follow a very peculiar methodology: participation is a time-varying demographically biased sample of the population, participants are effectively continuously answering different "survey" questions, and finally, participants can choose how often they respond, and to which questions.

Exactly how these data should be utilized to complement existing polls and surveys is an open question. We do not propose to have all of the answers, but by understanding the data we can redirect research and methodology towards this goal. The main results of this paper already show that online and media data can provide insights into the levels of interest and engagement among different population demographics. However, we propose continuously re-weighting the data to a specified target population; either a consistent representation of online or social media users or the voting population.

There are two key reasons to re-weight data to stable populations: to provide meaningful aggregated comparisons and create consistent correlations with key outcomes. For example, women are underrepresented in the raw volume of tweets, but tweet more often about politics than men. Thus, the average level of conversation increases when we weight by women representatively. If we want to provide an account of the general population, then it is key to re-weight by gender, age, geography, individual, etc. Since demographics of active users shift so much over time it is also key to provide a stable demographic sample if correlations to outcomes the campaigns care about are going to hold over time and elections. We opened this paper with one of the most celebrated examples of online data's usefulness; predicting influenza with search data. But, recent work by Butler (2013) shows that that relationship failed soon after the initial papers were published. Correlating topline results with outcomes is going to be fragile with input data that is in constant flux, but the more we are able to break the input data into its elements the more stable it will become. The methods for doing this are the same as the new methods being tested for non-representative surveys.

Another methodology for utilizing the data is to treat online and social media as a panel, where we can track the opinions of a set of users over a period of time. For instance, in Figure 9 we show the median days since the last election-related tweet by the same person for any tweet on a given day. For most days in the sample this distance is within a week—the median is 6 days. This means that if one picks a randomly selected tweet from a random user on a typical day, that user's last election-related post occurred less than a week earlier. However we see a drastic shift in this quantity around conventions and debates. For instance, more than half of the activity on the October 3rd debate came from individuals who had not previously posted about the election. These are likely to be a very different subset of the population than regularly

engaged users. At the same time, this repeated usage shows the promise of capturing when and how conversations change, not just between users, but within users.

*<**Figure 9.** Median days since the last election-related tweet by the same person for any given tweet on that day>*

We do not go into detail on sentiment in this paper for two key reasons. First, per-tweet sentiment is affected by the same selection issues as counts, thus we saw counting as more fundamental to describing the pertinent challenges and sentiment as a secondary issue to explore in later work. Second, we view co-occurring words as a crude, domain specific sentiment for domain experts. Future work is already underway in using co-occurring words to create continuously updating domain-specific sentiment.

These results should be interesting for campaigns or observers who are invested in following how ideas are flowing through the population with quantitative precision and low latency. We are not pursuing matching the ground truth of an election outcome or current level of support, but a new metric that could be used to study the impact of the billions of dollars campaigns spend each cycle on the interest levels and discussions during the campaign. Campaigns adjust their messages daily (or even hourly) and our concepts of low latency and quantitative indicators of interest can help determine the effectiveness of those messages within detailed demographic groups. Further, the methods outlined in the paper will also help target the individuals as they are either followed directly or as part of a demographic group.

References

Alfonso, Omar, Qifa Ke, Kartikay Khandelwal, and Srinivas Vadrevu. 2013. Exploiting entities in social media. In Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval (ESAIR '13). ACM, New York, NY, USA, 9-12.

Bermingham, Adam, and Alan F. Smeaton. 2011. "On using Twitter to monitor political sentiment and predict election results."

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam DI Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature*, *489* 7415:295-298.

Borondo, J., A. J. Morales, J. C. Losada, and R. M. Benito. 2012. "Characterizing and modeling an electoral campaign in the context of twitter: 2011 Spanish presidential election as a case study." *Chaos: an interdisciplinary journal of nonlinear science*, *22* 2:023138.

Boutet, Antoine, Hyoungshick Kim, and Eiko Yoneki. 2012. "What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election." In *ICWSM*.

Broder, Andrei. 2002. "A taxonomy of web search." In *ACM Sigir forum, 36* 2:3-10. ACM.

Butler, Declan. 2013. When Google got flu wrong. *Nature*, *494*(7436), 155.

Ceron, Andrea, Luigi Curini, Stefano M. Iacus, and Giuseppe Porro. 2013. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France." *New Media & Society*, 1461444813480466.

Chang, Jonathan, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. "ePluribus: Ethnicity on Social Networks." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. ACM.

Chen, Lu, Wenbo Wang, and Amit P. Sheth. 2012. "Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries." In *Social Informatics* pp. 379-392. Springer Berlin Heidelberg.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the present with google trends." *Economic Record*, *88* s1:2-9.

Choy, Murphy, Michelle Cheong, Ma Nang Laik, and Koo Ping Shung. 2012. "US Presidential Election 2012 Prediction using Census Corrected Twitter Model." *arXiv preprint arXiv:1211.0938*.

Conover, Michael D., Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2013. "The digital evolution of occupy wall street." *PloS one, 8* 5:e64679.

De Choudhury, Munmun, Scott Counts, and Michael Gamon. 2012a. "Not All Moods are Created Equal! Exploring Human Emotional States in Social Media." AAAI,

De Choudhury, Munmun, Michael Gamon, and Scott Counts. 2012b. "Happy, Nervous, or Surprised? Classification of Human Affective States in Social Media." AAAI.

Diaz, Fernando. 2009. "Integration of news content into web results." In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* 182-191. ACM.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. "More tweets, more votes: Social media as a quantitative indicator of political behavior." *PloS one*, *8* 11:e79449.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter." *PLoS ONE*, 6 12:e26752.

Gayo-Avello, Daniel. 2011. "Don't turn social media into another Literary Digest poll." *Communications of the ACM, 54* 10:121-128.

-----. 2012. "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*.

Gayo-Avello, Daniel, Panagiotis Takis Metaxas, and Eni Mustafaraj. 2011, July. "Limits of electoral predictions using twitter." In *ICWSM*.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014.

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. "Predicting consumer behavior with Web search." *Proceedings of the National Academy of Sciences* 10(741):17486-17490.

Goel, Sharad, Jake M. Hofman, and M. Irmak Sirer. 2012, June. "Who Does What on the Web: A Large-Scale Study of Browsing Behavior." In *ICWSM*.

Golder, Scott A., and Michael W. Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." Science 333.6051 (2011): 1878-1881.

Golder, Scott, and Michael Macy. 2012. "Social science with social media." *ASA Footnotes* 40.1 (2012).

Hampton, Keith, Lauren Sessions Goulet, Lee Rainie, and Kristen Purcell. 2011. "Social networking sites and our lives." http://www.pewinternet.org/files/old-media//Files/Reports/2011/PIP - Social networking sites and our lives.pdf

Hanna, Alexander, Ben Sayre, Leticia Bode, JungHwan Yang, and Dhavan V. Shah. 2011. "Mapping the Political Twitterverse: Candidates and Their Followers in the Midterms." In *ICWSM*.

He, Yulan, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong. 2012. "Quantising Opinions for Political Tweets Analysis." LREC 3901-3906.

Huberman, Bernardo A., Daniel M. Romero, and Fang Wu. 2008. "Social networks that matter: Twitter under the microscope." *arXiv preprint arXiv:0812.1045*.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. "Why we twitter: understanding microblogging usage and communities." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* 56-65. ACM.

Johnson, Christopher, Parul Shukla, and Shilpa Shukla. "On Classifying the Political Sentiment of Tweets." *cs. utexas. edu*.

Jungherr, Andreas, Pascal Jürgens, and Harald Schoen. 2012. "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment"." *Social Science Computer Review* 302:229-234.

Kairam, Sanjay R., Meredith Ringel Morris, Jaime Teevan, Dan Liebling, and Susan Dumais. 2013. "Towards supporting search over trending events with social media." *Proceedings of ICWSM 2013, the 7th International AAAI Conference on Weblogs and Social Media*.

Kıcıman, Emre. 2012, May. "OMG, I have to tweet that! A study of factors that influence tweet rates." In *Sixth International AAAI Conference on Weblogs and Social Media*.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. "Private traits and attributes are predictable from digital records of human behavior." *Proceedings of the National Academy of Sciences* 110.15 (2013): 5802-5805.

Lee, Jisue, Hohyon Ryu, Lorri Mon, and Sung Jae Park. 2013. "Citizens' use of Twitter in political information sharing in South Korea." *iConference* 351-365

Li, Xiao, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In Sigir '08: proceedings of the 31st annual international acm sigir conference on research and development in information retrieval, 339--346, New York, NY, USA, 2008. , ACM.

Lui, Catherine, Panagiotis T. Metaxas, and Eni Mustafaraj. 2011. "On the predictability of the US elections through search volume activity." In *Proceedings of the IADIS International Conference on e-Society*.

Metaxas, Panagiotis T., and Eni Mustafaraj. 2012. "Social media and the elections." *Science*, *338*6106 472-473.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ACM.

Moe, Hallvard, and A. Larsson. 2012. "Methodological and ethical challenges associated with large–scale analyses of online political communication." *Nordicom Review*, 331:117-124.

Monti, Corrado, Matteo Zignani, Alessandro Rozza, Adam Arvidsson, Giovanni Zappella, and Elanor Colleoni. 2013. "Modelling political disaffection from Twitter data." In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* 3. ACM.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose." *Proceedings of ICWSM*.

Mustafaraj, Eni, and Panagiotis Takis Metaxas. 2011. "What Edited Retweets Reveal about Online Political Discourse." In *Analyzing Microtext*.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM*, 11:122-129.

M Pennacchiotti and A Popescu. 2011. "A machine learning approach to twitter user classification." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ACM.

Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. "Truthy: mapping the spread of astroturf in microblog streams." In *Proceedings of the 20th international conference companion on World wide web* 249-252. ACM.

Shi, Lei, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, and Jacob Spoelstra. 2012. "Predicting US primary elections with Twitter." In *Proceedings of Social Network and Social Media Analysis: Methods, Models and Applications NIPS Workshop, Lake Tahoe, NV, December* Vol. 7.

Skoric, Marko, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim, and Jing Jiang. 2012, January. "Tweets and votes: A study of the 2011 Singapore general election." In *System Science HICSS, 2012 45th Hawaii International Conference on* 2583-2591. IEEE.

Soler, Juan M., Fernando Cuartero, and Manuel Roblizo. 2012. "Twitter as a Tool for Predicting Elections Results." In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining ASONAM 2012* 1194-1200. IEEE Computer Society.

Tang, Cong, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. "What's in a name: a study of names, gender inference, and gender behavior in facebook." In *Database Systems for Adanced Applications* pp. 344-356. Springer Berlin Heidelberg.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10:178-185.

Vallina-Rodriguez, Narseo, Salvatore Scellato, Hamed Haddadi, Carl Forsell, Jon Crowcroft, and Cecilia Mascolo. 2012. "Los twindignados: The rise of the indignados movement on twitter." In *Privacy, Security, Risk and Trust PASSAT, 2012 International Conference on and 2012 International Confernece on Social Computing SocialCom* 496-501. IEEE.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2013. "Forecasting Elections with Non-Representative Polls." *Submitted to International Journal of Forecasting*.

Wu, Shaomei, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. "Who says what to whom on twitter." In *Proceedings of the 20th international conference on World wide web* 705-714. ACM.

Yom-Tov, Elad, and Fernando Diaz. 2013. "The effect of social and physical detachment on information need." *ACM Transactions on Information Systems (TOIS)*, *31* 1:4

F Al Zamal, W Liu, D Ruths. 2012. "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors." In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ACM.

Zhao, Xin, and Jing Jiang. 2011. "An empirical comparison of topics in twitter and traditional media." *Technical Paper Series, Singapore Management University School of Information Systems*.

Appendix

We use the tweet author's first name to impute gender, based on gender distributions gathered from annual social security reports of births from the years 1880 to 2012.[3] This dataset provides us with gender probabilities for 91,320 names, with almost all names clearly indicating one gender or the other (91.7% of names are single gender with 99.0% probability). We leave the values as the percentage likelihood of being male rather than making a binary choice on gender. This dataset provides 58% coverage over Twitter user profiles and 54% coverage over messages. Twitter user names that are not covered by our technique include pseudonyms without first names, organizational names, and foreign names too uncommon in the US to be included in the list of SSN births; this is a common procedure (Tang et al., 2011).

We infer a tweet author's coarse-grained geographic location from their self-declared user profile (Kıcıman, 2012). While people often self-declare easily interpretable geographic locations, many use colloquial names ("deep south"), acronyms ("DMV" for DC-Maryland-Virginia) and terms that are unrelated to locations ("none of your business", "the universe"). To learn the physical locations of these self-declared location names, we collect 1 month of geo-located tweets---separate from our election-related tweets--- where individual tweets have been tagged with a specific latitude and longitude, in addition to the users' self-declared location names. From this, we are able to learn a mapping from self-declared location names to latitude-longitude ranges. We extend this learned mapping with a list of US cities. For our analyses in this paper, we translate user locations to US regions and divisions. International, ambiguous, and unrecognized locations are labeled separately. The location is determined in 48% of profiles and that account for 56% of tweets; we ignore the 21% of geography identified profiles are international, but they comprise just 13% of geography identified tweets.

To determine the positive or negative sentiment of each tweet, we use a machine learned affect and mood classifier (De Choudhury et al. 2012a, 2012b). This classifier measures the mood of the author, rather than their opinion about an entity – a much harder and unreliable task given sarcasm and rapid evolution of topic-specific pejoratives. The sentiment model we use is based on the research described in Choudhury et al. (2012a, 2012b). 203 different mood-

---

[3] The Social Security Administration makes these available at URL

indicating hashtags are mapped to a valence (V) and activation (A) score, based on the psychological literature and a crowd-sourced questionnaire. V and A range from 1-10. From the valence and activation score of each mood hashtag we computed a compound score V*A, normalized to a range from -1 to 1. We then trained a boosted decision tree regression model to regress on the compound score. We used 2.75 million tweets that contained one of the 203 mood hashtags at the end of the tweet for training. Each tweet is represented by a vector of unigram and bigram features. After eliminating features that occur fewer than 50 times in our corpus, we are left with 344,000 features. The resulting model achieves a Root Mean Squared Error (RMS) of 0.41, with a random baseline RMS of 0.77.

Tables

	Twitter	Search	Voters
Obama	52%	88%	52%
Romney	36%	11%	48%
Both	12%	0%	0%
Male	64%	53%	46%
Female	36%	47%	54%
New England	5%	4%	5%
Mid-Atlantic*	20%	22%	15%
East North Central	13%	14%	16%
West North Central	8%	5%	8%
South-Atlantic*	17%	21%	18%
East South Central	5%	4%	6%
West South Central	11%	10%	10%
Mountain	6%	5%	7%
Pacific	15%	14%	14%
URL?	40%	----	----
Sample Size	----	----	129m

	Search	Voters			Search	Voters
$0-$15k	15%	6%	**0-17**		9%	0%
$15k-$25k	11%	6%	**18-24**		28%	9%
$25k-$40k	12%	13%	**25-34**		20%	14%
$40k-$60k	15%	16%	**35-44**		15%	16%
$60k-$75k	23%	13%	**45-54**		14%	20%
$75k-$100k	12%	15%	**55-64**		9%	19%
$100k	11%	30%	**65+**		6%	22%

Table 1. Summary statistics of demographics associated with search and Twitter for the 2012 presidential election from August 1, 2012 through November 6, 2012 (Election Day)

Note: The gender, income, age of the voting population is from 2012 Current Population Survey. For income categories that do not perfectly align we assume a constant distribution within a category. The geographical divisions and support represent the true percentages from the 2012 presidential vote totals, with the support shown as percentage of support for each candidate among the two major party candidates. The geographical divisions correspond to the nine official Census Bureau divisions of the United States. We make one small change to divisions to make them more meaningful for politics, which is to shift DC, DE, and MD from South Atlantic to Mid-Atlantic. Demographics are counted by engagement, not be respondent, so a person's demographics are counted as many times as they engage.
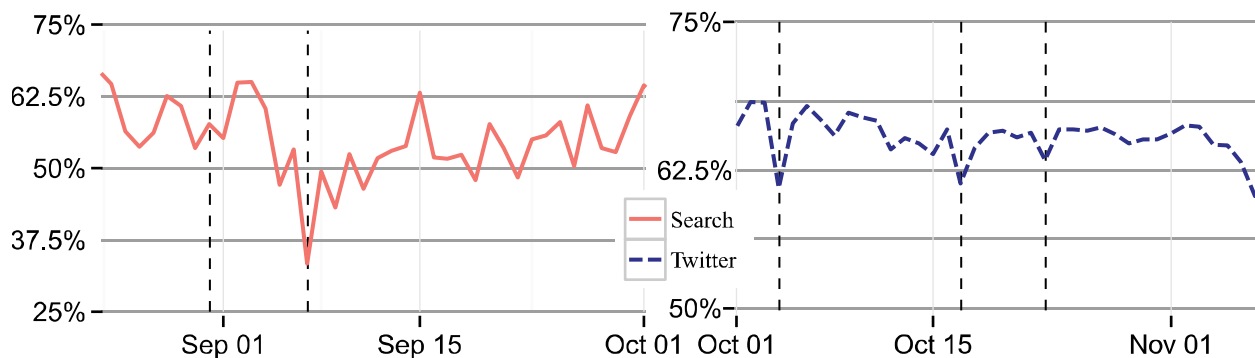
	Percent All Obama	Percent All Romney	Mixed
1 Tweet	54%	46%	0%
2 Tweets	34%	31%	34%
3 Tweets	23%	22%	55%
4 Tweets	27%	16%	66%
5+ Tweets	9%	8%	83%

Table 2. Percent of tweets, which mention only one candidate, on any given day by the same user, from August 1, 2012 through November 6, 2012 (Election Day), that reference Obama
Note: For example, if a user made 5 tweets in a day and 3 mentioned Obama, than the user would be in the 60% bucket.

Figures

Figure 1. Percent of overall search and Twitter discussion about the presidential candidates from August 1, 2012 through November 6, 2012 (Election Day) on search and Twitter, complied daily
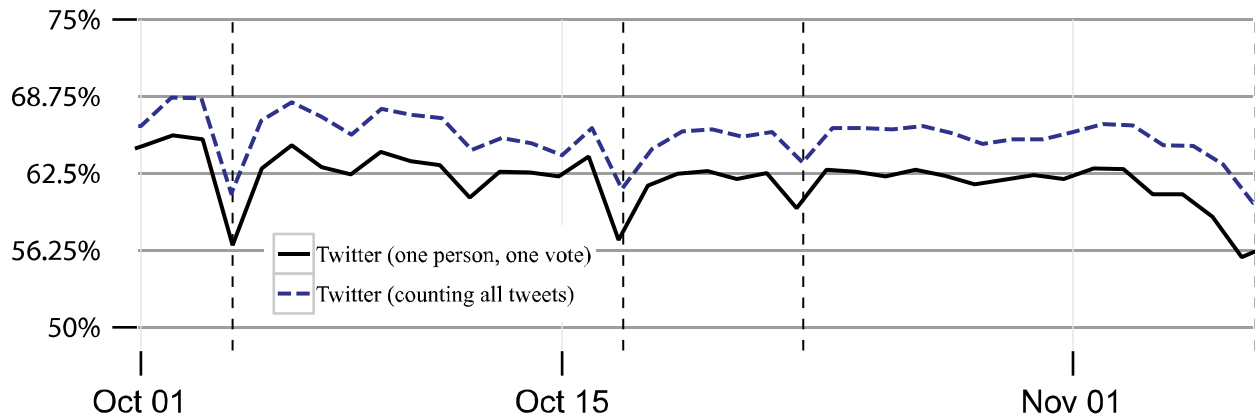*Note:* Each line combines any text that has Obama, Romney, or both on any of the two mediums. The charts sum up the total discussion about all three of the entities combined on any given day and divides by the total discussion. The vertical dotted lines in this figure (and almost all other figures) represent the following major events (in order): the Republican National Convention, the Democratic National Convention, and the three presidential debates.

Figure 2. Percent of search and Twitter discussion about the presidential candidates conducted by males from August 15, 2012 through October 1 (search), and from October 1, 2012 through November 6, 2012 (Election Day) (Twitter), complied daily
Note: Each line combines any text that has Obama, Romney, or both on any of the two mediums.

26

Figure 3. Percent of Twitter discussion about the presidential candidates conducted by geographical region from August 1, 2012 through November 6, 2012 (Election Day), complied daily

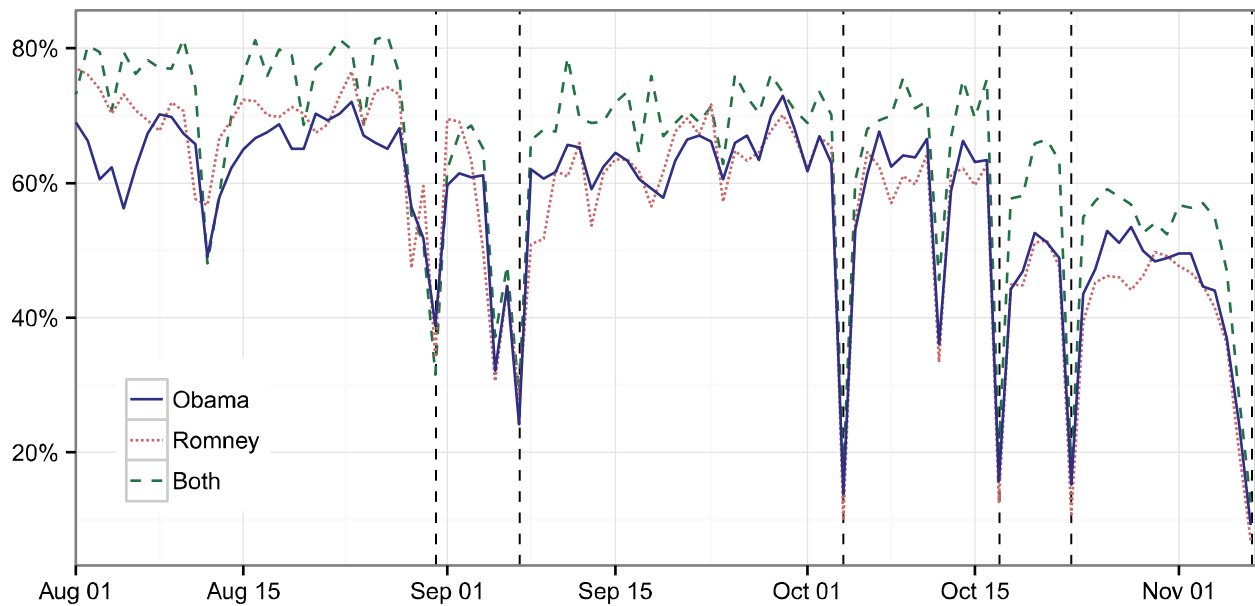*Note:* Each line combines any text that has Obama, Romney, or both on any of the two mediums.

Figure 4. Average number of tweets about the presidential candidates per user from August 1, 2012 through November 6, 2012 (Election Day), complied daily
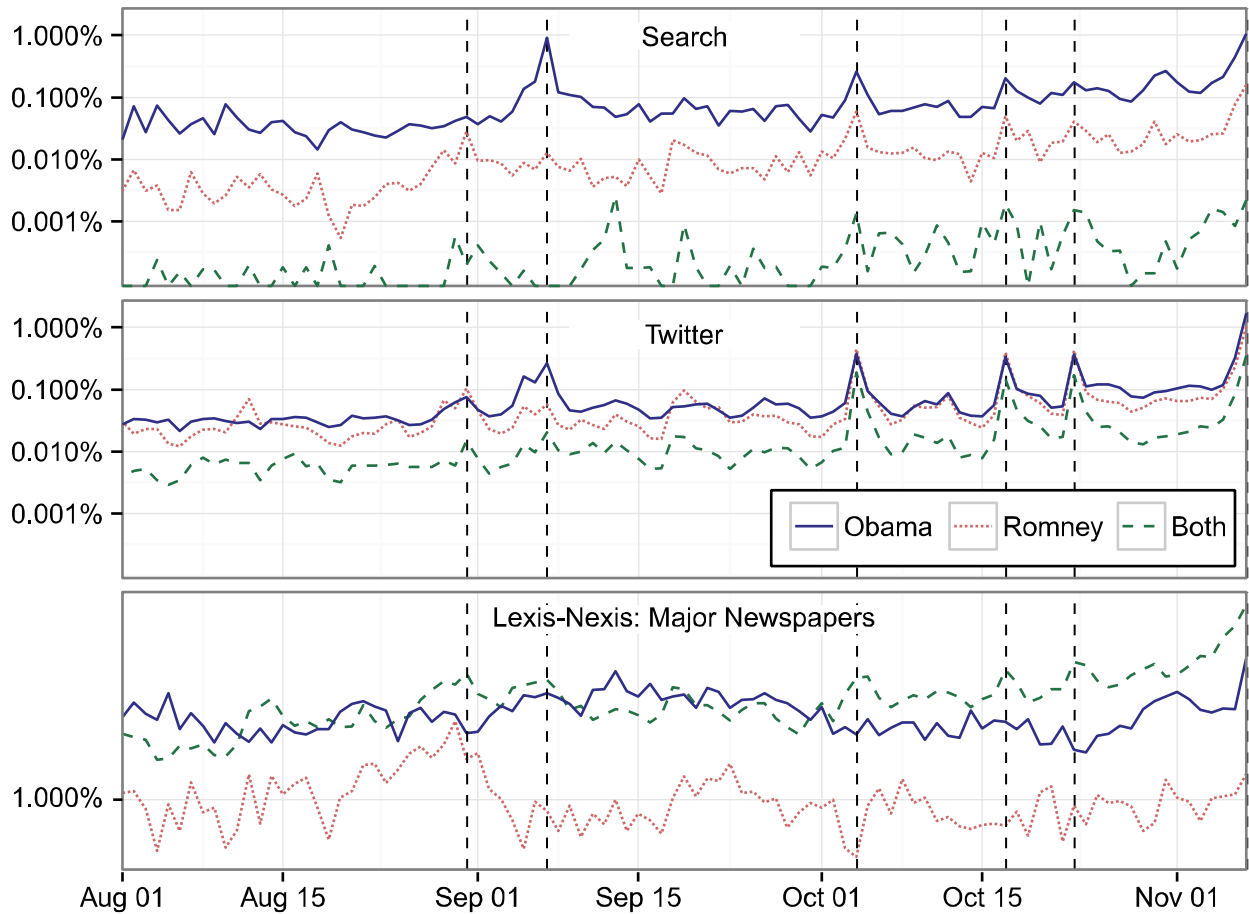
Note: Each line combines any text that has Obama, Romney, or both on any of the two mediums.

Figure 5. Percent of Twitter discussion about the presidential candidates conducted by males from October 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Each line combines any text that has Obama, Romney, or both on any of the two mediums. The top line is the same line as the right line from Figure 2.
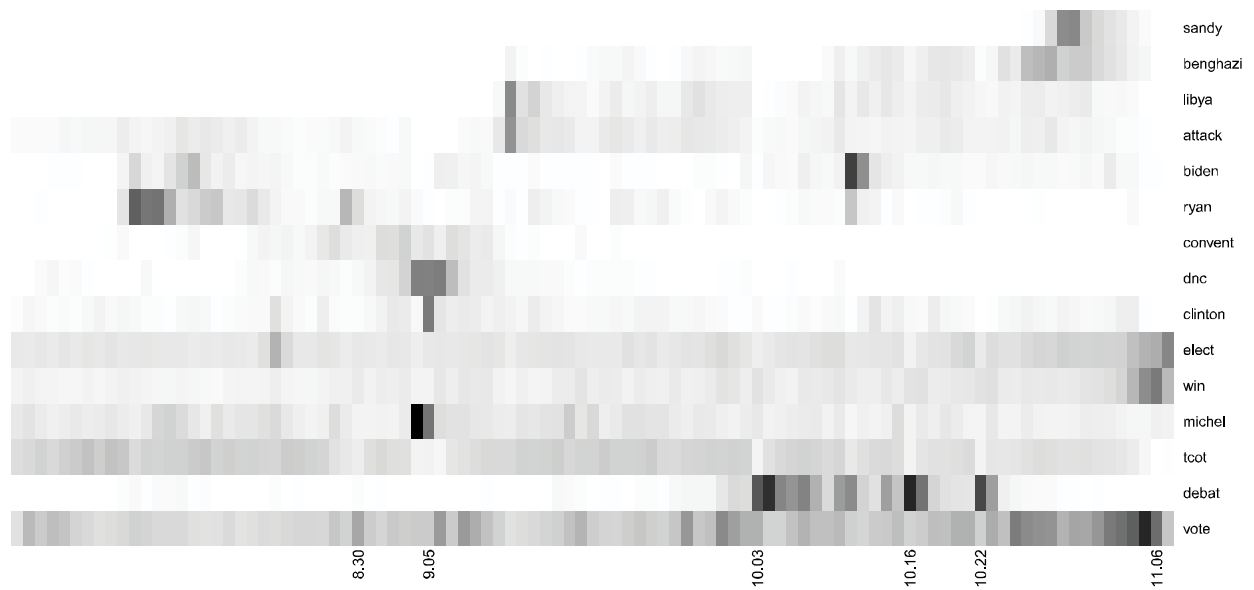

Figure 6. Percent of Twitter discussion about Obama, Romney, and both that contain a URL from August 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Each line shows the percentage of tweets with a URL about all three of the entities separately on any given day.

Figure 7. Percent of overall search, Twitter, and print media discussion about the Obama, Romney, and both from August 1, 2012 through November 6, 2012 (Election Day), complied daily

Note: Each line sums up the total discussion about all three of the entities separately on any given day and divides by the total discussion.
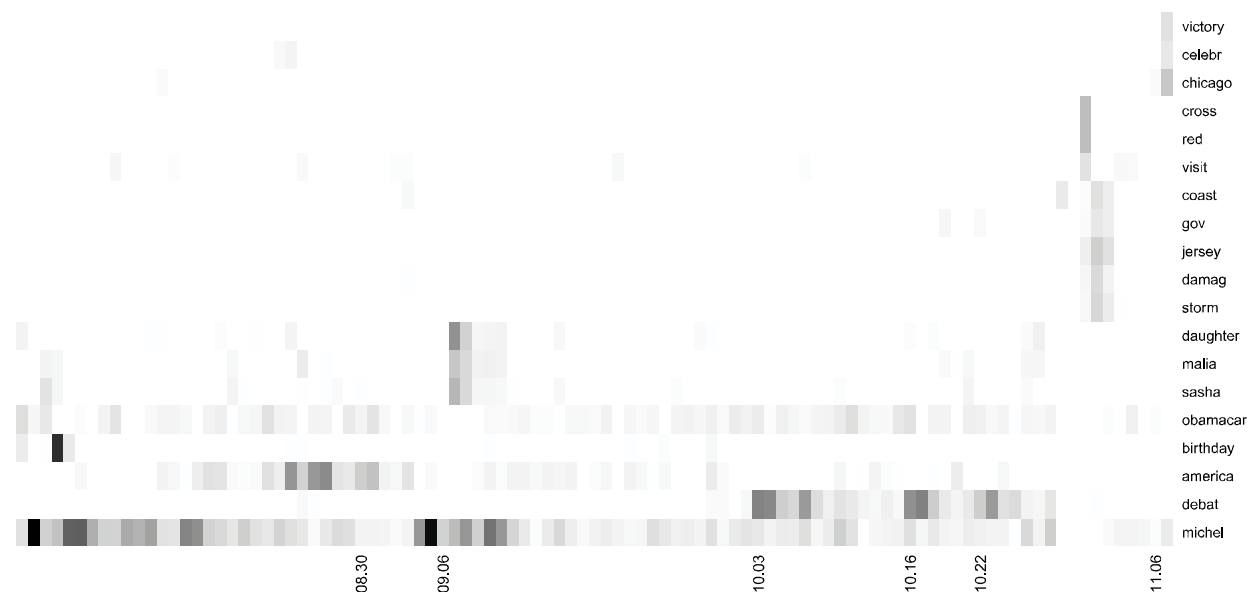
Figure 8. Top terms co-occurring with "Obama" on Twitter, from August 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Term rows are arranged by time series clustering.

Figure 9. Median days since the last election-related tweet by the same person for any given tweet on that day
Note: October 3 is actually infinity in that the median tweet is done by a person who did not tweet about the election in our sample.
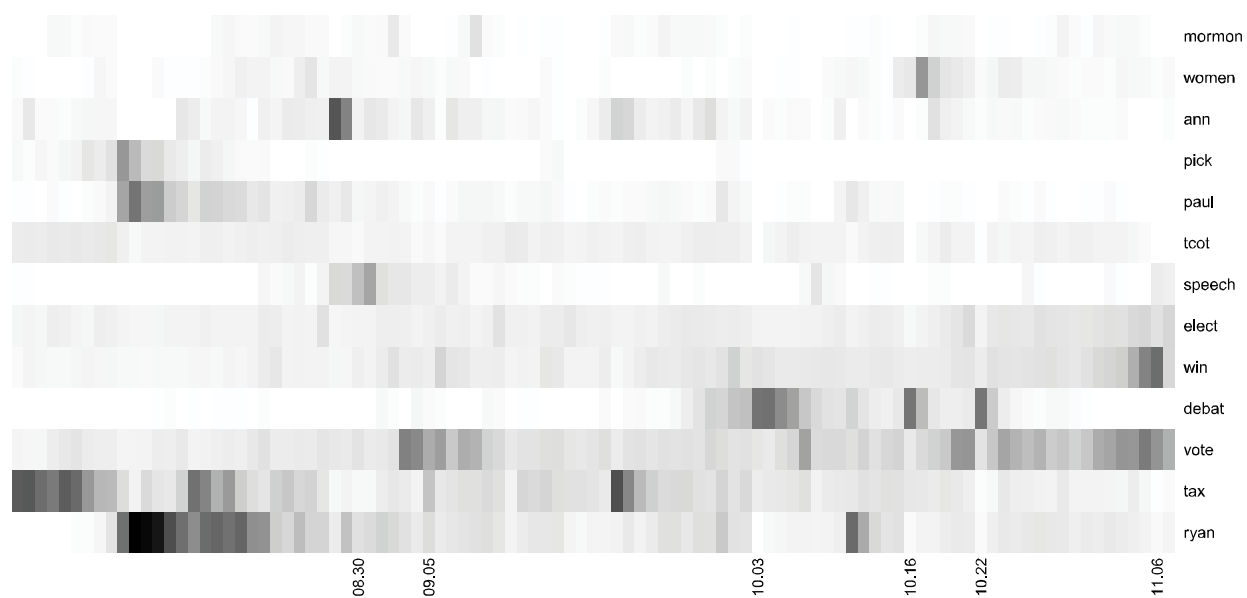
30

Appendix Figures

Figure A1. Top terms co-occurring with "Obama" on search, from August 1, 2012 through November 6, 2012 (Election Day), complied daily
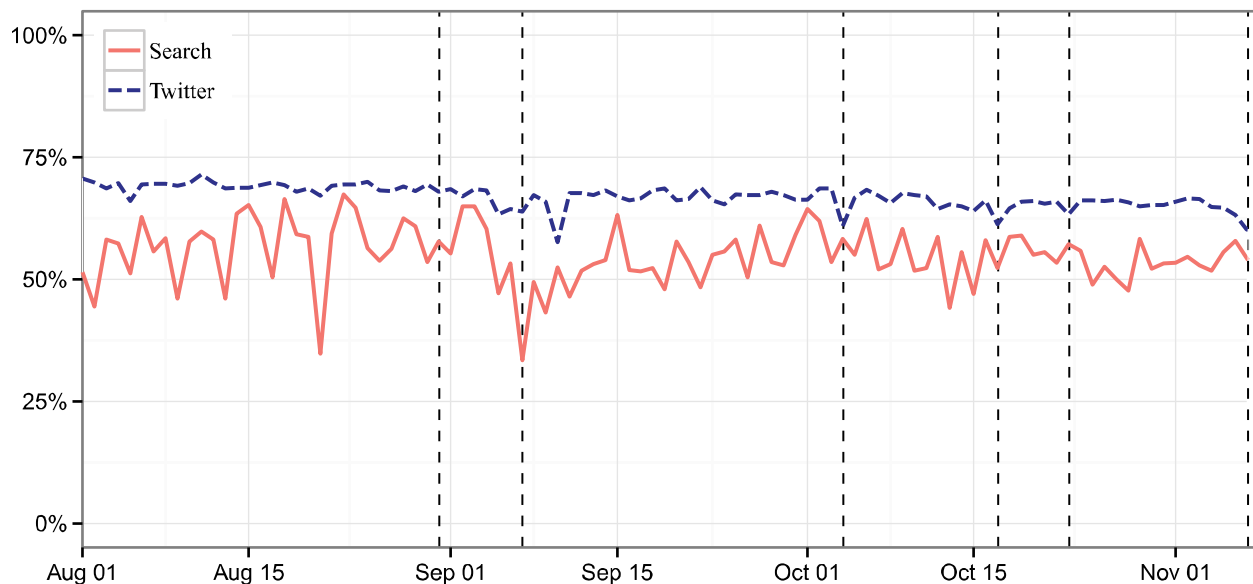*Note:* Term rows are arranged by time series clustering.

Figure A2. Top terms co-occurring with "Romney" on Twitter, from August 1, 2012 through November 6, 2012 (Election Day), complied daily
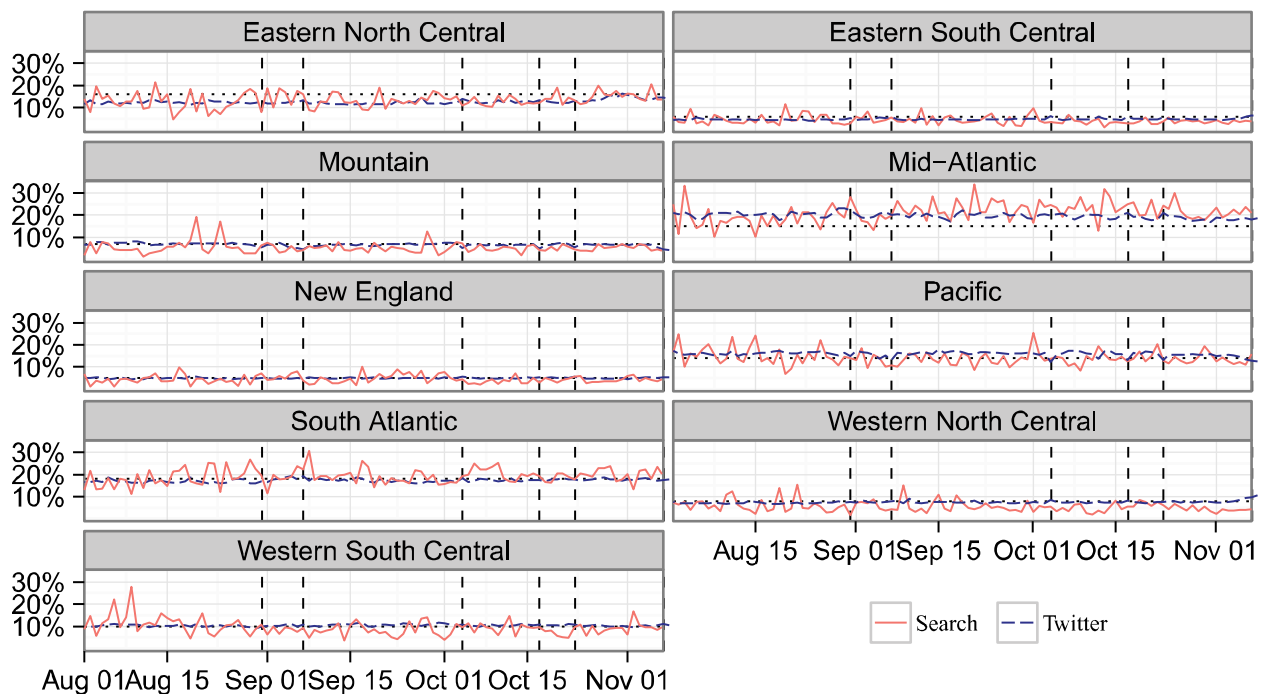*Note:* Term rows are arranged by time series clustering.

Figure A3. Top terms co-occurring with "Romney" on search, from August 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Term rows are arranged by time series clustering.

Online Appendix Figures

Figure OA1. Percent of search and Twitter discussion about the presidential candidates conducted by males from August 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Each line combines any text that has Obama, Romney, or both on any of the two mediums.

Figure OA2. Percent of search and Twitter discussion about the presidential candidates conducted by geographical region from August 1, 2012 through November 6, 2012 (Election Day), complied daily
Note: Each line combines any text that has Obama, Romney, or both on any of the two mediums.