Twitter Zombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere

Alan Black
Drexel University
Philadelphia, PA
aeblack@gmail.com

Christopher Mascaro
Drexel University
Philadelphia, PA
cmascaro@gmail.com

Michael Gallagher
Drexel University
Philadelphia, PA
michael.gallagher24@gmail.com

Sean P. Goggins
Drexel University
Philadelphia, PA
outdoors@acm.org

ABSTRACT

Social computational systems emerge in the wild on popular social networking sites like Facebook and Twitter, but there remains confusion about the relationship between social interactions and the technical traces of interaction left behind through use. Twitter interactions and social experience are particularly challenging to make sense of because of the wide range of tools used to access Twitter (text message, website, iPhone, TweetDeck and others), and the emergent set of practices for annotating message context (hashtags, reply to's and direct messaging). Further, Twitter is used as a back channel of communication in a wide range of contexts, ranging from disaster relief to watching television. Our study examines Twitter as a transport protocol that is used differently in different socio-technical contexts, and presents an analysis of how researchers might begin to approach studies of Twitter interactions with a more reflexive stance toward the application programming interfaces (APIs) Twitter provides. We conduct a careful review of existing literature examining socio-technical phenomena on Twitter, revealing a collective inconsistency in the description of data gathering and analysis methods. In this paper, we present a candidate architecture and methodological approach for examining specific parts of the Twittersphere. Our contribution begins a discussion among social media researchers on the topic of how to systematically and consistently make sense of the social phenomena that emerge through Twitter. This work supports the comparative analysis of Twitter studies and the development of social media theories.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Web-based Interaction

General Terms

Algorithms, Design, Experimentation, Standardization.

Keywords

Twitter, data collection, data management, methods, social media

1. INTRODUCTION

There are recognized empirical and theoretical gaps in the application of social science theories to raw, electronic trace data like that retrievable from Twitter [23]; a record of interaction through technology does not necessarily act as a proxy for social interaction. Such gaps are further exacerbated by the opaqueness of Twitter's API for retrieving data, and different choices researchers make about how to retrieve, store and analyze data from Twitter. Each peer reviewed study of Twitter interactions, in domains ranging from disaster relief, to sports viewing, political action and celebrity engagement with fans, presents an explanation of how data is captured, analyzed and related to findings within the individual study. The explicitness of these descriptions is variable. Consequently, comparisons across Twitter studies and aggregation of findings leading to more comprehensive theories of social media interaction are difficult because of the lack of a shared view of well understood and documented methods for gathering and analyzing electronic traces from social media, including Twitter.

A few papers have attempted to build community understanding of how to select and use different Twitter application programming interfaces (APIs) for research. Zhao et al. [50] , for example, contrast the three main APIs provided by Twitter – the REST, Search and Streaming APIs – but their paper is either out of date a year after publication, or incorrect in its interpretation of Twitter's API's.

Scholarship that references, documents or contrasts different social media platform APIs, including Twitter's, face the challenge of working to reverse engineer a system whose traits may be shifting over time. What we report on today with regards to the Twitter API may not be sustained over time by the Twitter platform. The two key gaps in analysis of social media generally, and Twitter in particular, then, are 1) Each study constructs it's own approach to gathering Twitter data and 2) Attempts to explain the Twitter API through analysis are difficult to verify because the data delivered by API may be changing over time.

We see three potential ways of overcoming the challenges we identify. First, social media vendors could make the completeness of data retrieved through their API's more transparent. This is stifled by privacy and competiveness concerns. Second, individual studies of Twitter may begin to follow a standard methodology for gathering data, appropriate to the problem context, and referencing articles focused on these methodological approaches. Third, the community might consider a standard architecture for the capture of social media, which would constitute a technical architecture to ensure consistency of social science results; in a sense, using computers for what they are good at, and people for what they are best at.

Fully developed solutions to these challenges are an important, long-range goal for the social media research community. In this paper we present a modest, but integrated methodological approach and technology architecture for the standard capture, social transformation and analysis of Twitter interactions using the Search API. We contrast this API with the other two, describe the results of experiments conducted using our tools, which we call Twitter Zombie, explain our process of social transformation, and describe a pilot visualization and analysis project using this methodological approach and toolset. We conclude with a road map for methodological enhancements to social media research.

2. PRIOR TWITTER WORK

2.1 Twitter and its Affordances

Twitter research has evolved from the time when Twitter was first introduced in 2006. Early research on Twitter attempted to characterize user behavior in the technology. Java et al. [26] found that individuals used Twitter to discuss daily routines and to exchange news. At the same time, other researchers attempted to characterize user behavior on Twitter and identify specific behavior around Twitter's numerous affordances [29].

The three most common affordances in twitter are the hashtag, retweet and @-mention. Hashtags are used to highlight streams of discourse for others to attend to [24; 42] and retweets are a mechanism of forwarding another user's message in one's own Twitter stream [8]. In addition to the inclusion of hashtags and retweets, the inclusion of @-mentions (@ followed by a username) signifies a direct addressal to or highlighting of a message to someone else and may be indicative of targeted information sharing or discourse [22].

In addition to these technological affordances, researchers are able to collect the device or application that a user utilized to send a tweet. Analysis of the device activity can highlight the utilization of different technological access mechanisms for different purposes [48]. We illustrate such differences in the description of our pilot study in section five. When coupled with geographic location from the user profile, or from where the tweet originated, researchers can identify geographically specific information such as localized discourse [39] or attempt to identify the overall happiness of people in certain geographic areas [41].

Unlike other forms of social media, only 22% of Twitter relationships are reciprocal [30]. This creates an environment of "context collapse" in which a user has multiple audiences for their tweets, and the user may not be aware of who is in those audiences [33]. As a result of this asymmetric network structure, information diffusion is significantly different on Twitter than in social networking platforms that have symmetric relationships [31]. Substantial prior research utilizes the follower/following relationships to characterize user behavior on Twitter [28].

Measuring the number of followers an individual has (as in-degree influence) illustrates popularity, but does little to measure their ability to influence others [9]. Though follower/following relationships are important for understanding initial information diffusion, collecting and characterizing actual user behavior such as retweet behavior, the number of mentions and reply-to's, and the content of tweets are much better indicators of influence [3; 6]. Influence in Twitter is also shown to be the result of long-term reputation building in a network of individuals [9].

One of the more commonly used affordances in Twitter is the retweet. A retweet is the process of an individual further propagating another's message by copying it into their Twitter stream. Prior studies illustrate that 16% of tweets on a daily basis on Twitter are retweets [38]. In specific domains, such as political discourse, the percentage of retweets can be as high as 56%, depending on the topic [38]. Kwak et al. [30] found that 75% of retweets occur within the first hour of the original tweet, but that 10% of retweets occur a month later. This illustrates a different intent and purpose of retweeting and a different trajectory of information diffusion. There are many reasons that individuals retweet messages, which include: to propagate information, to illustrate that they are "present" in the conversation or in the space, and to attempt to return favors to other individuals to prop up their twitter followers [8]. In addition to these reasons, retweets are seen as a mechanism of conversation that takes on different characteristics depending on the user's network and the content of the original tweet. One study found that positive messages are more likely to be retweeted than negative messages, illustrating how content can affect information diffusion through the network through the retweet mechanism [17].

Honeycutt and Herring [22] identify the technological affordance of @, which directs a message towards someone else, as a form of addressivity [46] in Twitter. In their analysis, they found that close to 90% of the instances of @ were someone addressing another individual in a conversation and these conversations on average lasted 3-5 messages. Of the larger sample that included tweets with both @ and without, they found that those tweets with @ tended to be more interactive in their content. For example, they found that messages that employed the @-mention affordance received a response 31% of the time, which is higher than previous studies of technologically mediated communication. This number is also a conservative estimate, as it does not take into account the possibility that a reply may have been sent in another channel [49].

2.2 Applied Twitter Analysis

One of the most significant areas of research on Twitter has been the use of sentiment analysis and other modeling techniques to examine, explain, or predict offline events [2; 4]. Some research has indicated that the brevity of Twitter messages affords more reliable sentiment classifications [5]. These approaches have been applied to predicting the direction of the US Stock market [7], analyzing debate performance in a 2008 US Presidential Debate [13] understanding the outcome of 2011 Portuguese Presidential elections [14] and identifying general public opinion [1]. Further research indicates that the volume of Twitter activity may mirror box office performance, but may not be as representative of stock market activity [34]. These findings indicate possible different uses of Twitter in different social domains, and also may represent a demographic difference in user activity. Understanding sentiment on Twitter has also been used to further understand sporting events such as the Olympics [17] and Brazilian Soccer Leagues [18].

The medical community has also studied Twitter as a way to understand whether or not the public adopts certain terminology in the context of a Pandemic [10]. In this instance, the researchers were interested in whether the public used the term "swine flu" or the more medically formal term, "H1N1." Early research analyzing medical information diffusion on Twitter has also attempted to identify the mechanisms that users utilize to judge trust and validity of medical information. This research has demonstrated that the originating user and the content of the message is likely to be a significant factor in how individuals assess the validity of information. [36]. Researchers have also attempted to identify influenza outbreaks through the monitoring of Twitter for certain keywords, but have had limited success [12].

One of the greatest areas of research on Twitter is the analysis of political activity and participation in Twitter. Researchers identify partisan clusters in retweet behavior illustrative of echo chambers of ideas in information diffusion in Twitter [11]. These researchers also find examples of "content injection" that identifies users adopting partisan hashtags or keywords to broadcast material that may be counter to the ideology of the party to proliferate a message. Similar activity is also been noted in the context of the conservative hashtag #tcot (Top Conservatives on Twitter) [32]. Additionally, researchers show how multi-

dimensional scaling can be used to classify users based on hashtag and @-mention usage [19].

Researchers in the U.K. found a difference in adoption of Twitter based on partisan affiliation of members of parliament (Williamson et al 2010). Additionally, many of the studies focused on the utilization of Twitter by the US Congress found that members used Twitter for self-promotion as opposed to communicating directly and specifically to citizens [16]. Analysis of political activity in Korea has found that "resource-deficient politicians" may be more likely to engage with followers and use it as a mode of connecting to citizens [28]. Research has identified similar behavior in the United in the 2010 US midterm election where the conservative minority was more effective at using social media to build support [32] and challengers tended to interact more with the public than incumbents [40].

Using hashtags to analyze political discourse on Twitter has been done across cultures as well. German researchers used politically oriented hashtags to identify 2009 election discourse [27]. During this election, German Twitter users were encouraged to use party related hashtags followed by + or – to illustrate agreement or disagreement with the message. Through this hashtag valence the researchers were better able to understand the network structure of individuals and identify "small worlds" of connected individuals had similar political viewpoints.

In addition to understanding how politicians, candidates and the public use Twitter in the political context, there has been a significant amount of work that has attempted to illustrate how social media may be able to predict elections [45; 47]. A review of this work illustrates fundamental flaws in the approaches and illustrates the lack of comparison to traditional mechanisms of prediction and analysis such as polling or historical evidence, illustrating that the incumbent wins close to 90% of the time in United States Congressional election [35]. Metaxas et al. further extend this critique by identifying that much of this "prediction" occurs after the election and may actually be worse than traditional models. When attempting to repeat experiments that "predicted" wins in electoral races, Metaxas et al. [35] were unable to reproduce the results.

The contrast between planned and unplanned events is one that has been explored in the context of crisis informatics on social media such as Twitter. Research that compared national political events and natural disasters have found that Twitter is used as a way to broadcast information out to the public and in the case of natural disasters Twitter is used as a frequent way to share links with the public [25]. Twitter has also been instrumental in understanding crisis events and natural disasters. One of the reasons that individuals use Twitter during a crisis event is to relay information from the place where the activity is happening and also to synthesize current information to proliferate it through the network of individuals [43]. Researchers have studied the use of Twitter for information diffusion and sense making related to unplanned, social and violent events like school shootings [21]. Research on the 2007 wildfires in California illustrated how social media – Twitter in particular – can be an important source of information for citizens and described how broadcast media turned to Twitter to get information about what was happening [44].

2.3 Extending our Collection Knowledge

Our survey of the existing literature that we discuss above reveals a significant variation in how individuals collect Twitter data. Currently, Twitter provides three API's to collect data, with the two most popular being the Search API and Streaming API. Our review of the literature shows that studies that look at specific affordances such as hashtags, @-mentions or other keywords contained within tweets tended to use the Search API to access data about specific events or topics [2; 8; 13; 17; 21; 30; 42], while other studies that attempt to look at longitudinal opinions of movies, politics and other domains use the Streaming API to access data [6; 19; 34]. The differences between how data is collected by these two API's may significantly alter the type of dataset collected by researchers, though these differences are not discussed in detail in empirical studies of Twitter.

In addition to lab developed access mechanisms that query the Twitter API directly, there are other tools such as NodeXL [20] that provide an interface for users to access data, and statistical libraries such as twitteR for the statistical program R [15]. In addition to these access mechanisms, Twitter also provides feeds of the Twitter stream to some organizations. These feeds are described by Twitter as a random sample of a percentage of the overall Twitter stream. We found only one explicit mention to this access mechanism in our survey of the literature and that was the Twitter "garden hose" which provides a sample of 10% of all tweets. This access was used by the individuals of the Truthy project at Indiana for a series of papers [11; 35; 37; 38]

Our review of the literature illustrates the domain breadth, affordance diversity and methodological approach differences associated with prior Twitter research. We show that social media research in general, and Twitter research specifically presents with a diverse set of approaches for gathering and analyzing socio-technical phenomena that share Twitter as a social media platform. Twitter literature to date illustrates that there is not a consistent, repeatable set of tools for collecting, analyzing and reporting on Twitter facilitated social groups. Further, different studies present and explain their methods of capture and analysis with an inconsistent level of clarity and specificity. These gaps make it difficult to draw comparisons across studies of similar phenomena in Twitter, and impair the development of broader social media theories.

3. TWITTER ZOMBIE

3.1 Twitter Data Collection Facilities

The contribution we make to address the challenges presented is a methodological approach and technical tool (Twitter Zombie) for Twitter data collection and analysis. Our Twitter Zombie system for capturing data from Twitter and the associated experiments we present provide a repeatable foundation for the community to use for social media capture verification and our pilot study illustrating our methodological approach illustrates the collection idiosyncrasies associated with certain collection parameters, and how different parameters can alter the collected datasets. If played out over time and across studies, these small differences may be significant and alter findings. Presently, as we noted, such differences are seldom surfaced in empirical studies of social phenomena on Twitter. We invite other researchers to share their experiences with their specific systems as an important methodological step in addressing the opaque nature of the Twitter API structure.

Twitter offers three primary methods for allowing software developers access to Twitter data: the Streaming API, the REST (Representational State Transfer) API and the Search API. The Streaming API relies upon a continuously open network connection between Twitter and the receiving host and is designed to support significant volumes of data transfer. By contrast, the REST API follows a typical client-server request and response

communication pattern where connections between Twitter and the requesting host are dynamically created on a per-request basis. Both APIs return data in JSON (JavaScript Object Notation) format, a compact human-readable data interchange format akin to an XML document representation, though less verbose.

Twitter Zombie utilizes the third publically available API, known as the Twitter Search API. The Search API employs a REST communications pattern and provides a mechanism to query the real-time index of tweets. The index contains tweets that are six or fewer days old and may include tweets up to nine days old. In addition to temporal limitations, the search API imposes a number of important performance constraints. First, a query request can be rejected if it is too complex, although complexity is not publically defined. Also, results from the Search API are rate limited. Unfortunately, the parameters related to these limitations are unpublished. Finally, queries submitted via the search API are limited to a maximum of the 1,500 results, which may contain less than the most recent six days of tweets depending on how prolific the user is.

Selecting an API is an important decision for researchers, but one that is often not specified in empirical work and not rationalized in the face of research questions as illustrated in our previous review of existing literature. For this application, the Search API offers a number of advantages over the REST or Streaming APIs. The Search API does not impose explicit rate limits as does the REST API. Perhaps most importantly, batch use of the Search API allows Twitter Zombie to maintain distinct result sets from each search, even when a unique tweet is returned multiple times in response to different query strings.

3.2 Twitter Zombie Architecture

We access the Search API using a software system we developed in PHP, called Twitter Zombie. Data collected by Twitter Zombie is stored in a MySQL relational database management system. Twitter Zombie is designed to gather data from Twitter by executing a series of independent search jobs on a continual basis, 24 hours a day, 7 days a week. The execution interval for each search can be controlled independently through a rudimentary job management system. Each search job can be programmed to execute once every n minutes (where n >= 1) using a run interval value. This allows us to run searches for high volume queries (those returning many tweets) more frequently than those associated with low volume result sets. High volume queries are typically run every minute or two, while some low volume search jobs are scheduled to execute only once each day (or every 1440 minutes). The search job control system also allows us to stop, restart, and change execution intervals on the fly.

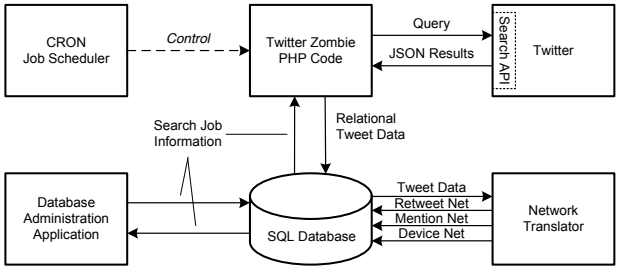

Figure 1 - Twitter Zombie System Diagram

The workflow for creating a new Twitter Zombie search job begins at Twitter's advanced search web page. This simple page provides a text input field for the user's query string as well as access to a handy pop-up reference that lists search operators along with usage examples. These operators include OR, the minus sign (-) for negation, the ampersand (@) for referencing people, the pound sign (#) for locating Twitter hash tags, "near" and "within" for geo-based searches, and others including attitudinal and temporal operators.

To develop a new search job, we begin by entering a query into the advanced search page and executing it. If no results are returned, this is generally an indication of a malformed query. If results are produced, they are inspected for face validity. This process is repeated until a suitable query is constructed and tested. At that point, the URL encoded search string is copied from our web browser's address field for later entry into the Twitter Zombie job control system. In effect, we are using Twitter facilities to help develop, pre-flight and encode our queries.

Once a new query string is successfully tested, it can be entered into Twitter Zombie's job table. This table contains one record for each search job (active or inactive). Each record represents a complete search job definition that includes a run interval value, a human readable job description, the ID of the last collected tweet for this search and the encoded query string. Storing the Twitter supplied ID of the most recently collected tweet allows Twitter Zombie to request only those tweets that have been created since the last time the search job was executed which greatly improves overall collection efficiency.

The Twitter Zombie PHP code is executed each minute by the Linux time-based job scheduler, cron. During each running of Twitter Zombie, the entire search job table is scanned to determine which queries should be executed. If a search job's run interval value indicates that it's time to execute the query, a request is made to Twitter. The Twitter Zombie control architecture can therefore be thought about as two loop constructs, one inside the other. The outer control loop responsible for repeatedly executing Twitter Zombie is managed by cron, while the inner search job loop is managed by the PHP code, which references the job table.

In addition to storing and manipulating search job control data, the MySQL database is also used to store the collected tweets and associated metadata. The database schema is optimized for insertion efficiency. This is important in the context of a system that may be called upon to handle sudden unanticipated surges in tweet volume. Unforeseeable surges are common in disaster relief scenarios or when a news story breaks that impacts a large number of Twitter users. In terms of a design tradeoff, we have chosen to optimize the efficiency of data writes at the expense of storage consumption by forgoing a more space conserving, fully normalized database schema.

Much of Twitter Zombie's utility as a research tool stems from its ability to capture the hierarchical relationships in the data returned by Twitter. The search results are run through a tool we call the "network translator" which performs post collection processing and records the results in the database. The complete tweet text and all related entities (e.g. hashtags and mentions) are stored separately. We explain this further in section 4.1, which describes our conceptualization of how to socially transform raw Twitter data to reflect interactions between people, and between people and artifacts. Preserving the data's original structure allows us to leverage the power of SQL (Structured Query Language) to perform post hoc data transformation in order to answer specific research questions.

3.3 Tweet Data

3.3.1 Character encoding and counting

Social media researchers make implicit or explicit choices about whether or not to include the full character set for non-English languages; or multi-byte languages like Arabic, Chinese or others at all. This is due to the way Twitter handles character encoding, and the subsequent handling of that encoding by common software tools. Twitter stores the text strings that comprise tweets and other data as UTF-8 encoded characters. This means that tweets may include a variety of characters not represented in the ASCII (American Standard Code for Information Interchange) encoding scheme. UTF-8 encoding allows Twitter to handle the entire Unicode character set, but this affordance comes at the cost of complexity. Because UTF-8 is a variable-width encoding scheme (where a single character may be represented by two or more bytes), visually counting characters does not necessarily reveal the number of bytes required to store a given string. This uncertainty is exacerbated by the fact that some words with accented characters can be encoded using more than one representation. In order to not disadvantage users of non-English characters, Twitter employs Unicode Normalization Form C[1] in order to compute character count. This reality has obvious implications for the Twitter Zombie database design. In order to ensure that the full text of a tweet is faithfully recorded, the field containing the tweet string must be able to store four bytes for each character for a total of 560 bytes (i.e. 140 characters * 4 bytes per Unicode code point).

In order to alleviate the need for all of our downstream analysis tools (and even some basic system utilities) to support UTF-8 encoding, Twitter Zombie is capable of performing transliteration. This process maps Unicode characters that cannot be represented in ASCII to a suitable character or character string substitute. For example the euro sign would be replaced with the string "EUR" when transliteration is enabled. In future versions of the tool, full support for UTF-8 will be developed. Our review of dozens of previous Twitter studies reveals no explicit mention of how multi-byte Tweets are handled.

3.3.2 Metadata

In addition to receiving the raw text of a tweet, Twitter provides a wealth of metadata that is captured by Twitter Zombie. This invaluable metadata includes the time and date of a tweet and the tweet language expressed as a two-letter code defined by the ISO 639-1 standard. Tweet search results also include a source field that names the application used to create each tweet. Some tweets (the vast *minority,* unfortunately) are returned with geo-location data expressed as a point in terms of longitude and latitude.

Entities such as hashtags, mentions, and URLs are returned as distinct elements within the JSON representation. Each entity is further described by metadata that identifies its exact location within the tweet text. The metadata indicates the beginning and ending character positions for each entity providing a simple mechanism to calculate entity length.

Finally, each tweet returned to Twitter Zombie carries information regarding the author (i.e. sender). A unique Twitter ID as well as a long and a short user name identifies the tweet's creator. Tweets that are directed to a particular Twitter user also contain ID and name data for the intended recipient.

[1] http://unicode.org/reports/tr15/#Norm_Forms

3.3.3 Duplicate tweets

Twitter employs processes to remove duplicate and near-duplicate tweets from search results. The duplication detection technique relies on the MinHash algorithm. A number of signatures are computed for each tweet. These signature sequences are only four bytes in length. A tweet is considered a duplicate if it shares a set of signatures with another tweet.

3.3.4 Result quality and relevance

Twitter filters the results delivered by both the Streaming and the Search APIs in order to exclude tweets that are deemed low quality. While the filtering algorithm is unpublished, and therefore, is likely to change without warning, Twitter does provide some insight into the filtering methodology. Frequent tweets that are considered repetitious are targeted for filtering. Twitter also filters tweets from suspended accounts and tweets that fail to meet other vaguely defined standards.

When working with the search API, the result set may have also been culled based upon relevance. Twitter returns only the most relevant tweets pertaining to the query based upon unpublished criteria. The relevance filtering process is not imposed on results returned from the Streaming API.

4. Experimental Results

We performed several experiments using Twitter Zombie in order to better understand the operational characteristics of the Twitter Search API. The search terms "sports" and "sex" were chosen as they represent high tweet volume subjects, each returning hundreds of tweets per minute during most hours of the day.

Table 1 - Experimental Job Parameters

Experiment	Collection interval (min.)	API Query
1	1	q=sex
1	2	q=sex
1	3	q=sex
2	1	q=sports
2	2	q=sports
2	4	q=sports
3	1	q=sex
3	1	q=sex

In experiment number one, the same query was performed with different collection time intervals. All three search jobs were started at the same time. The graph in Figure 2 shows that the number of tweets collected each hour tracks closely across the three search jobs over a 24-hour period.

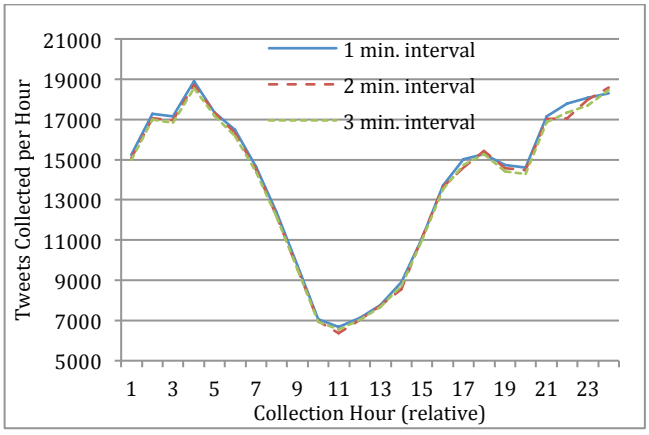

Figure 2 – Tweets per hour for q=sex at 1, 2, and 3 min. collection intervals

Despite the close agreement in hourly tweet counts, there appeared to be a slight but consistent drop-off in the number of tweets collected when using collection intervals longer than one minute (the baseline interval). Figure 3 shows the reduction in tweet counts for search jobs run at two and three minute intervals as a percentage of tweets collected with the same search performed each minute.

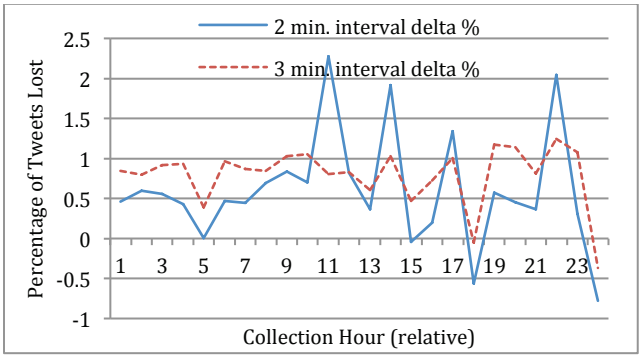

Figure 3 – Percentage drop-off for q=sex at 1, 2, and 3 min. collection intervals

As was true for experiment one, the number of tweets collected over a 24-hour period, this time using one, two and four minute collection intervals, tracked over time as shown in Figure 4.

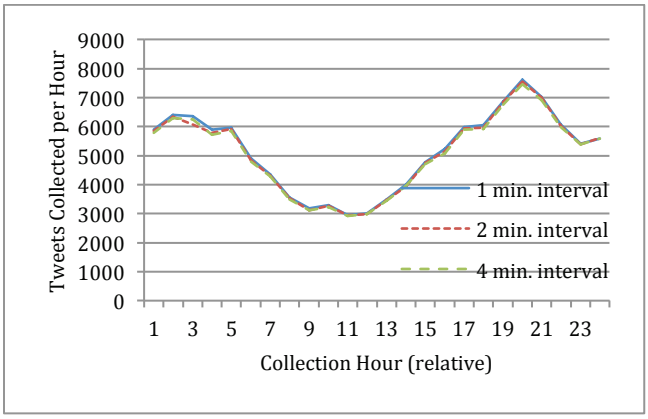

Figure 4 - Tweets per hour for q=sports at 1, 2, and 4 min. collection intervals

The pattern of recovering fewer tweets when using longer collection intervals appears again in experiment two. These results show the phenomenon more clearly than in experiment one due to the use of a longer maximum collection interval (4 min. versus 3 min. in experiment one).

While the mechanism or mechanisms responsible for reducing the number of tweets returned from Twitter when using longer collection intervals is unknown, one obvious potential source of the discrepancy is user deleted tweets. This highlights an important difference between the Search and Streaming APIs. The Streaming API provides tweet deletion messages that signal receiving software to discard previously recorded tweets. By contrast, the Search API provides no information regarding deleted tweets.

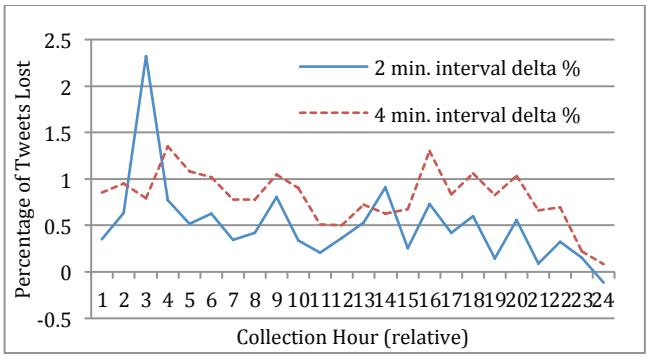

Figure 5 - Percentage drop-off for q=sports at 1, 2, and 4 min. collection intervals

Experiment three continues the effort to characterize the Twitter Search API. Twitter Zombie was used to run two identical search jobs both starting at the same time. Figure 6 shows how closely the hourly counts of returned tweets agree. The percentage differences are detailed in Figure **7**.

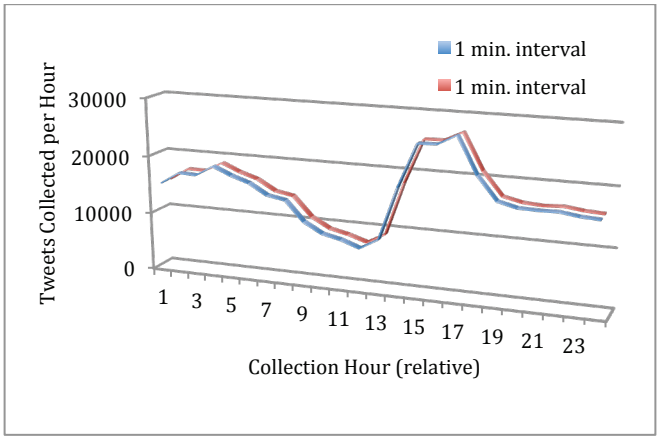

Figure 6 - Tweets per hour for q=sex at 1 min. intervals

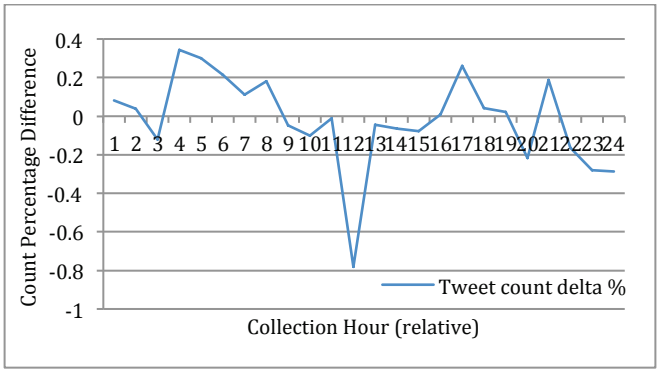

Figure 7 - Percent difference for q=sex at 1 min. and 1 min. collection intervals

4.1 Social Transformation of Tweets

We use the metadata of Tweets, described in Section 3.3.2 to transform the raw Twitter data into a representation of interactions. These social interaction representations take the form of social networks around four key affordances: one each for direct mentions, retweets, devices and hashtags. These representations surface the social structure that is implicit in tweets containing traces of these user affordances, and transform them into easily analyzable node pairs in a table. This aspect of

Twitter Zombie further enables statistical analysis of Twitter metadata with relative ease. Direct mentions and retweets are stored as node pairs of individual Twitter users in the analysis tables. Device information and hashtags are stored as bipartite networks, where one node type is a person, and the other is a device type or hashtag respectively.

4.2 Analysis of Socially Transformed Tweets

Our extraction of social metadata from the tweet string enables Twitter Zombie to swiftly visualize social and device information about Twitter activity related to planned and unplanned events that emerge in the Twittersphere. To illustrate the utility of this aspect of the Twitter Zombie Architecture, we conducted a pilot study, which we describe in following section.

5. PILOT STUDY

We utilized the Twitter Zombie Collection Architecture and methodological approach to study Twitter activity around the US Republican Party Presidential Primary Debate in South Carolina on January 16th, 2012. In an effort to focus our collection efforts only on data related to that specific debate we collected Twitter messages that contained the hashtag #SCDebate. In addition to that hashtag, the debate sponsor, FOX News, encouraged individuals to tweet the candidate's name along with the hashtag #answer or #dodge when a question was asked to identify whether the public believed that the candidate was providing an answer to the question or dodging the question. In order to facilitate this activity, FOX News created a page on their website where individuals were able to use a button specifically created to facilitate the tweeting of #answer and #dodge. The following findings illustrate the presence of different types of communication networks and the adoption of specific technological applications for different purposes.

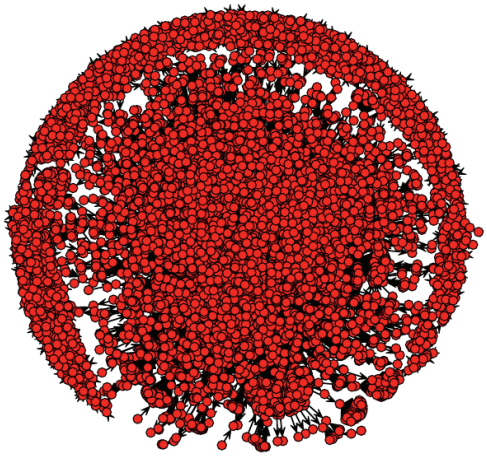

Figure 8 - Reply To Network

The methodological approach explained earlier in this article allowed us to identify several interesting characteristics of the public Twitter behavior surrounding the debate. Utilizing the #scdebate collection we are able to analyze reply-to and retweet networks to illustrate two distinctly different behaviors. Figure 2 illustrates the diffuse network of reply-to behavior in the #scdebate data. We see that there are a lot of disconnected groups of discourse indicative of an unstructured set of Tweets directed towards other individuals.

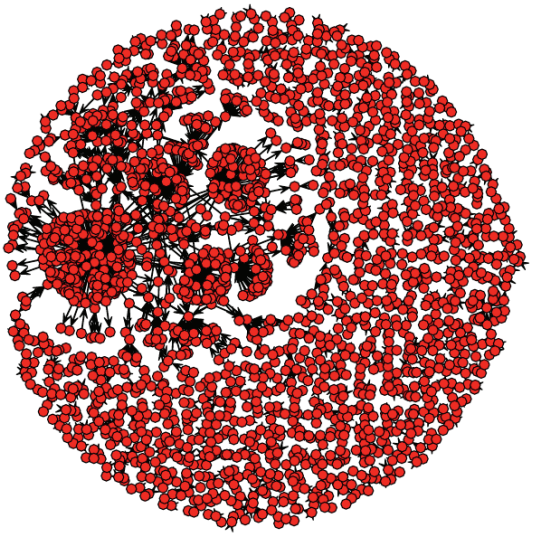

Figure 9 - Retweet Network

On the other hand, we see in the retweet network (Figure 3) that there is a significant amount of disconnected retweet behavior similar to the reply to network, but we also identify a set of clusters that illustrate concentrated retweet behavior. The most retweeted individuals are: BorowitzReport, TheFix, BretBaier and washingtonpost. These accounts all represent journalistic entities tweeting about the debate, with BretBaier being the moderator. The below distribution of devices that are used to tweet in our dataset illustrates that the "web" was the most popular mechanism for tweeting followed by Tweetdeck for the hashtag #scdebate.

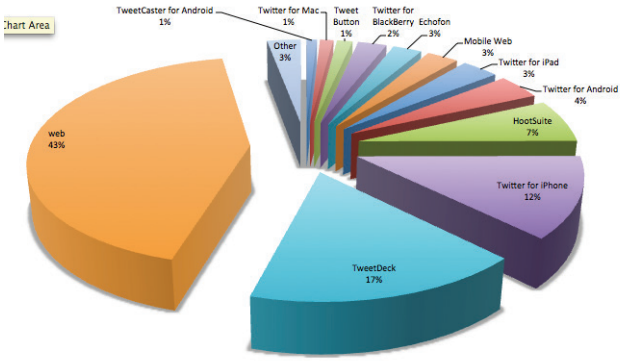

Figure 10 - #SCDebate Device Distribution

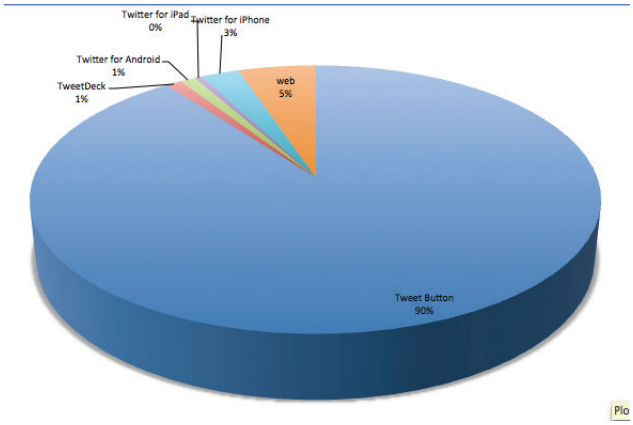

Figure 11 - #answer device distribution

In our second stage of analysis we examined the dataset of collected tweets using #answer and #dodge. The hashtags #answer and #dodge did not need to be tweeted with #scdebate indicating that there did not need to be overlap between the two datasets. The device distribution of #answer and #dodge is quite different from that of #scdebate. Figure five illustrates the device uses for the #answer button.

Figure six illustrates the device usage for the #dodge button. You can see that 90% of people answering, "answer" used the Fox News web page, while 85% of people saying "dodge" did. .

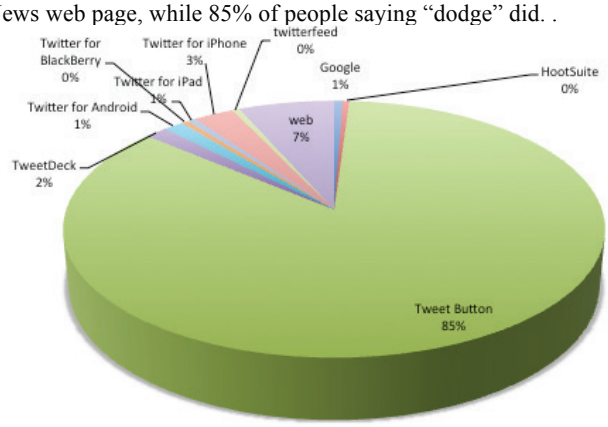

Figure 12 - #dodge device distribution

The Tweet Button on the http://www.foxnews.com website was the most dominant technique for tweeting using those hashtags. The contrast in device distribution between the dataset of #scdebate and #answer and #dodge identified two distinct areas of discourse that warranted deeper analysis of the individual activity within each dataset to highlight the differences of the behavior related to one event using different hashtags.

Table one, below, illustrates the number of unique participants relative to the number of tweets. The lower the percentage, the higher number the number of repeat participants. Based on the statistics, individuals using #answer and #dodge posted more times than those that used #scdebate. This is indicative of the more focused purpose of the #answer and #dodge hashtags, relative to the more general discourse happening with the #scdebate hashtag. This higher level of repeat participation could also be a result of a concentrated technological mechanism to utilize those hashtags such as the Fox News website.

Table 2- Hashtags in South Carolina Debate

Hashtag	Percent of Unique Tweeters
Answer	32.63%
Dodge	29.93%
SCDebate	55.02%

We wanted to extend this line of inquiry further and understand the participation rates between the #answer and #dodge hashtags. Our findings indicate that close to 37% of the individuals posted a tweet using both #answer and #dodge. This demonstrates that some participants were interested in participating in discourse relative to candidates using both #answer and #dodge, but that close to 63% of the total participants only participated using either #Answer or #Dodge.

In an effort to identify if there were distinct discourse communities we compared the user composition of the two datasets (#scdebate versus #answer and #dodge). We treated the individuals that contributed using #scdebate as one dataset and individuals using either #answer or #dodge as the other dataset. Our analysis identified that only **13%** of the individuals that used the #SCDebate hashtag participated in the #Answer vs. #Dodge exercise.

This low percentage of participation in both the general discourse and the #Answer versus #Dodge exercise illustrates that there were two different discourse communities active on Twitter, participating in social media in relation to the January 16th debate. This distinction is illustrated by the differences in device utilization as well as the differences in user overlap across the two distinct hashtag sets.

Our pilot study illustrates the power of the Twitter Zombie toolset for providing rapid, consistent analysis of planned and unplanned events as represented through Twitter discourse. The architecture, social transformation and analysis software system developed here helps to close an important gap of consistency and transparency in social media research.

6. DISCUSSION

Twitter is participatory mass media. Our architecture and methodological approach paves the way for other researchers to examine emergent, social phenomena on Twitter. We enable this continued inquiry with a tool called Twitter Zombie, which gives social media researchers, like other social science researchers before us, a transparent framework for data gathering, analysis and reporting. As Howison et al point out [23], electronic trace data is not representative of social interaction, even though people are responsible for its creation. How people interact through a particular socio-technical milieu of tools, affordances and practices is different in each instance. Our validation of the results of using our tool shows that differences and assumptions about data returned must be carefully and regularly examined. We do not know if our findings with relation to the Twitter API are a result of shifts in the API over time, or shortcomings in the methods of other researchers or, plausibly, us. The inability of the social media research community to speak with authority about these kinds of data completeness and quality issues is at once expected at the dawn of a research era, where we are now, but is also essential to address in order to ensure increasingly useful, valid and relevant results for society. Our demonstration of the use of the Twitter Zombie in a pilot study illustrates how powerful validated tools can be.

Some elements of research method complexity are specific to the social media platform. In the case of Twitter, the selection of API is demonstrated here to play a significant role in filtering data. For the social science researcher these choices, like decisions of methodological approaches such as survey sampling, ethnography site selection or theory for classic social science researchers, will influence the resulting findings and ensuing development of theory.

When a phenomenon is new, as social media has been for the past decade, inductive research methods are called for to define the salient constructs for future inquiry. In the case of social media research, which often includes the examination of electronic trace data, and other quantitative methods of inquiry, the road map for inductive research is ill defined from prior phenomena. We are now moving into an era where key constructs, like the socio-technical interaction, the technical interaction and others, are clearly defined by prior literature. We are also entering an era where we can look back on a decade of research on social media in general, and six years of research on Twitter, to discern the

inconsistencies in our approaches. These inconsistencies, while expected as inquiry begins, are important to iron out as inquiry advances.

We do not claim to have built the ultimate, or even the penultimate data collection, transformation and analysis software system and methodological approach for Twitter research. Instead, we put forth this example as a reference for other researchers to use. The work presented here should inspire critique from over 100 researchers whose work we build on, but upon whom we also call to advance the field.

7. CONCLUSION

We make our collection, transformation and analytical processes for Twitter available for scrutiny and use by other researchers[2]. Our hope is that the social media research community, as a whole, will rise up to the challenge of building corpora of comparative studies looking at social computational phenomena across the Twittersphere. Through this approach, our findings will grow, theory will emerge and our contribution to an increasingly technologically mediated world will possibly find easier translation.

8. ACKNOWLEDGMENTS

We would like to thank Leysia Palen for her assistance in understanding the challenges associated with data collection and management in Twitter research. Thanks to Nora McDonald for providing numerous reviews of this paper's drafts. We also thank Scott Robertson and Ravi Vatripu for their insights on data collection and management in social media research.

9. REFERENCES

[1] Akcora, C. G., Bayir, M. A., Demirbas, M., and Ferhatosmanoglu, H. 2010. Identifying Breakpoints in Public Opinion. Workshop on Social Media Analytics.

[2] Bae, Y. and Lee, H. 2011. A Sentiment Analysis of Audiences on Twitter: Who is the Positive or Negative Audience of Popular Twitterers. ICHIT 2011. 732-739.

[3] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. WSDM.

[4] Barbosa, L. and Feng, J. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. International Conference on Computational Linguistics. 36-44.

[5] Bermingham, A. and Smeaton, A. 2010. Classifying Sentiment in Microblogs: Is Brevity an Advantage? CIKM. 2010.

[6] Bigonha, C. A. S., Cardoso, T. N. C., Moro, M. M., Almeida, V. A. F., and Goncalves, M. A. 2010. Detecting Evangelists and Detractors on Twitter. Brazilian Symposium on Multimedia and the Web (WebMedia).

[7] Bollen, J., Mao, H., and Zeng, X. 2011. Twitter mood predicts the stock market. Journal of Computational Science. 2, 2011, 1-8.

[8] boyd, d., Golder, S., and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. Hawaii International Conference on System Sciences. 43.

[9] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM. 4th.

[10] Chew, C. and Eysenbach, G. 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS One. 5, 11.

[11] Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., and Menczer, F. 2011. Political Polarization on Twitter. ICWSM. 5th.

[12] Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. Workshop on Social Media Analytics. 1st.

[13] Diakopoulos, N. A. and Shamma, D. A. 2010. Characterizing Debate Performance via Aggregated Twitter Seniment. CHI.

[14] Fonseca, A. 2011. Modeling Political Opinion Dynamics Through Social media and Multi-Agent Simulation. First Doctoral Workshop for Complexity Sciences.

[15] Gentry, J. 2012. twitteR package for R.

[16] Golbeck, J., Grimes, J. M., and Rogers, A. 2010. Twitter Use by the U.S. Congress. Journal of American Society for Information Science. 61, 8, 1612-1621.

[17] Gruzd, A., Doiron, S., and Mai, P. 2011. Is Happiness Contagious Online? A Case of Twitter and the 2010 Winter Olympics. Hawaii International Conference on System Sciences. 44th.

[18] Guerra, P. H. C., Veloso, A., Meira, W., and Almeida, V. 2011. From Bias to Opinion: a Transfer-Learning Approach to Real-Time Sentiment Analysis. KDD. 2011.

[19] Hanna, A., Sayre, B., Bode, L., Yang, J., and Shah, D. 2011. Mapping the Political Twitterverse: Candidates and Their Followers in the Midterms. ICWSM. 2011.

[20] Hansen, D., Schneiderman, B., and Smith, M. A. 2011 Analyzing Social Media Networks with NodeXL. Elsevier.

[21] Heverin, T. and Zach, L. 2011. Use of Microblogging for Collective Sense-Making During Violent Crises: A Study of Three Campus Shootings. Journal of American Society for Information Science. 10.1002/asi.21685.

[22] Honeycutt, C. and Herring, S. C. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. Hawaii International Conference on System Sciences. 43.

[23] Howison, J., Wiggins, A., and Crowston, K. 2012. Validity Issues in the Use of Social Network Analysis for the Study of Online Communities. Journal of the Association of Information Systems. 12, 2.

[24] Huang, J., Thornton, K. M., and Efthimiadis, E. 2010. Conversational Tagging in Twitter. HT.

[25] Hughes, A. L. and Palen, L. 2009. Twitter Adoption and Use in Mass Convergence and Emergency Events. ISCRAM. 6th.

[26] Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. WEBKDD/SNA-KDD Workshop.

[27] Jurgens, P., Jungherr, A., and Schoen, H. 2011. Small Worlds with a Difference: New Gatekeepers and the Filtering of Political Information on Twitter. WebSci.

[2] link included in final version (blind review omitted)

[28] Kim, M. and Park, H. W. 2012. Measuring Twitter-Based political participation and deliberation in the South Korean context by using social network and Triple Helix indicators. Scientometrics. 90, 1.

[29] Krishnamurthy, B., Gill, P., and Arlitt, M. 2008. A Few Chirps about Twitter. WOSN.

[30] Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a Social Network or a News Media? WWW.

[31] Lerman, K. and Ghosh, R. 2010. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. ICWSM. 4th.

[32] Livne, A., Simmons, M. P., Adar, E., and Adamic, L. A. 2011. The Party is Over Here: Structure and Content in the 2010 Election. ICWSM. 5th.

[33] Marwick, A. E. and boyd, d. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Society. 13.

[34] Meador, C. and Gluck, J. 2009. Analyzing the Relationship Between Tweets, Box-Office Performance and Stocks. Methods.

[35] Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. 2011. How (Not) To Predict Elections. International Conference on Social Computing. 165-171.

[36] Murthy, D., Gross, A., and Oliveira, D. 2011. Understanding Cancer-based Networks in Twitter using Social Network Analysis. IEEE International Conference on Semantic Computing. 5th.

[37] Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. T. 2011. Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail. International Conference on Social Computing.

[38] Mustafaraj, E. and Metaxas, P. T. 2011. What Edited Retweets Reveal about Online Political Discourse. Workshop on Analyzing Microtext.

[39] Naaman, M., Becker, H., and Gravano, L. 2011. Hip and Trendy: Characterizing Emerging Trends on Twitter. Journal of American Society for Information Science. 62, 5, 902-918.

[40] Pole, A. and Xenos, M. 2011. Like, Comments and Retweets: Facebooking and Tweeting on the 2010 Gubernatorial Campaign Trail. State Politics and Policy Conference. 11th.

[41] Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. 2012. Tracking "Gross Community Happiness" from Tweets. CSCW.

[42] Romero, D. M., Meeder, B., and Kleinberg, J. 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contation on Twitter. WWW.

[43] Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. 2010. Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. CSCW.

[44] Sutton, J., Palen, L., and Shklovski, I. 2008. Backchannels on the front lines: Emergent use of social media in the 2007 Southern California fire. Information Systems for Crisis Response and Management Conference.

[45] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM. 4th.

[46] Werry, C. C. 1996 Linguistic and interactional features of Internet Relay Chat. In Computer-mediated communication: Linguistic, social and cross-cultural perspectives, S. C. Herring,Ed John Benjamins.

[47] Williams, C. and Gulati, G. J. 2008. What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries. American Political Science Association.

[48] Wohn, D. Y. and Na, E. K. 2011. Tweeting about TV: Sharing television viewing experiences via social media message streams. First Monday. 16, 3.

[49] Zelenkauskaite, A. and Herring, S. C. 2008. Television-mediated conversation: Coherence in Italian iTV SMS Chat. Hawaii International Conference on System Sciences. 41.

[50] Zhao, S., Zhong, L., Wickramasuriya, J., and Vasudevan, V. 2011. Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. rxiv preprint arXiv:1106.4300.