Technologically Mediated Discourse and Information Exchange through Medium Specific
Syntactical Features: The 2012 Presidential Election on Twitter

A Thesis

Submitted to the Faculty

of

Drexel University

by

Christopher M. Mascaro

in partial fulfillment of the
requirements for the degree

of

Doctor of Philosophy

May 2014
# Table of Contents

List of Tables .......................................................................................................................... vii  
List of Figures .......................................................................................................................... iv  
Abstract ..................................................................................................................................... x  

CHAPTER 1: INTRODUCTION ................................................................................................. 1  
Political Discourse on the Internet .......................................................................................... 1  
Statement of Problem ............................................................................................................... 3  
Purpose of Study ...................................................................................................................... 5  
Statement of Research Questions ............................................................................................ 7  
Contributions to Knowledge .................................................................................................... 7  
Limitations ................................................................................................................................ 8  

CHAPTER 2: LITERATURE REVIEW ....................................................................................... 10  
Theoretical Frame .................................................................................................................... 10  
  The absence of political participation ..................................................................................... 10  
  Moving from the public sphere online ................................................................................... 12  
  Tying it together: applying theory to the socio-technical environment ................................ 13  
The Internet’s Role in American Elections .............................................................................. 14  
  The Internet is introduced: 1992 - 2004 ............................................................................. 15  
  From static content to dynamic engagement: The 2008 Presidential Election ................. 17  
  The 2012 Presidential Election ........................................................................................... 18  
Twitter as Technology ............................................................................................................ 20  
  Establishing a baseline understanding of Twitter ................................................................. 21  
  Twitter’s social connections ................................................................................................ 22  
  Twitter syntactical features .................................................................................................. 23  
  The nature of retweets .......................................................................................................... 24  
Politics and Twitter ................................................................................................................ 32  
  Elected members and candidate use of Twitter ................................................................. 32  
  Categorizing Twitters by political language and affiliation ............................................. 34  
  The study ............................................................................................................................ 39  

CHAPTER 3: DATASET CREATION AND ANALYTICAL METHODOLOGY ....................... 40  
Data Collection ......................................................................................................................... 40  
  Original TwitterZombie design ......................................................................................... 40  
  Redesigned Technical infrastructure of TwitterZombie-n .................................................. 41  
Dataset Creation ....................................................................................................................... 43  
  Job selection ....................................................................................................................... 44  
  Candidate dataset creation ............................................................................................... 45  
  Event dataset creation ....................................................................................................... 46  
  Campaign promoted hashtag dataset creation ................................................................. 49  
Collection summary ................................................................................................................. 49  
  Dataset provenance and partitioning ............................................................................... 51  
Dataset Description ............................................................................................................... 53  
  Syntactical feature overview ............................................................................................. 53  
Analytical Methods ................................................................................................................ 56  
  Question 1 .......................................................................................................................... 58
Question 2 .................................................................................................................. 61
Question 3 .................................................................................................................. 65

Methods Summary .................................................................................................... 70

CHAPTER 4: FINDINGS .............................................................................................. 73

Overview ................................................................................................................... 73

Question 1 .................................................................................................................. 75
Finding 1a: .................................................................................................................. 76
Finding 1b: .................................................................................................................. 79
Finding 2: .................................................................................................................... 80
Finding 3a: .................................................................................................................. 82
Finding 3b: .................................................................................................................. 84
Finding 4: .................................................................................................................... 85
Finding 5: .................................................................................................................... 87

Question 2: .................................................................................................................. 90
Finding 5a: .................................................................................................................. 90
Finding 6b: .................................................................................................................. 92
Finding 7a: .................................................................................................................. 94
Finding 7b: .................................................................................................................. 95
Finding 7c: .................................................................................................................. 97
Finding 8: ................................................................................................................... 99
Finding 9: ................................................................................................................... 100

Question 3: .................................................................................................................. 101
Finding 10: .................................................................................................................. 102
Finding 11: .................................................................................................................. 104
Finding 12: .................................................................................................................. 106
Finding 13a: .................................................................................................................. 107
Finding 13b: .................................................................................................................. 108
Finding 14: .................................................................................................................. 109

Looking forward ....................................................................................................... 110

CHAPTER 5: Discussion .......................................................................................... 112

Theoretical Implications .......................................................................................... 113
Democratic theory ..................................................................................................... 113
The Internet’s effect on political processes ............................................................... 116
Twitter as a new public sphere .................................................................................. 117
Moving to the private sphere .................................................................................... 119
Deliberative discourse not realized .......................................................................... 120

Socio-technical Implications ................................................................................... 121
Syntactical Feature Usage ....................................................................................... 121
Information Exchange ............................................................................................... 122

Practical Implications ............................................................................................. 123
Citizenry .................................................................................................................... 125
The influence of the technological platform ............................................................. 126

Methodological Implications for Social Media Research ......................................... 127
Dataset Construction ............................................................................................... 128
Unit of Analysis Choice ............................................................................................ 129
Conceptualizing the Approach Clearly .................................................................... 130

Conclusion and Next Steps ..................................................................................... 130

List of References .................................................................................................... 133
| Appendix A: Number of Tweets Collected by Query | 146 |
| Appendix B: Datasets Used for Analysis | 148 |
| Appendix C: TwitterZombie Data Fields | 149 |
| Appendix D: TwitterZombieAnalysisScript.r Detailed | 150 |
| Appendix E: Shortened URL Decoding Script Detailed | 151 |
| Vita | 152 |
List of Tables

Table 1: Prominent 2012 Electoral Events .............................................................. 18
Table 2: Syntactical Features Enumerated .............................................................. 24
Table 3: Candidate Queries .................................................................................. 46
Table 4: Key Election Season Events and Related Queries .................................. 48
Table 5: Promoted Hashtags ............................................................................... 49
Table 6: Dataset Descriptions ............................................................................. 52
Table 7: Syntactical Feature Conceptualization .................................................. 54
Table 8: Raw Count of Syntactical Features ......................................................... 55
Table 9: Syntactical Feature Presence in the Complete Dataset ............................ 55
Table 10: Candidate Mention Percentage ............................................................ 56
Table 11: Syntactical Feature Percentage Presence Summary by Day ................. 56
Table 12: Overview of Research Question 1 Datasets, Methods, Scripts and Output Files .................................................................................................................. 58
Table 13: Dates included in Debate Days Dataset ............................................... 58
Table 14: Overview of Research Question 2 Datasets, Methods, Scripts and Output Files .................................................................................................................. 61
Table 15: Overview of Research Question 3 Datasets, Methods, Scripts and Output File .................................................................................................................. 62
Table 16: URLs by Dataset and Date Decoded ...................................................... 67
Table 17: Overview of findings, datasets, methods, scripts and output files ........... 70
Table 18: Overview of RQs, Datasets, Methods and Findings ............................. 73
Table 19: Syntactical Feature Overview of Debate Days ....................................... 79
Table 20: Most Frequently Occurring Hashtags ................................................... 80
Table 21: Most Frequently Occurring Hashtags Not Explicitly Collected .............. 81
Table 22: Percentage of Overall Tweets with #debates ........................................ 85
Table 23: Top Hashtags by Debate ..................................................................... 85
Table 24: At-mention occurrences by code and tweets ........................................ 91
Table 25: Candidate At-Mention Summary Statistics .......................................... 93
Table 26: At-reply Out-degree .......................................................................... 96
Table 27: At-reply in-degree ............................................................................. 96
Table 28: Types of Individuals Who Received At-replies ..................................... 98
Table 29: Top recipients in Conversational Pairs ............................................... 99
Table 30: Summary Statistics for those that both sent and received an at-reply ............................................................. 100
Table 31: Number of at-replies based on conversational activity .......................................................................................................................... 101
Table 32: Most Frequently Occurring URLs by Tweet Type ...................................................................................................................... 102
Table 33: Top 50 Base URL Codes by Dataset ........................................................................................................................................... 104
Table 34: Percentage of Overall Tweets with URLs by Code .................................................................................................................. 104
Table 35: Syntactical Feature Contribution to Overall URL Presence ........................................................................................................... 105
Table 36: Percentage of URLs unable to be Decoded By URL Shortener ........................................................................................................ 106
Table 37: Retweet In-Degree ................................................................................................................................................................. 108
Table 38: Via @ In-degree ................................................................................................................................................................. 110
Table 39: Tweets by Job ................................................................................................................................................................. 144
List of Figures

Figure 1: #debates Usage by Day ................................................................. 47
Figure 2: Collection and Dataset Construction Overview ................................ 50
Figure 3: Dataset Overview ........................................................................ 51
Figure 4: TZScript Analysis Process ............................................................. 57
Figure 6: URL Decoding Process ................................................................. 66
Figure 7: Tweets Per Day ........................................................................... 77
Figure 8: Syntactical Feature Usage by Day ............................................... 78
Figure 9: Frequency of #election2012 Tweets by Day .................................... 83
Figure 10: Frequency of #debates by Day ................................................... 84
Figure 11: Temporal Distribution of Promoted Hashtags ............................. 88
Figure 12: Temporal Distribution of #cantafford4more ................................ 89
Figure 13: Candidate At-Mention Percentage by Day ................................... 92
Figure 14: At-reply Percentages By Day ..................................................... 94
Abstract

Technologically Mediated Discourse and Information Exchange through Medium Specific Syntactical Features: The 2012 Presidential Election On Twitter

Christopher M. Mascaro

Political discourse has been historically constrained by geographic proximity of participants. The introduction of the Internet and specifically social media has altered these geographic constraints and political discourse is now one of the most prevalent activities in social media. The increasing use of technology to acquire political information and participate in the political process in the United States creates a gap between what is understood about political activity in a democratic society and the specific technological features people use. As more individuals begin to use technology for political activity, understanding how the technology is used becomes increasingly important.

Previous research exploring political discourse on social media has focused on one discrete event or a narrow time period. This narrow focus limits the understanding of the complex environment that comprises an election. This study takes a longitudinal approach and uses network analysis, co-occurrence analysis and temporal frequency analysis to examine a 53 million Twitter message (tweet) corpus collected during the 2012 Presidential Election (August 20, 2012 – November 13, 2012) to understand how individuals use Twitter to engage in political discourse. The queries used to compose the dataset were theoretically informed based on democratic theory and previous socio-technical research.

This study makes three contributions to the existing literature. First, this study identifies that individuals use syntactical features differently in the context of an acute event such as a debate. Second, this study indicates that, although candidates and media are the most talked about and talked to, these interactions elicit no response. Third, this study reveals that information shared through URLs was predominantly user-generated content from Twitter and mass media information suggesting a reflexive information-sharing environment.

This study illustrates that even with the availability of the numerous technological and syntactical features to facilitate interactions and share information, there is still a limited realization of the promise that technologies such as Twitter afford. Instead of fundamentally changing the political discourse process by having individuals use it for two-way communication, Twitter amplifies the existing political environment where there is limited cohesive discourse and communication is one-way.
CHAPTER 1: INTRODUCTION

The exchange of ideas, discussion of issues relevant to the electorate and debate about how to allocate the limited resources of government are essential for a functioning, democratic society. Political discourse, the interaction between politicians and citizens and among citizens about political matters, is integral in developing new ideas and addressing problems facing society (van Dijk, 1997). How and where this political discourse takes place is changing. As a result of these changes the electronic trace data that represents this discourse needs to be examined to further understand the changes and impact on society.

One of the greatest benefits of Internet is that it reduces the costs of obtaining information and participating in the political process (Bimber, 2001). Citizens that utilize the Internet are more likely to be civically engaged than those that do not and this is likely the result of the lower barrier of entry to participate (Best & Krueger, 2005; Krueger, 2006). The costs that elected officials or candidates have to communicate with the public have also been reduced with the introduction of the Internet to the political process. This has led to the greater use of technology in campaigns.

The rapid growth of technology for acquiring information and participating in the political process in the United States creates a gap between what is understood about political participation in a democratic society and the specific technological features people use. As more individuals begin to adopt technology to acquire political information and participate in the political process, understanding the technology that enables discourse along with the networks of discourse that result becomes increasingly important (Lazer, 2011). This understanding can be developed through contextual studies of political activity in various technologies. The following study addresses this gap by examining activity in Twitter during the 2012 Presidential election.

Political Discourse on the Internet

Political discourse in the physical world was historically constrained by the geographic proximity of participants — people talk about politics, politicians and social issues with others whom they already know and whom they are in close geographic proximity (Bearman & Parigi, 2004; Huckfeldt & Sprague, 1987). Much of this discourse also results from informal social interactions. These interactions contribute to individual political awareness and participation (McClurg, 2003). New technology, such as social
networking sites (SNS), facilitates discourse independent of geography and pre-existing relationships. In this new medium, citizens form ties with individuals they already know, but are also able to engage with a diffuse, growing and geographically diverse, new group of people. The ability to form new ties and participate in discourse with existing ties in a new medium circumvents prior geographic constraints and increases the possibility of interactions.

Political discourse is one of the most common forms of activity on social networking sites. The 2012 Election was the most active to date on social media. Pew Research found that 66% of social media users (39% of all adults in the United States) participated in a civic or political activity using social media in the 2012 election (Rainie, Smith, Scelzoman, Brady, & Verba, 2012). These activities included “liking” or promoting political material (38%), encouraging others to vote (35%), posting general political commentary (34%), reposting others information to include retweeting\(^2\) (33%), posting a URL to a political story (28%) and following a candidate (20%). These statistics illustrate the high amount of activity related to the election in social media.

The 2012 election also saw a change in how individuals accessed social networking sites. In 2012, 27% of registered voters used their phone to consume election or political news with 5% signing up to receive information from a candidate or other organization involved with a campaign (Smith & Duggan, 2012b). Pew found that over half (55%) of registered voters watched a political video during the campaign and 52% of registered voters had others recommend a video for them to watch (Smith & Duggan, 2012a). These videos were a mix of news reports, previously recorded speeches, campaign videos and live press conferences. The increase in the ability to access political information in numerous ways and digest many forms of content helps contribute to the increase in popularity of political activity on the Internet -- specifically within social media.

Social networking sites are influencing relationships surrounding political matters. In a study conducted near the end of the election and published on Election Day, Pew found that 22% of registered

---

1 The definition of a social networking site is broad, but for the purposes of this study I use Ellison (2007) (Ellison, 2007), which would apply to the commonly accepted websites such as Facebook, Instagram and Twitter, the subject of this study. The terminology social media and social networking sites (SNS) are used interchangeably in this study.

2 Retweeting is the act of sharing another individual’s tweet with one’s own network. This term is further explored later in chapter 2.
voters announced their vote on social media. The partisan preference of these individuals was not considerably different as 25% of these voters voted for President Obama and 20% claimed to have voted for Mitt Romney (Rainie, 2012). In an earlier study, 10% of users have blocked, un-friended or hidden another individual’s comment as a result of political content (Rainie & Smith, 2012). It is currently unknown what the implications of announcing one’s vote on a SNS is, but this activity signals a new mode of announcing personal political preferences.

This increasing reliance of citizens on social networking sites to acquire political information and participate in the civic and political activities such as donating to a campaign, sharing information about a candidate or encouraging others to get involved requires a further understanding of how these activities occur in social media. There currently lacks research to examine the actual activity that occurred in the technology. In this study, the 2012 election is used to understand how individuals used syntactical features such as retweets, hashtags, URLs, at-reply’s and at-mentions in Twitter to participate and engage in political discourse.

**Statement of Problem**

Previous research of political discourse on social networking sites has focused on one discrete event or a narrow time period such as the immediate run up to an election, one debate or a press conference (Diakopoulos & Shamma, 2010; Robertson, Vatrapu, & Medina, 2009; Shamma, Kennedy, & Churchill, 2009; Williams & Gulati, 2007). These approaches are inadequate for examining both sides of the discourse if they exist (the citizens and candidates or the citizens and elected officials) and how this discourse reflects normal time periods without an event that draws increased interest from the public. As a result, previous research focuses only on political discourse in social media surrounding acute events without situating these analyses in the context of a larger event – an election.

The following study takes a longitudinal approach that focuses on more than just one event to understand how individuals use the numerous syntactical features of a technology to engage with each other during both acute and non-acute events (Ancu & Cozma, 2009; Sweetser & Weaver Lariscy, 2008). The longitudinal approach used in this study is defined as taking place over the complete election period

---

3 In this dataset 1,476,786 tweets contained the text “voted for” on election day, illustrating a significant amount of individuals announcing their vote.
starting with the conventions and ending shortly after election day. All of the events that make up an electoral time period are treated as observations in the context of the time period.

The focus on only one event as is the case in most research of political activity in social media takes the analysis of that event out of context. Many individuals who participate in social media activity during a debate or other acute event may be using keywords or hashtags that may not be immediately identified as being related to the debate. This narrows the focus of the analysis, but also limits broader insights that could be gained from examining discourse occurring at the same time using different keywords of hashtags.

The longitudinal approach and dataset construction used for this study allows for the examination of a series of events that are related through the greater context of the election by situating them in the electoral context using a temporal dimension. This approach allows for a systematic analysis that considers events as related to the larger context and not isolated. The findings related to research question one illustrate how limiting analysis to what is believed to be the only hashtag related to events may lead to a lack of collection of important pieces of data.

As discussed in the literature review (chapter 2), the recent decade has brought about a number of theories of democratic engagement that have been established by examining technologically mediated activity in the context of one event. This narrow scope can lead to different findings than examining discourse longitudinally. Opinions and ideas can be shaped contextually as a result of one event and engagement during an event may differ from other times during a time period.

The focus on a narrow timeframe in political discourse research on social networking sites extends beyond just electoral processes in the United States. Research into the use of Twitter by the electorate in Singapore examined a corpus containing 110,815 political tweets filtered from a larger dataset that spanned 11 days (Skoric, Poor, Achananuparp, Lim, & Jiang, 2012). A similar study of 104,003 tweets from a five-week period that examined a German national election further illustrates the tendency of narrow time periods of analysis in electoral political research (Tumasjan, Sprenger, Sandner, & Welpe, 2010). In both cases, the datasets were constructed by collecting data from Twitter using hashtags, keywords or candidates involved in national level election as filters. The collected terms often amount to only a few and as a result can create bias in the findings as a result the narrow dataset collected.
In Canada, similar research has centered on local elections that have limited participation and this research has also used similar narrow filtering of criteria to produce narrow corpus of data (Elmer, Langlois, & McKelvey, in press; Small, 2010). In Australia, an analysis of the #ausvotes hashtag, the hashtag associated with the national election, over the course of approximately 6 weeks collected 415,009 tweets (Bruns & Burgess, 2011). This analysis yielded interesting insights into how individuals used Twitter to engage with others, but the narrow time period studied coupled with the narrow dataset limited the findings from such analysis.

Although these studies illustrate an interest in studying political discourse and activity on Twitter, their findings are limited since they only focus on one discrete event, the actual election in a narrow timeframe. Narrowing the analysis of this specific timeframe is important in understanding election-specific discourse (Mascaro, Novak, & Goggins, 2012a), but limits insight about who the central participants and subjects of the discourse are over time and how interactions evolve. The longitudinal nature of the data in this study helps to identify who the most important actors are throughout the whole election and not just during one slice of time. This has numerous implications for the development of political strategy and engaging with the public on a number of issues in a more efficient manner from the perspective of the candidate and also the media. This research also has numerous implications for understanding how citizens go about acquiring and exchanging political information.

Contemporary political discourse has been radically transformed by technology, but may still promote century-old issues of democratic participation and action that are reviewed in chapter 2. Even though technology exists to facilitate communication, the ongoing evolution of social media makes consistent and effective harnessing of this power in the new medium an elusive goal.

**Purpose of Study**

The following study examines how citizens and candidates engage in a technologically mediated environment using medium specific syntactical features to include hashtags, at-mentions, at-replies, retweets and the inclusion of URLs in the text. These syntactical features allow users to engage in discourse, talk about and talk to candidates and exchange information with others. This study of the 2012 Presidential Election is situated in two distinct sets of literature, the emerging literature that examines how social media facilitates information exchange and personal interaction and how traditional theories of democratic
participation in contemporary political discourse are being realized or not (Dewey, 1927; Habermas, Lennox, & Lennox, 1974; Habermas, 1984; Habermas, 1991; Lippmann, 1925; Papacharissi, 2002; Papacharissi, 2004; Papacharissi, 2010).

One of the previous limitations of research on technologically mediated political discourse is the lack of clarity about how tools that collect data from technologies, such as Twitter, and other social media data are implemented. Goggins et al. (2012) propose a general ontology and methodological approach for analysis of electronic trace data and Black et al. (2012) identify a system for collecting and analyzing Twitter data. Most studies of Twitter to date fail to specify how data is retrieved and how the use of different user defined syntactical features influence the subsequent analysis. Through incorporation of an established analytical approach and a documented collection platform, the findings of this study extend the existing literature related to social media data collection, management and processing.

The data collection for the study occurred over an 85 day time period from before the Republican National Convention in August 2012 through the week following Election Day in November 2013 (August 20, 2012 – November 13, 2012). This time period is often treated as the official electoral period since the candidates are not officially nominated until the conventions and that marks the beginning of the official campaigning. The 68 queries used to create the approximately 53 million tweet corpus are informed by both theory and the events that unfolded. As described in the methods chapter (chapter 3), the initial set of collection terms evolved as more terms were identified to be relevant to the election.

The use of TwitterZombie for collection, coupled with a diverse combination of analytical methods borrowed from a wide spectrum of previous work on Twitter detailed in Chapter 2 allowed for a comprehensive Twitter dataset related to a Presidential Election to be collected and analyzed as part of this study. This approach allows the findings to be applicable to both political discourse and discourse more broadly as the nature of the activity extended beyond just political discourse alone.

---

4 This statement is based on a current review of the literature as of November 2013 and does not take into account the fact that Twitter has all of the tweets and can conduct analysis on everything during the time period. This statement is made in the context of datasets external to Twitter.
Statement of Research Questions

The following research questions guide the examination of the 53 million tweet corpus by approaching the dataset from a number of different perspectives that take into account the numerous syntactical features of Twitter.

1. How does the political discourse related to the 2012 Presidential election manifest itself in a technologically mediated environment?
   a. How does this discourse in Twitter differ during acute events such as debates compared to long-term discourse that occurs throughout the election?
   b. How does the use of politically oriented hashtags in Twitter facilitate this discourse?

2. How does the political discourse that occurs in Twitter using the syntactical features of the at-mention and at-reply surrounding the 2012 Presidential Election identify an emerging participatory public engaged in political discourse?

3. To what extent do the URL and retweet syntactical features in Twitter facilitate information exchange surrounding the 2012 Presidential Election?

Contributions to Knowledge

The 14 findings from this study make three significant contributions to the existing body of knowledge. The findings from the first research question (findings 1-5) demonstrate that individuals use syntactical features in Twitter including retweets, at-replies, at-mentions, URLs and hashtags differently in the context of an acute event such as a debate compared to other times during the election. This contribution is unique to this study given the nature of the dataset and the longitudinal collection of both acute and non-acute event related discourse. The findings from the second research questions (findings 6-9) illustrate that although candidates and media are the most talked about and talked to, these social media interactions elicit no response. Further, there are no repeated interactions among citizens using the election specific syntactical features. This finding illustrates that even though technologically facilitated interactions are possible, there is little adoption of these features for conversational interactions.

Finally, the findings from the third research question (findings 10-14) reveals that information shared through URLs was predominantly user-generated content and mass media information and that most of the information that was shared was done through retweets. Retweets allow users to share information
with ease as all a user needs to do is click on a retweet button on the website or in the application to share the original tweet to his or her followers. This illustrates that the activity is within Twitter is reflexive and propagated using technological means. Although the inclusion of external content is possible, it is less common than information from within Twitter, suggesting that the Twitter environment is an “echo chamber” (Pariser, 2011). These findings help to contribute to the understanding of political discourse in social media and how technologically mediated interactions are constructed using technology specific syntactical features.

**Limitations**

This study examines a Twitter dataset collected pertaining to the 2012 United States Presidential Election. The theoretically informed selection of terms as they emerged during the election allowed for a theoretically informed snowball sample that afforded a more complete understanding of the activity in Twitter. The use of hashtags, at-mentions and keywords as selection criteria allowed for comparative analysis of the type of syntactical features used. Even with the richness of the data and the theoretically informed collection and analysis, this study has two limitations worth noting.

The first limitation is the unknown nature of all the tweets during that timeframe. The TwitterZombie infrastructure (Black, Mascaro, Gallagher, & Goggins, 2012) provides a documented collection platform, but during acute events it is likely to have missed a percentage of tweets. This percentage is unknown since Twitter does not publicly provide details about the number of tweets that fit into specific criteria. Therefore, it is impossible to triangulate the number of tweets collected using TwitterZombie and the actual number of tweets during the studied time period.

These tweets are not believed to be different in content or syntax than the tweets collected as part of the dataset since a sample of the tweets during the time period were still collected, just not at the rate that they were produced. This gap in data is not unique to this study and is a limitation that is part of many Twitter studies identified throughout the literature review in chapter two. The difference with this study is that the collection platform, method and terms used to construct the dataset are detailed to allow for an

---

5 Twitter has recently started a program to allow researchers to apply for a grant to get historical data for research from the company. At the time of this study, this program was not in place and the “awards” are not going to be given until the middle of 2014. Therefore, this limitation may be overcome in many future Twitter studies and is an area of future research.

6 As discussed in chapter 2, some activity garnered hundreds of thousands of tweets per minute. This rate of collection is impossible for anyone external to Twitter to collect without privileged access by the company.
understanding of the collection approach whereas the methodological approach is often obfuscated in other studies.

The second limitation that is inherent in this study and all studies of social media is the lack of understanding of the user population that is being studied. There has been some research (Keeter, Horowitz, & Tyson, 2008; Lenhart, Purcell, Smith, & Zickuhr, 2010; Rainie, 2011; Rainie & Smith, 2012) that has attempted to categorize social media users, but it is unknown the true demographics of the individuals that tweeted and were part of this sample. Further, many individuals maintain multiple social media presences and may participate differently in each technology differently depending on their network and persona they have chosen to take on in that specific social networking technology. Therefore, the findings of this study cannot be generalized beyond social media and most likely not beyond Twitter.
CHAPTER 2: LITERATURE REVIEW

Social media is unlike any technology used to facilitate interaction in the past. Even though social media is radically different than previous communication technologies such as the newspaper, telephone, radio, and television, there are still theoretical frameworks that can be used to interpret the activity occurring in social media. Technologically mediated communication research in the past has focused on synchronous activity in chat rooms, instant messages or comments and replies in discussion forums (Fiore, Lee-Tierman, & Smith, 2002; Fisher, Smith, & Welser, 2006; Kelly, Fisher, & Smith, 2005; Viegas & Smith, 2004). Twitter’s technological specific syntactical features and large user base provide a richness of discussion that gives users a greater ability to exchange information in distinctly different ways depending on the purpose of the exchange. This contrasts the technology of social media and Twitter specifically from earlier implementations of communication technologies. The archival nature and semi-synchronous ability to exchange information resembles both a chat room and a discussion board with millions of individuals posting approximately 350 million posts (tweets) a day.

Theoretical Frame

There is a long history of political science theories of democratic participation that help to further understanding the new technological environment. These theories were established throughout the 20th century as new technology was developed and adopted by the public. Although these theoretical frameworks are decades old, they are still applicable to the current technological environment.

The absence of political participation

In Lippmann’s seminal work from over 85 years ago, The Phantom Public (Lippmann, 1925), he argues that the ideal of the public as being present in the political sphere is false. Lippmann argues that the public is a “phantom” and that traditional democratic theory that argues that public opinion influences elected leaders has many flaws. Lippmann argues that the public is mostly absent in politics although they often participate in elections. This electoral participation is often uninformed as the public is unclear about what they may be voting for since many electoral issues are framed by political insiders and candidates and this framing may limit true understanding of what is at stake. In the context of political discourse, these events (elections in this case) may lead to greater excitement and different types of discourse than on a
normal basis. Therefore, it is important to examine discourse longitudinally beyond just the time period surrounding the actual Election Day to develop a better sense of how the public feels and the type of information they are interested in.

John Dewey responded directly to Lippmann’s argument two years later in his book, *The Public and its Problems* (Dewey, 1927). One of the principal arguments made by Dewey is that the public is not absent nor a “phantom” as Lippmann argues, but instead lacks cohesion and is distracted by the technological and societal improvements of the day such as entertainment. The public’s interest in non-civic activities detracts from their interest in politics. Dewey argues that technological improvements can lead to greater communication that may help the public (or multiple disconnected publics) find cohesion and a greater interest in politics. This greater interest can translate to more engagement and involvement in political activity that would bring the public out of its “phantom” state.

A more recent analysis of democratic theory comes from Jurgen Habermas and his concept of the bourgeois public sphere (Habermas, 1984; Habermas, 1991; Habermas, 1989). Habermas argues that transformations in history led to the ability for the public to engage in critical discourse that put into balance the power of societal leaders. The public sphere emerges from the congregation of individuals in publics that can form opinions representative of the individuals that comprise them. Habermas notes specific technologies starting with the 15th century printing press through newspapers, radio and television as the media of the public sphere in his earlier writings (Habermas et al., 1974). The nature of the public sphere as defined by Habermas may be changing, with the introduction of the Internet.

Although Habermas identifies the public sphere as one entity, there are many issues that are discussed within the public sphere. These issues are of interest to many different individuals and as a result of the narrow context of some of these topics, “issue entrepreneurs” emerge (Agre, 2004; McCarthy & Zald, 1987). In addition to these issue entrepreneurs, these issues form interlocking spheres that have also been identified as “subaltern counterpublics” by others (Fraser, 1990). Fraser identifies these counterpublics as publics that are often shut out of the public sphere as a result of their minority status and provides the example of the late-20th century U.S. feminist movement. These counterpublics expand the political discourse space by creating a sphere for further discourse. This ability is especially the case in electronic communication as the ability to create alternative discursive spaces is significant.
Moving from the public sphere online

Recently scholars have attempted to apply Habermas’ concept of the public sphere to electronic communication (Dahlberg, 2001; Robertson, Vatrapu, & Medina, 2010). Dahlberg (2001) argues that in order for the public sphere as conceptualized by Habermas (1984) to be realized online and for online deliberations to facilitate rational-critical discourse, certain social and technical requirements must be met. These requirements include:

- “Exchange and critique of reasoned moral-practical validity claims”
- Reflexivity
- Ideal role taking
- Sincerity
- Discursive Inclusion and equality
- Autonomy from state and economic power

At the time of Dahlberg’s initial research no public, virtual space that met these requirements existed. Political theorists supported Dahlberg’s focus on the importance of enabling deliberative discourse through technology; for example, they established criteria for effective political deliberation and discourse that can now be achieved through social networking technology (Fishkin, 1997; White, 1989).

Scholars such as Zizi Papacharissi have examined how social media and new technology affects political discourse. The “virtual sphere” as Papacharissi titles the Internet has great potential for facilitating discourse, but with this potential brings about the possibility of easier discursive fragmentation and the adaptation of the technology to the existing discursive environment (Papacharissi, 2002). Further, technologically mediated discourse and information brings about access issues that may disenfranchise and limit access by some.

Papacharissi (2004) has examined the Internet’s ability to foster political discourse and provide a virtual public sphere (Papacharissi, 2004). She identified that conversations on newsgroups can be impolite yet still civil. When the conversations became uncivil and as a result harmed civic engagement attempts, such activity was viewed in a negative context. Even with the possibility for the Internet to foster a virtual public sphere, Papacharissi further examines how more recent technologies contribute to the creation of a
“private sphere” where one’s activities in technology are focused on themselves even in the context of politics and civic involvement (Papacharissi, 2010). Twitter is one such technology as individuals are able to create their own environment.

Extending on Papacharissi’s work, Freelon (2010) has proposed a framework to examine online political discourse that is grounded in traditional democratic theories (Freelon, 2010a). Through an extensive review of both political literature and socio-technical literature, Freelon identifies “indicative metrics” of three of the most common types of democratic communication: liberal individualist, communitarian and deliberative. Freelon proposes that each of the three models of democratic discourse and their constituent metrics help to establish a framework with three models that can be applied to understanding specific types of discourse in a system. This bridging of two significant and important literatures helps to extend and allow for comparative analysis of discourse in socio-technical systems.

Technological systems that foster or expose individuals to discourse such as those studied by Papacharissi, Dahlberg and Freelon along with other scholars (Foot & Schneider, 2002; Freelon, 2010b; Gonzalez-Bailon, Kaltenbrunner, & Banchs, 2010; Robertson et al., 2010; Wojcieszak & Mutz, 2009), bring about significant promise for the realization of democratic discourse online. Even though there has been progress in facilitating democratic discourse on the Internet and an interest from the public there is still technological development (Walton, 2007) that needs to occur. This is evident in the discussion of the findings of this study in chapter 5.

Tying it together: applying theory to the socio-technical environment

Socio-technical systems allow for the study of the criteria that Dahlberg established along with identifying the existence of other types of discourse and participation by the public in the political process as identified in Freelon’s (2010) framework. Although the access to the technology may help to facilitate more discourse, this discourse may not be constructive – though with new models that allow for systematic analysis of this type of discourse this type of activity can be more clearly analyzed. The writings of Lippmann, Dewey and Habermas touch on three areas of democratic theory, but when applied to the context of the Internet and a specific technology such as Twitter, it is possible that the writings may be used as a theoretical framework to further an understanding of how multiple iterations of technological innovation since the early nineteenth century have all identified similar societal problems.
Lippmann’s concept of the phantom public can still exist even with such widespread communication technologies that allow the public to engage with anyone they wish. Even though engagement and discourse are easy to undertake, this discourse may be shallow and not constructive, and thus does little to address the “phantom” nature of the public. As Lippmann argues, the public does participate when given the opportunity to place another party in control. Therefore, in an election year it is likely that the public may not be as phantom as in other years.

On the other hand, Dewey’s notion of the need for better communication technology has been realized with the introduction of the Internet, but as these technologies facilitate better communication, they also facilitate more distraction. Therefore, it is possible that these technologies provide more outlets for individuals to participate in the public process, yet things such as entertainment, as noted by Dewey, may distract them. Finally, the notion of Habermas’ public sphere is also likely to be realized in the context of social media and with the presence of certain criteria established by scholars such as Dahlberg.

The durable work of Lippmann, Dewey and Habermas and the recent additions to this body of literature by Papacharissi reflect three distinct theoretical frames for examining the relationship between the public and the democracy they ostensibly participate in. These theoretical approaches serve as a theoretical framework for understanding how citizens engage with each other and their elected officials in technologically mediated settings. Through analysis of Twitter activity in the context of an election in which the public is actively involved and the politicians appear to actively participate, these theories of engagement can be examined in the context of a large dataset that spans time and is geographically independent.

**The Internet’s Role in American Elections**

The Internet has played a role in every election since its introduction to the public in the 1990’s. This role has evolved with the technology, as campaigns have wanted to find an edge to allow them to further disseminate their message to a broader audience. As technology has allowed for more engagement, the use of the Internet has also evolved to where individuals are now able to get personalized updates from candidates and engage with them using Twitter.
The Internet is introduced: 1992 - 2004

One of the earliest studies of a computer network’s use in a Presidential election examined the “computer lists” of George H. W. Bush, Bill Clinton and Ross Perot (Hacker, Scott, Howl, & Steiner, 1996). These lists were similar to early bulletin board systems and discussion forums. In this study, the researchers found that the campaigns employed different strategies. The Clinton campaign used the computer lists to post more information whereas the Perot campaign used it to assert opinions. The most significant finding of this research was that it illustrated the legitimacy of using technology in national level campaigns.

The first widespread use of the Internet in national level politics in the United States began in 1996, when the Internet began to be available to the mass public. In 1996, most candidate websites were merely a virtual campaign brochure with basic information about the candidate and limited information about candidate positions on policy issues (Dulio, Goff, & Thurber, 1999). As the technology has matured, so have the strategies that have been employed on the Internet. It was not until the 1998 midterm election cycle that political candidates started to harness the power of the Internet and began to focus on developing strategies to succeed online.

In 1998, more than two-thirds of candidates for Congressional seats had websites. Of these websites, more than 70 percent of the candidates used the websites to perform some sort of fundraising activity (e.g., allowing supporters to make credit card donations on the website) (Dulio et al., 1999). The year of 1998 marked the emergence of the Internet as a method to reach out to the public. In 1998, Jesse Ventura used e-mail to attract supporters for his bid for the Minnesota Governor seat in what was one of the first uses of the Internet to actively target and mobilize a group of voters (Napoli, 2003). The ability for individuals to donate online did not lead to a quick shift to online donations, but it introduced the concept that would be built upon in years to come.

In the 1998 race for the California Senate Seat, less than 1 percent of all campaign donations were made online (Dulio et al., 1999). This figure is limited in contrast to the hundreds of millions of dollars being raised online today, but it presaged the future of campaigning on the Internet. Even though most candidates in major elections had a website in 1998, the benefits of having an online presence were not realized because Internet access was not as ubiquitous as today. As Internet access entered more homes in
the early 21st century, candidates realized the need to develop a more attractive web presence. The elections of 2000 were the first Presidential election cycle where the Internet played a role in fundraising and mobilizing support. John McCain carried out the first success of fundraising on the Internet in 2000, when he raised more than $5 million online (Napoli, 2003).

The widespread proliferation of the Internet and political candidates’ adoption of the Internet as a campaign tool meant that people were turning to the Internet for political information. During the 1998 and 2000 election cycles, potential voters were asked what they used for primary media source for election information. From 1998 to 2000, the number of respondents indicating that they used newspapers as their primary source of information dropped by 14 percent, while the number claiming that they used the Internet increased by about the same percentage (Lupia & Baird, 2003). This statistic illustrates the shift to the Internet for political information. This phenomenon would be evident in the 2008 election, as studies have indicated that at least 46 percent of Americans used the Internet to obtain a large amount of election news during the primary season (Smith & Rainie, 2008).

The 2000 election was the first election in which the Internet was widely used to disseminate information and accept campaign donations. Even though individuals were using the Internet more to participate in political activity, the content disseminated by candidates was still limited. From 2000-2004, the Internet began its transformation from static to dynamic content and with the transformation came the emergence of participatory media such as social networking sites and blogs. These technologies would usher in an age that allowed candidates to conduct endless amounts of mobilizing and help to build communities online and offline.

In the 2004 Presidential Election, Howard Dean was the first presidential candidate to use blogs to provide dynamic content. The blogging platform also facilitated discourse among supporters (Trippi, 2004; Williams, Trammell, Postelnicu, Landreville, & Martin, 2005). The use of technology to engage with the public in the 2004 election, coupled with increasing use of Facebook by the general public in 2006, found the wider adoption of social networking sites for political purposes in the 2006 mid-term election (Williams & Gulati, 2007).
From static content to dynamic engagement: The 2008 Presidential Election

The 2008 Presidential Election built on earlier efforts of using the Internet through the utilization by all three of the Democratic primary candidates of social media to include Facebook, Twitter and other technology such as e-mail. President Barack Obama was the first Presidential Candidate to fully embrace the Internet and had a dedicated staff of 200 individuals working on social media strategy and engagement. These efforts led to the collection of over 3 million supporters on his Facebook page along with over 15 million other supporters on other platforms.

All three of the major candidates in the 2008 US Presidential Primary and General Election, Barack Obama, John McCain and Hillary Clinton, depended on social media for mobilization and fundraising. Immediately following the November 2008 election victory, Barack Obama’s Facebook page had over 3 million supporters. He also had five million supporters on 15 other social networking sites, including Black Planet, a social network focused on the African American Community (Vargas, 2008). At that time, Obama was also number one in Twitter followers with over 100,000. In addition to the use of social networking websites, Obama’s campaign sent one billion messages to a list of 13 million supporters that had been amassed from campaign rallies and other online activity (Borins, 2009).

Foundational research that has examined the Facebook walls of the 2008 Presidential Candidates surfaced characteristics of the specific posts and patterns of activity among participants (Robertson et al., 2009; Robertson et al., 2010). These findings illustrated that SNS’s are shown to approximate a socio-technical representation of Habermas’ (1984) public sphere. More generally, SNSs provide unique avenues for political discourse. This work focused on the technical capabilities and potential of Facebook to support such discourse, but not on the network of individuals involved, the extent of their involvement in political discourse or how this network evolved over time.

Obama placed the Internet at the center of his campaign’s information, communication and coordination infrastructure. In the 2008 election, at least 46 percent of Americans used the Internet to get a significant amount of election news during the primary season (Smith & Rainie, 2008). Further, 40% of Internet users with profiles on social networking sites used them to receive political information or participate in political activity. More specifically, half of those under 30 used their social networking accounts for political activity. Most of the political activity included attempting to mobilize support and the
sharing or seeking out of information from individuals on the site. Obama’s election is often attributed to his community organizing experience in the physical world, but his campaign’s skillful use of social media enabled him to build up grassroots support more rapidly than what was witnessed in previous campaigns (McGirt, 2009).

In addition to Presidential campaigns, Congressional members increasingly used social networking sites such as Facebook and Twitter to engage with the public. Websites such as tweetcongress.org have publically encouraged members of Congress to utilize Twitter to further civic engagement in the legislative process (Netherland & McCroskey, 2010). The increasing usage of Twitter by Congress and the perceived campaign benefits led to a dramatic adoption of technology for purposes of civic engagement in the 2010 election (Schaper, 2010).

The 2012 Presidential Election

The 2012 Presidential Election provides a unique context for the use of technology and social media as many individuals had previously identified that they used technology for political purposes in the 2010 election. The larger availability of technology and increased public reliance on technology in 2012 would help to demonstrate unique uses of technology by both the candidates and citizens.

The 2012 Presidential race effectively began following the April 25th declaration by the RNC that Mitt Romney was the presumptive nominee after a long set of primaries. Paul Ryan was officially named Governor Romney’s Vice Presidential candidate on August 11, 2012 just over two weeks before the start of the Republican National Convention (RNC) (Hunt, 2012). The VP announcement drew significant media excitement and the announcement came with the unveiling of Paul Ryan’s Official VP Twitter account (@PaulRyanVP). During the announcement of Paul Ryan as his VP candidates the tweets per minute referencing Paul Ryan were close to 3,800 (Twitter, 2012).

<table>
<thead>
<tr>
<th>Table 1: Prominent 2012 Electoral Events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dates</strong></td>
</tr>
<tr>
<td>August 27 - August 30, 2012</td>
</tr>
<tr>
<td>September 4 - September 6, 2012</td>
</tr>
<tr>
<td>October 3, 2012</td>
</tr>
<tr>
<td>October 11, 2012</td>
</tr>
<tr>
<td>October 16, 2012</td>
</tr>
<tr>
<td>October 22, 2012</td>
</tr>
<tr>
<td>November 6, 2012</td>
</tr>
</tbody>
</table>
The Republican National Convention occurred two weeks following the announcement of Paul Ryan as the vice presidential nominee and kicked off the general election season (Table 1). There was an official convention account (@gopconvention) and also an official hashtag (#gop2012) promoted by the Republican Party Account (@gop). The Republican National Convention had numerous prominent Twitter moments including the public’s creation of the parody account, @InvisibleObama, after Clint Eastwood gave a speech where he spoke to an empty chair (pretending that President Obama was sitting in it). This account received over 20,000 followers in just 45 minutes and was mentioned over 10,000 times in that time period.

The biggest moment of the convention occurred near the end of Romney’s speech when he formally accepted the nomination when there were over 14,000 tweets per minute (Sharp, 2012a). In total, Twitter estimates that there were over four million tweets related to the RNC that occurred during the 4-day convention.

The Democratic National Convention occurred a week later and also had an official account (@demconvention) and hashtag (#dnc2012). In total, Twitter estimates that there were 9.5 million tweets related to the event (Sharp, 2012c). The focus of the convention was President Obama’s speech that at one point had over 43,000 tweets per minute referencing the speech. This was more than 3 times the maximum number of tweets per minute during Romney’s speech.

This difference in tweet count illustrates a contrast in the activity for the two conventions. There were over twice as many tweets for the DNC than the RNC and the tweets related to the DNC were more focused on the acceptance speech. In fact, Vice President Biden’s acceptance speech on the penultimate night of the convention peaked at approximately 18,000 tweets per minute illustrating that there was a greater interest in the speech of the Vice President as compared to the Republican Nominee.

The conventions kicked off a two-month race to the November 6th, 2012 election with the four most prominent electoral events being the debates. There were three presidential debates and one vice-presidential debate. The first presidential debate occurred on October 3rd, 2012 at the University of Denver and focused on domestic issues. The second presidential debate was on October 16th, 2012 at Hofstra University in New York and focused on a mix of issues in a town hall format. The final presidential debate was held on October 22nd, 2012 at Lynn University in Florida and focused on foreign policy and security.
issues. The vice-presidential debate occurred on October 11th, 2012 at Centre College in Kentucky and focused on a broad range of topics in a traditional debate format. Although there was a number of other third party candidates running none were polling high enough to be included in these debates and none received any substantial portion of the vote.7

Twitter estimates that approximately 28 million tweets were sent related to the debates, with the first debate being the most popular with close to 10 million tweets. The vice-presidential debate was the least popular debate on Twitter with approximately 4 million tweets.8 The significant number of tweets across all of the debates illustrates the excitement surrounding the debates. The first debate was identified by Twitter as the most tweeted about event in U.S. Politics until Election Day, which had over three times as many tweets.

The amount of activity on Twitter on Election Day was large. Network television called the race around 11:15 PM EST on November 6th and Romney conceded at 1:00 AM EST on November 7th. President Obama won the election with 332 electoral votes (65,907,213 popular votes) to Governor Romney’s 206 electoral votes (60,931,767 popular votes).

When the networks called the election there was a spike of activity on Twitter that was measured at a peak of 327,452 tweets per minute (Sharp, 2012b). Following the declaration of Obama’s victory, President Obama tweeted a photo of himself and First Lady Michelle Obama with the caption “Four More Years.” This tweet was retweeted approximately 800,000 times in twelve hours.

The high number of tweets on Election Day (approximately 31 million) illustrates the excitement and prominence of the event in social media and identifies an area for study beyond just tweets per minute. The significant amount of activity warrants an exploration of the type of connections, syntactical features and recent political work on Twitter to form a foundation for the following study.

**Twitter as Technology**

Twitter is unlike any other social media platform. Its 140-character message maximum requires minimal effort on behalf of the user to contribute to the discourse. The asymmetric user relationships also allow for more transient network structures. This transience creates a social network that differs in shape to

---

7 No third party candidates were included in the analysis since none were polling at any number to be included in a debate they were considered to be inconsequential in the overall election.

8 This statistic was created by combining the four figures provided by Twitter on their blog recapping the debates.
other social networking platforms because reciprocity is not required to form a tie (Golder & Yardi, 2010). Even with this simplicity, Twitter is powerful in the number of syntactical features it offers. Within the 140-character “tweets”, individuals can reply to others, mention others, include hashtags, retweet information from others, and include links to pictures, videos or to other websites. Academic research on Twitter is plentiful since the company provides three API’s that provide large amounts of data to the research community. As a result of the access to data, Twitter research has evolved and increased from the time when Twitter was first introduced in 2006.

**Establishing a baseline understanding of Twitter**

Early research on Twitter activity attempted to understand user behavior (Krishnamurthy, Gill, & Arlitt, 2008) and user intent for using certain syntactical features. This research established a “baseline” for understanding how individuals behaved in a new socio-technical environment. Findings from these early have limited application to understanding today’s environment as the technology and user base has evolved. The most recent research on user syntactical feature usage has found differences between organizations, journalists/bloggers and ordinary people in how they utilize Twitter in both event specific and news that is of wider appeal to the public. De Choudhury et al. (2012) found that organizations and journalists tend to retweet more and include more URLs in their tweets. Further, ordinary users are more likely to utilize at-mentions when discussing a specific event such as a concert. The categorization of users allows for an understanding of how user roles may change depending on the context (De Choudhury, Diakopoulos, & Naaman, 2012). Other research has identified that the extremeness of an event is a predictor of whether someone will tweet about it (Kıcıman, 2012).

The evolution of user syntactical feature usage is likely attributable to the content and types of users (early adopters versus early majority) on Twitter. Early research found that individuals used Twitter to talk about their daily routine, to converse with others, to share information and to report news (Java, Song, Finin, & Tseng, 2007). In recent years, organizations have adopted Twitter as an information dissemination and engagement mechanism. More recent research has found that Twitter is useful for covering topics that are not widely covered in traditional media and allows for coverage in traditional media to be augmented since Twitter users are active at spreading breaking news (Zhao & Jiang, 2011).
In an attempt to establish a new baseline for the current environment, research has attempted to build upon earlier findings by developing a further understanding of how Twitter specific syntactical features such as hashtags (Huang, Thornton, & Efthimiadis, 2010; Romero, Meeder, & Kleinberg, 2011), retweets (boyd, Golder, & Lotan, 2010; Starbird & Palen, 2012), and URLs (Hughes & Palen, 2009; Java et al., 2007) are used. In contrast to previous technologies where participants can control or know their audience, Twitter’s public nature creates different perceptions of the audience that the tweets are intended for and as a result this creates a “context collapse” where multiple audiences become one (Marwick & boyd, 2011).

**Twitter’s social connections**

Like any social networking technology, the fundamental activity on Twitter is to create relationships in order to share or consume information. Users can follow each other, be followed by another individual or create lists of individuals to follow. A follower relationship is asymmetrical and therefore there is no need for the followed person to agree to the follower unless one of the users has a privacy setting enabled – something used by few Twitter users. Once an individual follows another, the tweets of the followed individual are included in the following users’ stream of tweets. The asymmetry of this relationship allows for numerous combinations of relationships, but makes traditional network analysis of relationships on Twitter more difficult as ties can signify a number of things compared to other social networks where ties are more likely to represent a real world connection (Golder & Yardi, 2010).

There have been limited studies that have examined the nature of relationships on Twitter. Kwak et al. (2010) crawled the entire public Twittersphere in 2009 and found that only 22% of relationships were reciprocal (Kwak, Lee, Park, & Moon, 2010). This asymmetry is likely attributable to the use of Twitter as a mechanism to get daily news feeds by following journalists or other news organizations. Although the numbers from Kwak et al. (2010) are currently a few years old, current user behavior is likely to be similar to that in 2009 as many users still utilize Twitter as an information consumption technology.

Relationships in Twitter are asymmetric and studying these links can be used to understand network position and how the establishment of these relationships factor into being influential or popular. Measuring the number of followers an individual has (in-degree influence) illustrates popularity, but does

---

9 The statistics for how many Twitter users employ a security setting to hide their tweets is currently unknown.
little to identify the ability of these users to influence others (Cha, Haddadi, Benevenuto, & Gummadi, 2010). Geographic analysis of the ties of Twitter users illustrates that many existing ties are geographically specific, but that factors such as common air travel between two locations can be attributable to the ties between individuals that exist over distance (Takhteyev, Gruzd, & Wellman, 2012). These links may be attributable to a number of factors that include corporate activity between cities with high air travel and shared cultural values and languages. Additionally, some research has illustrated that ranking users based on the number of followers, network statistics such as PageRank and retweet frequency can have different outcomes in identifying who is more influential (Kwak et al., 2010). This illustrates that influence and popularity on Twitter are context and measurement specific.

**Twitter syntactical features**

Although previous research examined how relationships on Twitter can establish network position and possibly influence others, I extend the research by looking at interactions and discourse behavior surrounding specific events. I do not focus my analysis on existing follower/followee relationships and instead focus on how syntactical features create and shape discourse surrounding an event or series of events. Through analysis of the specific syntactical features in Twitter, that I now operationalize and discuss, I am able to identify and characterize user behavior differently than just analyzing follower/followee relationships. This analysis coupled with longitudinal analysis of a series of events help to address gaps in the literature.

Political discourse begins with some kind of technological transport mechanism or platform. The printing press, broadcast media and newspapers are historical mechanisms for the support of discourse. With social media, the communication of ideas is more interactive and diffuse, both geographically and topically. The different syntactical features of each technology constrain and enable political discourse, and for those reasons it is important to understand the menu of interaction types in the media.

Twitter has many syntactical features that allow for a richer social media experience. These syntactical features include: at-replies, retweets, URLs, at-mentions, hashtags and direct messages. Most of these features such as the hashtag emerged organically from the user base and have evolved through varied usage by different sets of users. Therefore, each syntactical feature has a unique and varied use that requires specific conceptualization in different contexts to properly research and compare the findings from
different datasets. Table 2 illustrates the commonly used Twitter syntactical features with a description of each. Following the table, I review the existing literature that has studied the utilization of each one in specific contexts. As illustrated by the variety of studies, the findings of a study that examines Twitter activity in the context of syntactical features are highly dependent upon context and therefore it is imperative that the context be understood. It is for this reason that I narrowly define my conceptualization of Twitter’s syntactical features. Though there may be many other conceptualizations, my review of the literature identifies these as the most common and accepted by Twitter users.

Table 2: Syntactical Features Enumerated

<table>
<thead>
<tr>
<th>Syntactical feature</th>
<th>Syntax as applied to this study</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>At-Reply</td>
<td>@[username] at first position of tweet text</td>
<td>To directly address another individual in a public manner</td>
</tr>
<tr>
<td>At-Mention</td>
<td>@[username] at any point in tweet text</td>
<td>To highlight a tweet to another individual or to talk about someone. Mentioning them will inform them of the tweet</td>
</tr>
<tr>
<td>Retweet</td>
<td>RT @[username] “tweet text”</td>
<td>To further disseminate another individual’s tweet.</td>
</tr>
<tr>
<td>URLs</td>
<td><a href="http://t.co/%5B6-10">http://t.co/[6-10</a> characters]</td>
<td>To include external information in a tweet. Note: Twitter uses a URL shortener</td>
</tr>
<tr>
<td>Hashtags</td>
<td>#[alphanumeric text]</td>
<td>To tag a message with a conversational marker or to add a tweet to an existing stream of discourse independent of a follower/followee network</td>
</tr>
</tbody>
</table>

The nature of retweets

Retweets have been one of the most studied syntactical features of Twitter. A retweet is the act of an individual copying another individual’s tweet into one’s own Twitter stream for distributing to that individual’s followers. The nomenclature for retweets is ever evolving and there is not a standard agreed upon mechanism for retweeting, although Twitter has recently added a retweet button that allows an individual to retweet a message using a common format. The common format for retweets is “RT @[original username] | [original tweet].”

Even though there is a common format for retweets there are many other ways to retweet messages and there are other syntactical features that have been created by users such as “MT,” which stands for “modified tweet.” A modified tweet is a form of retweet in which the user usually shortens a message to add commentary. Analysis of
One of the problems with the analysis of retweets is the practice of editing these tweets and then retweeting them. This can alter the meaning of the original tweet and requires further analysis of the context of the tweet in the discourse. The editing of the tweets in the retweets also makes finding the original tweet difficult in large amounts of data. Retweets are used for a variety of purposes, but mostly focus on information sharing. A retweet network can be used as measure of information dissemination and the network surrounding such an action could be used for the identification of information legitimization or to influence another individual.

The practice of retweeting has been examined to identify user intentions and the type of information a user retweets (boyd et al., 2010). There are many reasons that individuals retweet which includes 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 increase their twitter followers (Stieglitz & Dang-Xuan, 2012). In addition to these purposes, retweets are seen as a mechanism of conversation that embodies different characteristics depending on the user’s network and the content of the original tweet.

The number of tweets that get retweeted varies depending on the dataset and domain. In analysis of representative samples of the overall Twitter stream, it has been found that 16% of tweets on a daily basis on Twitter are retweets (Mustafaraj & Metaxas, 2011). In a more focused analysis of datasets related to crisis events and political discourse, the number of retweets was found to be higher, ranging from 42%-56% of the total dataset (Mustafaraj & Metaxas, 2011). These statistics are highly variable and illustrate just one of the ways that drawing conclusions about socio-technical behavior in the context of one dataset may lead to different conclusions, especially in an environment such as Twitter.

The timing of when a tweet gets retweeted depends on the content of the tweet and also on the network of users that the original author has. Gruzd et al. (2011) found that positive messages are more likely to be retweeted than negative messages (Gruzd, Doiron, & Mai, 2011). This illustrates a different intent and purpose of retweeting and information diffusion. In a comparison of Twitter with other social news sites such as Digg, research has found that the less dense network means that Twitter stories initially spread slower than in a dense network, but they persist through time and spread further through the network (Lerman & Ghosh, 2010).

---

data captured using TwitterZombie has identified that “modified tweets” are rare (< .5% in most datasets) and as a result are not examined in this study.
Research has identified that retweets have more URLs, and hashtags, indicating that tweets with more structured information and information that is external to Twitter may be more likely to be retweeted (Suh, Hong, Pirolli, & Chi, 2010). Counter to intuition, users that followed more people were retweeted more often by their followers. This is likely the result of the user being exposed to more types of information and then being able to pass that along to their followers. This finding illustrates that retweets may be an indication of information diffusion. Finally, retweets tended to come from users that were more ‘veteran’ Twitter users indicating that long term reputation and familiarity with others may contribute to an individual being retweeted.

Twitter has also been instrumental in communication during crisis events and natural disasters. Retweets have played a significant role in these types of events. 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 (Starbird, Palen, Hughes, & Vieweg, 2010). In addition to information exchange during a natural disaster, researchers have attempted to use what has been learned from disaster information exchange to be able to detect disasters in near real-time through “sensors” (other Twitter users) (Sakaki, Okazaki, & Matsuo, 2010). Notification of disasters as a result of these systems is believed to be quicker than official notifications, although they are susceptible to difficulty in detecting valid threats.

**Conversational activity on Twitter**

An at-reply is the presence of an at-mention at the beginning of the tweet text. This syntactical feature allows users to direct a message towards another individual and to create a public conversation. Through network analysis of at-reply messages it is possible to identify conversational discourse similar to that which can be done with email (Klimt & Yang, 2004a; Klimt & Yang, 2004b). The inclusion of at-mentions is less studied than the other syntactical features, but there is a lot of information that can be derived from the inclusion of specific at-mentions. An at-mention is the inclusion of a username (similar to a at-reply) that is not at the first position of the tweet text. This inclusion of a username can be done to alert others of the presence of a certain tweet since when an at-mention is included; other individuals are technologically notified of the presence.
Honeycutt and Herring (2009) apply the construct of addressivity from earlier forms of computer-mediated communication (Werry, 1996) to the Twitter environment (Honeycutt & Herring, 2009). They identify the syntactical feature of @, which directs a message towards someone else, as a form of addressivity 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 since there was the initial intent of interaction as a result of the presence of @.

Honeycutt and Herring (2009) also found few instances (~3%) where another person was intended for a message, but there was no @ directing that message towards them. In these instances, the researchers searched for usernames to identify the intent of directing the message to someone else. This finding runs counter to other studies of large-scale computer mediated communication in other socio-technical environments where the utilization of @ is much more limited (Mascaro, Novak, & Goggins, 2012b). The researchers also found that those messages that were initially directed towards someone received a response 31% of the time. This is higher than previous studies of technologically mediated communication, but does not take into account the possibility that a reply may have been sent in another channel. This would make the total number of responses higher although they would not be reflected in the same data (Zelenkauskaite & Herring, 2008).

Analysis of the aftermath of the shooting of a late-term abortion doctor in Kansas found that individuals with similar viewpoints were more likely to engage conversationally with each other in the context of a controversial event (Yardi & Boyd, 2010a). Even though individuals were more likely to engage with others with a similar viewpoint, about a third of the reply-pairs identified were between individuals with differing viewpoints. This engagement ranged from confrontation to defensive regarding the actions of the shooter and how pro-life individuals did not agree with his actions. Finally, in the context of information diffusion, it was discovered that individuals frequently retweeted news sources, but even though some individuals attempted to engage with these news organizations, few responded.

In a study that built on the earlier work by comparing the shooting with the collapse of a parking garage in Atlanta (both relatively local events, but one with a more controversial wider appeal) Yardi and
boyd (Yardi & boyd, 2010b) found that the network of activity in Atlanta was denser than Wichita as a result of the politically divisive nature of the Wichita event. Further to the point, the more central Twitter users tended to be more geographically focused near the event, especially in the case of the parking garage collapse. This illustrates the difference between events of wide appeal and local appeal and how conversational activity can identify the differences in these types of events. Geography does not play as significant of a role in national electoral discourse, but this type of finding may be important for Congressional and local elections.

Though relationships are important in understanding initial information diffusion, collecting and characterizing actual user behavior such as retweet behavior, the number of mentions and at-replies and the content of tweets are much better indicators of influence or conversational behavior (Bakshy, Hofman, Mason, & Watts, 2011; Bigonha, Cardoso, Moro, Almeida, & Goncalves, 2010). This influence is also something that does not occur quickly and is often the result of extended interaction or reputation in the network (Cha et al., 2010). To further examine at-reply networks Querica et al. (2012), used follower/followee networks in combination with one-way and two-way reply networks to support earlier research that at-replies signify stronger relationship ties (Huberman, Romero, & Wu, 2008; Quercia, Ellis, Capra, & Crowcroft, 2012). Further analysis of public discourse that uses the at-reply is important in understanding a shift in how individuals are publicly communicating about numerous topics.

**Hashtags as conversational markers**

Hashtags have many meanings and are a way for a user to identify a tweet as being part of a larger stream of discourse. Users first implemented Hashtags soon after Twitter was introduced. They borrow many characteristics from tagging behavior on blogs and websites. Hashtags in Twitter are used for fundamentally different purposes than tagging in the past. In the past, tagging artifacts was used for archival purposes and for easy retroactive access. Tagging through the use of hashtags in Twitter helps a user identify and find topically relevant streams of discourse (Huang et al., 2010). The ability for hashtags to be used as markers of discourse allows for the existence of “searchable talk” where users employ hashtags as an affiliation to be associated with certain streams of discourse (Zappavigna, 2011). These hashtag communities are emergent and can traverse multiple topics and help users go outside existing network structures to find others to engage with (Lumezanu, Feamster, & Klein, 2012).
Hashtags play an important role in identifying streams of discourse as they allow individuals to seek out or contribute to streams of discourse. Following certain hashtags can allow users to be made aware of a certain area of discourse that they would not be exposed to by just looking at the individuals that they follow. In this manner, hashtags form a topical community and network of interactions that differs from just the follower/followee networks that persist throughout Twitter on a daily basis (Rossi & Magnani, 2012). These networks expose users to others interested in similar topics and may lead to shifts in existing relationships as individuals may choose to follow more individuals or may gain more followers as a result of their participation in this new stream of discourse.

One of the limitations of using hashtags is that they require adoption and proliferation throughout the network. Since users create hashtags, their greater adoption by the Twitter population may be subject to network specific characteristics such as promotion by one individual or organization. Analysis of hashtag adoption has used the Diffusion of Innovation (Rogers, 1962) framework to understand the characteristics of successfully and unsuccessfully adopted hashtags as a new hashtag can be viewed as a form of topical innovation in the Twitter space (Chang, 2010). By following a specific hashtag a user can become part of an emergent group of discourse. Therefore, hashtags can help to establish personal and group identity, conversational discourse, or can just be used to identify streams of discourse to other interested individuals.

In crisis informatics, researchers have used hashtags to identify streams of discourse to both share information and gather information from witnesses on scene. Heverin and Zach (2011) used hashtags to identify tweets associated with three separate school shootings to understand how information diffused through the network and applied the framework of sense-making to understanding how individuals used this information (Heverin & Zach, 2011). Similarly, hashtags and other shared keywords have been used to identify other crisis situations so individuals could discover important information and share information around a specific topic outside the context of their network of existing relationships (Bruns & Burgess, 2011; Sreenivasan, Lee, & Goh, 2011; Sutton, Palen, & Shklovski, 2008).

Based on an analysis of political hashtag utilization, Conover et al. (Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer, 2011a) found that hashtag usage and network analysis based on retweets and mentions of users were reliable predictors of political affiliation of Twitter users. In some cases, where the affiliation was determined to be wrong or ambiguous, it was determined that “content
injection,” the inclusion of a hashtag in an unrelated tweet, was the reason for the mischaracterization. This illustrates how hashtag utilization may represent a form of group identity or promotion of an affiliation by a user.

One of the limitations with analysis of discourse surrounding a particular hashtag is that hashtag adoption is limited in some contexts. This lack of adoption may be the result of individuals not being aware of the existence of a hashtag or preferring to not use a # before a keyword. In an analysis of the tweets that contained the word flu during the swine flu epidemic of 2009, only 10% of the tweets had hashtags and only 1% were retweets (Szomszor, Kostkova, & St. Louis, 2011). This illustrates that certain types of tweets may be lacking in the adoption of certain syntactical features that would be helpful for identifying the discourse to the broader community.

Since hashtag utilization is limited in some contexts, researchers have analyzed tweet content using keywords. For example, Mathioudakis and Koudas (2010) have attempted to create infrastructure that would identify “bursty” keywords that would indicate new streams of discourse that may not have a hashtag associated with them (Mathioudakis & Koudas, 2010). The medical community has also studied Twitter as a way to understand if the public adopts certain terminology in the context of a pandemic (Chew & Eysenbach, 2010). In this research, the researchers were curious if the public used the terms “swine flu” or the formal term “H1N1.” The researchers found that over time, Twitter users adopted H1N1 as opposed to swine flu. The researchers attributed this finding to the suggestion by the World Health Organization to use this term.

Although hashtag adoption may be limited in many contexts, they still offer the best alternative to keyword and follower/followee analysis in that the hashtag represents discourse about a specific event or series of events that is intentional. On the other hand, keyword analysis may capture tweets that are not actually related to the event of interest. Therefore, using the hashtag along with at-mentions of individuals who are key players in the event is one of the best ways to get a dataset representative of the election that can be used as a focal point for analysis. Further, analysis of the hashtags that co-occur with each other may help to identify related events that can be used to provide further context to the study. These approaches are incorporated into this study.
**Information exchange through URLs**

URLs play an important role in sharing information online. The examination of URLs and hyperlinks on the Internet has been studied extensively (Park, 2003). In the political domain, analysis of hyperlinks has been used to assess the interactivity of campaign websites (Trammell, Williams, Postelnicu, & Landreville, 2006). Other research has identified hyperlinks as an indicator of campaign positions and links to sites that are external to a candidate indicate specific relationships (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003). In the United States and other countries, links between websites of candidates and elected individuals can illustrate shared political ideology and partisan connections (Kim, Barnett, & Park, 2010; Park, Thelwall, & Kluver, 2005; Park & Thelwall, 2008; Park & Perry, 2007).

In journalism, hyperlinking between newspaper websites has been demonstrated to facilitate the sharing of information and has lead to an increase in page views as audience size increased (Weber, 2012). Analysis of hyperlinks between media and political websites in Spain has also identified that the political orientation of media outlets can be identified by examining the candidate websites that they link to (Romero-Frias & Vaughan, 2012).

In the context of social media there has been limited analysis of URLs and how they pertain to interactions. Robertson et al. (2009) found that in Facebook there is a high concentration of domains that account for a majority of the URLs that are shared (Robertson et al., 2009). Further, these URLs were used for a number of purposes to include: evidence, rebuttal, action, joking and ridicule, and direct address.

Recent research on the spread of breaking news on Twitter found that a small number of organizations influence much of the discourse in the context of URLs shared. In the case of the death of Osama Bin Laden, these organizations fit into three groups, journalists, mass media organizations and celebrities. The type of information that these individuals shared varied, but in this case about 10% of all tweets related to the death of Bin Laden contained a URL and 26 websites accounted for close to 60% of the URLs shared. Approximately two-thirds of the sites that were linked to were mass media websites and the other third were to websites that hosted user-generated content (Hu et al., 2012). Other research has identified that tweets are deemed to be more credible if the base URL included in the tweet is from a known source (Castillo, Mendoza, & Poblete, 2011).
Shortened URLs are important in socio-technical systems that privilege the amount of characters. Research into shortened URLs has identified some interesting characteristics of their usage. In one study, researchers identified that approximately 80% of URLs in one technology were spam-related (Klien & Strohmaier, 2012). Other research has found that shortened URLs tended to point to two categories of information, high quality information and spam (Kandylas & Dasdan, 2010). Research to develop systems to identify these different types of URLs has had some success (Yang, Chitturi, Wilson, Magdy, & Fox, 2012). In many cases, shortened URLs that are not of high quality may point to malicious software hosted on websites. In one study, it was determined that 8 percent of 25 million URLs posted to Twitter in 2010 pointed to a malicious website (Devesa, Cantero, Alvarez, & Bringas, 2011).

One of the problems with shortened URLs is that the original source of information may no longer be linked to after some amount of time. The disappearance of URLs shared in social media is a topic that has only recently begun to be examined. Recent research has found that approximately 11% of the resources shared in social media no longer resolve after one year. This number increases by about .02% a day (SalahEldeen & Nelson, 2012). This research echoes earlier research of the Internet that found web pages change locations online or disappear as time passes (Koehler, 1999). Understanding the rate of decay and the types of URLs that disappear is important from the perspective of archiving and understanding historical interactions. In the context of Twitter, it is also important to understand what types of tweets that the different URLs are a part of and this analysis is done in the following study.

Politics and Twitter

Elected members and candidate use of Twitter

One of the most significant areas of Twitter research has been the analysis of political activity and political participation in Twitter. Political activity on Twitter has two components, the politicians, either elected or running in the election, and the citizens involved in the discourse or consumption of the information. The syntactical features of Twitter allow for unique interactions as citizens can mention politicians or directly address them. Mentioning and directly addressing elected officials represent different modes of engagement. Mentioning someone is intended to highlight a message to someone or to have the message show up in searches for that username. Utilizing the at-reply function indicates that the individual is directly addressing the politician. This is illustrative of an engagement between an elected official and
the citizenry. Similarly, elected officials or candidates for office can do the same, which creates an interesting and unprecedented dynamic in political communication, even though this engagement has not been identified at the national level as identified in this study.

Profile creation and identity are essential factors to engagement on social networking sites. In Canada, politicians have been found to establish separate identities from their party and party leaders to create a personal brand (Small, 2010). Even with the separate accounts the two groups use it to broadcast information to the public. Golbeck et al. (2010) had similar findings in that members of Congress in the United States used Twitter for self-promotion and information broadcasting as opposed to a tool for discourse between themselves and the public (Golbeck, Grimes, & Rogers, 2010). With an increasing focus on Twitter for both campaigning and governing there is still a lack of research that analyzes the different uses between how elected officials use it to engage with the public about official matters as opposed to campaign issues. This area is difficult to examine given the existing technological environment and the duration of collection needed.

The current status of the party in regard to the public may also specifically influence technology adoption. Partisan affiliation can affect adoption of technology as certain parties may feel the need to engage with citizens more and express their message further (Williamson, 2009). Analysis of political activity in South Korea has found that “resource-deficient politicians” (those with limited financial means or name recognition) may be more likely to engage with individuals using social media such as Twitter to connect with constituents (Kim & Park, 2012). Similarly, in the 2010 US midterm election, the conservative minority used social media and specifically Twitter more effectively to build support (Livne, Simmons, Adar, & Adamic, 2011; Netherland & McCroskey, 2010; Schaper, 2010).

In a comparative analysis of incumbents and challengers, it was found that challengers tended to interact more with the public using social media (Pole & Xenos, 2011). Although one sided, this activity has been illustrated to occur both before and during electoral periods illustrating that once the technology is adopted the technological connection can persist (Macnamara, 2011). This illustrates how powerful of a tool social networking sites can be as they lower the costs of engagement and allow for challengers to get more exposure to constituents. This engagement may lead to electoral success, but there is a lack of research on this topic.
The number of times a candidate tweets has been correlated to the number of followers, the competitiveness of the race, the amount of money they spend and the size of the home state (Ammann, 2010). Although more tweets by a candidate may not be the reason that they get more votes, some research in other technologies such as Facebook has found that popularity on social networking sites such as Facebook may be a proxy for votes (Williams & Gulati, 2009; Williams & Gulati, 2007). Another reason is that the more money that candidates raise, the more staff they are able to hire to manage their social media presence.

**Categorizing Twitters by political language and affiliation**

Through analysis of retweet behavior, researchers at the Truthy project at Indiana have been able to identify partisan clusters illustrative of echo chambers of ideas in information diffusion in Twitter (Conover et al., 2011b). In addition to trying to understand general political activity, these researchers have been able to identify the proliferation of false memes in the context of Twitter (Ratkiewicz et al., 2011). The researchers were able to identify false memes that they deemed “truthy” memes by analyzing the network graph of parties attempting to spread information at abnormal rates and through analysis of the content of the tweets. Graphs with a limited number of connected components, star-like properties with high average degree and weighted edges between dyads would indicate possible “astro-turf” campaigns. Additionally, many “truthy” memes would attempt to game the inclusion of URLs in the tweets by adding random strings at the end of the URL so the shortening service would think that the URL was unique compared to others. This would lead users to believe that these URLs were unique.

These researchers also found the presence of “content injection,” which signifies users adopting partisan hashtags or keywords to broadcast material that may be counter to the ideology of the party to proliferate a message. Content injection techniques have also been identified by Livne et al. (2011) who in an analysis of the run up to the 2010 midterm election identified high usage of the conservative hashtag #tcot (Top Conservatives on Twitter) by Democratic candidates (Livne et al., 2011). The researchers conclude that this usage was an effort to expose #tcot followers to Democratic ideas. This finding may have implications for large-scale automated analysis of these types of datasets. As identified in this study, #tcot was one of the most frequently occurring hashtags, even though it was not explicitly collected.
Using a dataset of individuals running for the US House of Representatives, Senate and governorships in the 2010 midterm elections, along with a random sample of their followers, researchers used multi-dimensional scaling to classify users based on hashtag and mention usage (Hanna, Sayre, Bode, Yang, & Shah, 2011). They found that the frequency of mentions and hashtags that were classified as being associated with one partisan affiliation can be useful in identifying partisan leanings of the individuals utilizing the technology specific syntactical features. Although the researchers were able to identify a majority of the individuals, there were some ambiguous clusters that they claim may be the “poaching” hashtags. This “poaching” is similar to the “content injection” identified by other researchers (Conover et al., 2011b).

Research has been conducted in the context of elections to identify different characteristics of users that participate in discourse based on the number of contributions they made (Mustafaraj, Finn, Whitlock, & Metaxas, 2011). Users were broken into five categories based on number of contributions to discourse during the special election in Massachusetts to replace the late Ted Kennedy. It was found that the more “vocal” Twitter users (those posting more than 50 times (n=574) in a dataset of about 235,000 tweets) tended to use more hashtags, links and retweets than the other groups identified with the greatest difference occurring between the vocal Twitterers and those that only contributed once. In addition to those characteristics, it was discovered that users who were more vocal, were newer to Twitter.

German researchers used politically oriented hashtags to examine 2009 election discourse (Jurgens, Jungherr, & Schoen, 2011). During this election, German Twitter users were encouraged to use previously determined party related hashtags followed by + or – to illustrate agreement or disagreement with the message and the party. This additional metadata allowed the researchers to extend their analysis one step further by determining whether the hashtag was used in conjunction with a positive or negative reaction to the party. Such an analysis helps to address the possibility of “content injection” although individuals who would use content injection techniques would likely also misrepresent the valence. From a network perspective, Jurgens et al. (2011) were able to identify “small worlds” of connected individuals that had similar political viewpoints. This research illustrates the possibilities associated with analyzing hashtags that are promoted by the media or campaigns.
**Analyzing political events**

Analysis of political discourse on Twitter has focused on issues, candidate debates and elections. Each context differs in both the type of data and the implications of the findings. A study of the 2010 Australian Federal Election that examined the hashtag #Ausvotes collected 415,009 tweets from 36,287 users over a 6 week time period (Bruns & Burgess, 2011). Bruns and Burgess (2011) found that the discourse in this corpus was candidate centric. A majority of the mentions collected with the hashtag were of politicians running in the race and also prominent journalists, as they were the ones integral in covering the campaign. The researchers also identified through collection of candidates running in the race that certain politicians did not use the #ausvotes hashtag at all even though they were involved in the campaign.

The researchers also constructed networks of at-replies and at-mentions to identify the most central actors based on degree centrality. Bruns and Burgess extended this analysis by also applying the measure betweenness centrality to identify individuals who may not have been the most prolific (high in-degree centrality), but who based on the network of replies and retweets played an integral role in information dissemination and bridging in the network based on their position. The most highly central individuals in this study were politicians and media personalities. The researchers did not use follower and followee networks in constructing their networks since these networks are a result of “longer-term affinities and affiliations between users.” The point of using a hashtag is to encounter individuals that are not already in a network to participate in a distributed conversation with and as a result this means that more discourse oriented networks that are constructed around hashtags and are relevant to studying political activity.

Analysis of technologically mediated discourse surrounding political debates and speeches has been an area of burgeoning study and methodological approaches are still being developed and in some areas are titled “live research” methods (Elmer et al., in press). Research on one of GOP Primary Debates from the 2012 United States General Election has identified that the promotion of hashtags to allow individuals at home to announce whether they thought the candidate answered the question or dodged it created different communities of discourse. Only 13% of the individuals that participated in the #answer versus #dodge exercise promoted by the host of the debate, FOX News, participated in the general discourse related to the debate hashtag #scdebate. Additionally, the device usage between the communities
was different illustrating how different user communities may use devices or hashtags differently (Black et al., 2012).

Further research has identified differences in the types of discourse that occur surrounding these types of events. Hu et al. (2012) found that discourse that preceded or followed debates tended to be more “steady” in that it spoke more generally about the event (Hu, John, Seligmann, & Wang, 2012a). During the debate or the press conference this discourse shifted to more “episodic” discourse that addressed actual events as they occurred even though the timeline did not always correlate 100% to the events. The researchers use this lack of 100% correlation as a caveat to other approaches employed in studying debate discourse (Shamma et al., 2009).

Using hashtags and user accounts of those participating in a 2008 Canadian federal election among the federal leaders, Canadian researchers were able to identify spikes in temporal activity as a result of statements of wide appeal or controversy. By capturing the activity before and after the debate, the researchers were able to identify the emergence of different hashtags based on partisan affiliation (Elmer et al., in press). The dataset for this analysis was small, but illustrates a methodological approach that had previously not been attempted.

Shamma et al. (Shamma, Kennedy, & Churchill, 2010b) analyzed 53,712 tweets from the Twitter public timeline during President Obama’s inauguration to identify a set of 13,370 inauguration related tweets. They found that certain activity such as at-mentions dropped when important parts of the inauguration such as President Obama taking an oath and Vice President Biden taking the oath occurred, but increased over time as important events were not occurring. A decrease in the average word count of tweets during this time was also identified. Using this analysis, the researchers were able to segment the broadcast events and further understand the conversation that was occurring (Shamma, Kennedy, & Churchill, 2010b). The researchers conclude that as individuals pay more attention to the onscreen activity they are less likely to be tweeting extensively and using syntactical features such as at-mentions to highlight or engage with others. The following study provides further evidence to support these previous findings.
**Predicting elections?**

Researchers have found some ability to “predict” elections based on collecting tweets related to elections and that Twitter activity mirrors polling activity. Researchers in Germany collected 104,003 tweets that preceded the 2009 Federal election of the national parliament in 2009 and were able to find that the number of mentions of political parties and candidates were correlated to the vote count (Tumasjan et al., 2010). In addition to these correlations, researchers were able to identify that the frequency of conversational tweets mirrored earlier research of large datasets (Honeycutt & Herring, 2009). Tumasjan et al. (2010) also found that close to 20% of their sample were retweets, which also mirrors previous research.

Extending the analysis beyond just single mentions, the researchers found that the mention co-occurrence of politicians in a single tweet tended to be individuals from the same party or similar on the partisanship spectrum. This type of finding could be transferrable to other parliamentary systems looking to identify possible viable coalition governments as the co-occurrence of politicians may indicate public perception of stances on issues. Comparison of the differences in sentiment analysis of the tweets was also able to identify parties that were more closely aligned with each other. Though this research was exploratory, these findings have many possible implications for understanding government and political use of Twitter.

In an analysis of elections in Singapore, Skoric et al. (2012), found a correlation between the number of mentions on Twitter for a party and the number of votes it received in an election at the national level which was similar to earlier research (Skoric et al., 2012). This correlation was weaker at the constituency level, but illustrates that Twitter activity at the macro level may be indicative of possible vote share in some capacity. The correlation between vote share and mention rate or popularity on other social networking sites may also be a proxy of other activity such as spending or media activity.

The promotion of the idea that SNS’ can be used to “predict” elections has been met with some criticism. A review of the “predictive” work of other researchers mentioned above conducted by Metaxas et al. (Metaxas, Mustafaraj, & Gayo-Avello, 2011) identified flaws in the approaches of previous researchers. One of the fundamental flaws is the lack of comparison by these researchers to traditional mechanisms of prediction and analysis such as polling or historical research that illustrates that the incumbent wins close to 90% of United States Congressional elections. Metaxas et al. (2011) 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. (2011) were unable to reproduce the results. In fact, the researchers were only able to
produce results that were only slightly better than chance in predicting elections.

Metaxas et al. (2011) identify demographic skewing as one of the possible problems with using
Twitter to predict elections. These findings lead to a call on behalf of the researchers to perform more in-
depth analysis of the actual activity since the tweets and other social media data are susceptible to
manipulation and also tie the analysis to more established theories or in the context of these theories to
understand if predicting elections based on social media approach is better or worse. Other researchers have
followed up in other socio-technical environments such as Facebook and in light of traditional polling have
been unable to do any better, yet they claim that as more individuals participate in these environments for
political activity, that the methods will get better (Giglietto, 2012). This is an area of future research that
would have significant implications if such a correlation could be identified before an actual election took
place.

The Study

The election occurred over a year ago and there has been limited published research on how social
media was used in the context of the election (Conway, Kenski, & Wang, 2013; Zhang, Seltzer, & Bichard,
2013). None of the research that has been published has examined such a large corpus in a longitudinal
fashion using a comprehensive dataset that is theoretically informed. Instead, the research has examined
one event or set of events using the candidates as a collection seed (Lin, Keegan, Margolin, & Lazer, 2013).
The following study addresses a significant gap in the understanding of how Twitter was used in the 2012
Presidential Election.
Data collection of electronic trace data should be theoretically informed and take into account the topical and technological context (Goggins, Mascaro, & Valetto, 2012a; Goggins, Galyen, & Laffey, 2010; Goggins, Laffey, Amelung, & Gallagher, 2010; Goggins, Mascaro, Mascaro, & Gallagher, 2012b; Goggins, Valetto, Blincoe, & Mascaro, 2012b). As a result of the unique nature of each study of technologically mediated activity, it is imperative to describe data collection processes and the provenance of the data. This description helps to guide the reader through the complete research process from dataset construction through the findings and discussion. In studies that use traditional methodological approaches such detail may be superfluous, but with the evolving APIs, tools and methodological approaches used in social media research, transparency is required to help others understand not just the methods used to examine the data, but also how the dataset was constructed as each decision is material to the overall findings (Karpf, 2012).

The following chapter begins with a description of the data collection approach followed by a discussion of the queries included in the construction of the dataset. Following this discussion, a brief overview of the dataset is provided to situate the composition of the dataset in the larger context of Twitter research. Finally, the methods used to examine the numerous datasets are provided in the context of each research question.

**Data Collection**

The data collection for this study was done using a modified version of the TwitterZombie (Black et al., 2012) infrastructure. The software that conducted the management of the queries and collection was the same as the original TwitterZombie infrastructure that was previously validated in Black et al. (2012). Given the nature of the collection requirements for what was anticipated to be a high volume event (in this case the 2012 Presidential Election), the technical infrastructure of the original TwitterZombie system was redesigned.

**Original TwitterZombie design**

The original TwitterZombie infrastructure utilized one collection head that stored data in one database. This technical architecture was deemed to be insufficient for high volume collection for two reasons. First, Twitter limits the amount of tweets that can be collected at one time for each query using the
SEARCH API to the 1,500 most recent tweets. This constraint limits the overall number of tweets that can be collected by each job during each collection run. In turn, this requires jobs with high volume to be run frequently in order to collect the data. Second, each query can take anywhere from 2-45 seconds to run in the TwitterZombie system based on the volume of data that is being collected and each query must finish before the next one begins.\(^1\)

One of the design decisions for the software architecture used in the original TwitterZombie was to run queries sequentially. Once one query finished, the next would begin. This design decision allowed for TwitterZombie to recover from any system or network interruptions or technical issues that may be present in the Twitter network. If one query failed, the next would execute successfully because there was no dependency. Although this design decision increased the resiliency of TwitterZombie, it also limited some of its applicability to high volume events. In combination, these two limitations were deemed to be prohibitive for collecting a comprehensive dataset that used multiple queries during a high volume event.

During the initial testing and development of the TwitterZombie system identified in Black et al. (2012) it was identified that high volume events would overburden TwitterZombie as it was possible that a job would take too long to complete. This would delay the collection of another job that was next in sequence. For example, if three jobs were supposed to collect every minute on one head and the first two jobs ran over one minute, the third would start collecting and delay the first two jobs from running on schedule. This would create a cascade effect that would lead to a backup that may never be overcome and this might lead to significant gaps in collection. Therefore, a new technical infrastructure was developed to account for high volume events.

**Redesigned Technical infrastructure of TwitterZombie-n**

As a result of this need for high volume collection, the TwitterZombie technical infrastructure was rebuilt using the Amazon Web Services (AWS) cloud infrastructure. The cloud infrastructure provided collection elasticity that allowed for scalability based on collection requirements such as the occurrence of an acute event such as a debate. The new system, “TwitterZombie-n”, was an adaptation of the original Twitterzombie infrastructure for the cloud environment.\(^2\) In the new infrastructure, multiple collection

---

\(^1\) This limitation is not specific to the TwitterZombie system and merely reflects the delay that would exist for collection between any collection system and the Twitter API.

\(^2\) The “n” in Twitterzombie-N stands for the possibility for the addition of multiple collection heads.
heads were used to collect data. This allowed for the maximum possible collection based on the available resources given financial and technical restraints and for a higher volume of data to be collected. Although there are multiple collection heads in the new design, the infrastructure was built using one central “job” table\textsuperscript{13}. The centralization of the job table limited the possibility of system failure at one of the collection points and provided one systematic point of control.

The distribution of multiple queries across multiple collection heads limited the possibility of collection leading to overburdening heads as jobs were evenly distributed among heads. In the early part of the election cycle, four heads were utilized, but this was expanded to five during October and November to account for higher traffic volume from the debates and election day as identified in finding 1a. The system was still limited by the limit of 1,500 tweets per minute for each job, but the distribution of queries amongst the collection heads allowed the system to run as efficiently as possible given its tasking.

The TwitterZombie-N collection system was run at an Amazon Web Services colocation facility in Northern Virginia for the duration of the collection cycle (August 20, 2012 – November 13, 2012). In an effort to eliminate any possibility of collection failure during high volume events, another duplicate system was run during the last 2 weeks of the election cycle (October 28 – November 13) in another AWS facility in Oregon. The geographic separation assured availability of resources to limit any collection faults that were the result of a loss of network activity or other unanticipated technical failure\textsuperscript{14}. This redundancy ensured that a dataset that covered the whole time period without any gaps would be collected

As a result of the high volume collection of many of the queries during the last two weeks of the election, it was decided to combine the two datasets (Northern Virginia and Oregon) into one dataset. This combination was done using the “tweet_id” string that is included as part of each tweet to ensure no duplicate tweets existed in the combined dataset. As a result of the addition of the Oregon dataset to the Northern Virginia dataset an additional 2 million tweets were collected\textsuperscript{15}.

\textsuperscript{13} In Twitterzombie, a job and a query are synonymous and used interchangeably.

\textsuperscript{14} The AWS location in Northern Virginia where the original system was being run had multiple documented occasions in the previous year where service was lost for a multiple hour time period. This led to the loss of availability of popular websites such as Netflix. The addition of the AWS site in Oregon was intended to guard against this possibility and in turn led to greater collection.

\textsuperscript{15} An analysis of the tweet_id strings was conducted to identify any duplicates after the datasets were combined and there were no duplicated identified.
These additional tweets are the result of high frequency Twitter activity during the conventions, debates and on Election Day that the two systems (Northern Virginia and Oregon) were able to collect based on a slight offset of clocks on the two systems. Black et al. (2012) identified the possibility for differences in collection based on the time and this phenomenon was identified in the TwitterZombie collection infrastructure. Although the job tables were offset by a matter of seconds, they collected a number of different tweets during high volume events. This redundancy and elasticity helped to ensure a more representative collection sample, even though there are likely data gaps as a result of limits placed on public API access by Twitter.

Dataset Creation

The TwitterZombie architecture was designed to limit the need to disaggregate collected data. In the system, each tweet that meets the selection criteria for a query is collected as a distinct entry in the database. Therefore, the tweet: “@BarackObama Good work in the debate Mr. President beat @MittRomney #debates #election2012” would be collected five times by the TwitterZombie jobs that were currently collecting data for the election (@BarackObama, @MittRomney, #debates, #election2012 and “barack AND obama”).

This design allows for easy access to tweets that meet certain criteria, as it is possible to query the database by “job number” as opposed to querying the database for a string of characters. On one hand, the discrete job selection is powerful as one is able to get data quickly when examining tweets for a specific query. On the other hand, the creation of a dataset that represents the discourse about a larger topic that traverses multiple jobs during a specified time period becomes costly in terms of processing time. In turn, this approach requires a clear set of research questions and description of why specific jobs were included in the larger dataset. The following section describes the construction of the dataset for this study in terms of three categories of selection criteria: event related queries, candidate related queries and queries related to campaign promoted activity in Twitter.

The dataset used for this study was the result of combining 68 theoretically informed queries that are described in the next section. The complete dataset after the combination of Northern Virginia and

---

16 Although there are likely significant gaps in the dataset, this dataset is the most representative dataset that is publicly known for an American election. Twitter (the company) is likely to have the most comprehensive dataset since they have access to all of the data, but this type of access was not possible.
Oregon datasets and the 68 queries amounted to 62,806,682 tweets. After using the SQL `distinct` command on the unique tweet identifier (tweet_id) of each message to eliminate duplicates that would occur as a result of multiple syntactical features being present in a tweet, the final dataset amounted to 52,487,179 tweets\(^{17}\). This is a reduction of 10,319,503 tweets, approximately 16% of the original dataset.

The 16% overlap highlights the fact that search queries were highly correlated as one tweet was collected by multiple queries. This is likely a result of the theoretically informed query selection (Goggins et al., 2012a) that was done to ensure the widest amount of political discourse related to the 2012 Presidential Election. It is intuitive that tweets about President Obama during the election would also likely mention a debate or his opponent Mitt Romney and would be collected by multiple jobs resulting in overlap. The next section identifies the selection of jobs and inclusion of many queries that were used to collect political discourse during the time period of interest to afford a level of transparency in the construction of the dataset (Karpf, 2012).

**Job selection**

In a big social data environment a representative sample of the data is desired. Unfortunately, this type of sample is difficult to collect as it is unclear as to whether the technology provider is limiting access to data or whether individuals are using appropriate terms for comprehensive collection. This is a problem that is not only a factor with this study, but also with all studies that rely on API-based collection from technology services that are not under the control or management of the researcher or without other data sharing relationships established between the researcher and the company.

To date, there has been limited analysis of the specifics of the different Twitter APIs and this is definitely an area for future research, but outside the scope of this work (Morstatter, Pfeffer, Liu, & Carley, 2013). In an effort to collect as much political discourse related to the election during the time period, a total of 68 queries were collected on in the Twitterzombie-n infrastructure. These queries were a combination of Twitter handles, hashtags, and keywords related to the candidates and temporal events of the election such as the conventions, debates and Election Day.

---

\(^{17}\) The identification of duplicate tweets was validated by using R to identify unique tweets based on created_at, tweet text and from_user. The numbers in each dataset were the same, which indicates the validity of these approaches for dataset creation.
The initial set of queries was established by identifying all of the handles associated with the candidates and the campaigns, including the wives of the candidates since they were also participating in campaign events. In addition to these handles, the first and last name of the candidates were added to the collection in an effort to capture anytime an individual references one of the candidates without using their Twitter handle. Additionally, specific hashtags that were used in the candidates Twitter feeds were collected as these represented areas of discourse.

Throughout the data collection period, new hashtags were identified as associated with electoral events and promoted by candidates. These were not included in the initial set of queries, but were added to the collection infrastructure to attempt to collect a comprehensive dataset. All of the queries that were added during the election were kept on collection through the end of the collection period (November 13, 2012). The sustained collection of queries once they were identified and put into the system allows for temporal analysis of the data.

Since many of the queries emerged during the election they were not present at the initial commencement of the study. As a result of the fact that temporal analysis of the syntactical features is provided in the findings chapter, the date of the first tweet collected for each job is provided in Appendix A along with the total number of tweets collected for each job. Many of the jobs that were added during the election came from emergent events and as a result of this, the date of the first tweet collected also coincides with the date that the job was initially entered into the TwitterZombie system. This date allows the reader to identify when collection of specific queries began and puts temporal analysis into context for each of the queries.\[18\]

**Candidate dataset creation**

The queries related to the major Presidential and Vice Presidential candidates formed the foundation of the dataset. Each of the candidate’s official Twitter handles was collected as part of the collection process. In the case of Paul Ryan, his Congressional handles were also collected since he was concurrently running a Congressional campaign and it was likely that the discourse would bleed over

\[18\] A “First Seen” date of 8/20/2012 represents a query that was part of the initial set of queries loaded into TwitterZombie-n
between the two campaigns. In addition to each handle, the hashtag for each of the candidate’s full names, first names and last names was tasked. Hashtags for Paul Ryan were not collected due to the common nature of his name and the identification of completely unrelated discourse related to his name in an early sample of the data.

In addition to the handles for the candidates, a keyword combination of the candidate’s full names was tasked to ensure collection of data for when the handle or hashtag was not used with the candidate’s name. The keyword query followed the format of [candidate first name]%20[candidate last name]. In addition to the full names, the query “president %20 obama” was used to collect tweets that referred to President Obama by formal title instead of his name.

Twitter handles for the candidate’s wives along with the official White House account @WhiteHouse were also collected. Finally, the hashtags for the Presidential and Vice-Presidential candidates and “2012” (e.g. #obamabiden2012) were also tasked to collect data from those who may have used the hashtag as an affiliation. Table 3 illustrates the queries for the candidates and hashtags that represented campaign specific discourse.

Table 3: Candidate Queries

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Campaign</th>
<th>Hashtag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td></td>
<td>@BarackObama</td>
</tr>
<tr>
<td>Joe Biden</td>
<td></td>
<td>@Joe_Biden</td>
</tr>
<tr>
<td>Mitt Romney</td>
<td></td>
<td>@MittRomney</td>
</tr>
<tr>
<td>Paul Ryan</td>
<td></td>
<td>@PaulRyanVP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#Obamabiden2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@Obama2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@JoeBiden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#MittRomney</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@RepPaulRyan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#romneyryan2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#BarackObama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@VP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#Romney</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@PaulRyanPress</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#obama2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#Obama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#JoeBiden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mitt and romney</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@PaulRyan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#romney2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>barack and obama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>#Biden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>paul AND ryan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>President AND obama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>joe and biden</td>
</tr>
</tbody>
</table>

Event dataset creation

In addition to the candidates there were a number of emergent events where hashtags evolved and were used to mark specific discourse. The #election2012 hashtag was used throughout the time period and was also used before the studied time period. In total, #election2012 was used in 1,632,995 tweets in the

---

19 Even though Paul Ryan was actively running for reelection to Congress, his Congressional Twitter account garnered limited mentions illustrating that the primary focus of activity for him during the election cycle was his Vice Presidential handle (@PaulRyanVP).

20 “%20” represents the ASCII representation of the Boolean term AND.
dataset. The #electionday hashtag emerged on November 6 and was used in 221,610 tweets along with #vote (562,937 tweets), #govote (95,841 tweets) and the keyword syntax of “voted for” (1,476,786 tweets), where individuals were identifying who they voted for publicly on Twitter.

Events such as the conventions and the debates often did not have a predetermined hashtag or set of hashtags until right before the event and were difficult to collect until the day before the event. In the case of the debates, Twitter identified the official hashtag of #debates the day before the first debate even though some press releases for the first debate indicated that the hashtag would be #dudebate or #denverdebate as opposed to #debates. The use of one specific hashtag for all of the debates allowed individuals to be aware of the hashtag to identify in and participate in discourse for an extended amount of time. The use of the same hashtag for multiple discrete events (each debate) made the possibility of differentiation of discourse between debates difficult. Figure 1 illustrates that the usage of the #debates hashtag was concentrated the days immediately surrounding the debate making time a valid unit of analysis for examining debate specific discourse.

In the case of the first debate, the press release for the debate from the University of Denver identified the hashtag as being #dudebate two weeks before the debate, but this hashtag was only used 850 times. Additionally there were other hashtags identified by members of the local press such as #debatedenver and #denverdebate, but these were used in a limited capacity compared to the official
#debates hashtag promoted by Twitter. Although these debate specific hashtags were not extensively used, they are included in the dataset for completeness. This behavior differs from previous analysis of primary debates where there were specific hashtags used for each debate in addition to some cable network specific hashtags such as #cnndebate or #answer and #dodge (Black et al., 2012). Similarly, other hashtags promoted by the local event organizers emerged during the other two Presidential debates and the Vice-Presidential debate and are included in the dataset.

Table 4 includes all of the hashtags and the respective debate that they were collected in relation to. Even though the hashtags were related to the specific event, the collection of them occurred through the end of the election. This allows for the temporal analysis of the presence of the hashtag. In most cases the specific debate hashtags were added to the collection system the day of or before the specific debate. It is important to note that in Figure 1, there is limited presence of #debates before the day of the debate, 10/3/2012. Although there was limited activity, collection began the week before as identified in Appendix A. This illustrates a limited adoption until the actual day of the debate.

The hashtags that were common for all of the debates were #debate, #debates, #presidentialdebate, #cnndebate, #debate2012, #presdebate and as a result of their commonality across the four debates they are not identified as being associated with a specific debate, in Table 4. The hashtags and Twitter handles that were promoted by the respective parties for the conventions are included in Table 4.

Table 4: Key Election Season Events and Related Queries

<table>
<thead>
<tr>
<th>Dates</th>
<th>Event</th>
<th>Syntactical Feature Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 27 - August 30, 2012</td>
<td>Republican Convention</td>
<td>#gop2012, #2012rnc, @gopconvention</td>
</tr>
<tr>
<td>September 4 - September 6, 2012</td>
<td>Democratic Convention</td>
<td>#dnc2012, #2012dnc, @demconvention</td>
</tr>
<tr>
<td>October 3, 2012</td>
<td>Presidential Debate #1</td>
<td>#dudebate, #denverdebate, #debatedenver</td>
</tr>
<tr>
<td>October 11, 2012</td>
<td>Vice-Presidential Debate</td>
<td>#vpdebate, #centrevpdebate</td>
</tr>
<tr>
<td>October 16, 2012</td>
<td>Presidential Debate #2</td>
<td>#hofdebate</td>
</tr>
<tr>
<td>October 22, 2012</td>
<td>Presidential Debate #3</td>
<td>#lynndebate</td>
</tr>
<tr>
<td>November 6, 2012</td>
<td>Election Day</td>
<td>#electionday, #govote, #ivoted, “voted AND for”</td>
</tr>
</tbody>
</table>
Campaign promoted hashtag dataset creation

There were a number of promoted hashtags by the campaigns. A promoted hashtag is one where an organization pays Twitter to promote a hashtag as trending on its front page. This can be done by any organization, but was used primarily by the candidates to attempt to shape the discourse on Twitter. During each morning of data collection, Twitter was checked to see if there was a promoted hashtag from a campaign. This occurred at 6 am to ensure collection began before most of the east coast had woken up. If a promoted hashtag was identified, it was added to the job table for collection.

As a result of emergent events there were a number of other hashtags that were not promoted, but were trending as a result of the large amount of discourse and events in the news that were identified in the same manner. For example, a video of Mitt Romney stating that 47 percent of Americans were reliant on government support spawned the hashtag #47percent. Table 5 identifies the promoted and emergent hashtags collected as a result with promoted hashtags have a (p) next to them\textsuperscript{21}.

Table 5: Promoted Hashtags

<table>
<thead>
<tr>
<th>Promoted Hashtags</th>
</tr>
</thead>
<tbody>
<tr>
<td>#16trillionfail</td>
</tr>
<tr>
<td>#47percent</td>
</tr>
<tr>
<td>#romneyshambles (p)</td>
</tr>
<tr>
<td>#forwardnotback (p)</td>
</tr>
<tr>
<td>#forward (p)</td>
</tr>
<tr>
<td>#forward2012 (p)</td>
</tr>
<tr>
<td>#cantafford4more (p)</td>
</tr>
</tbody>
</table>

Collection summary

In total, the collection of candidate and event specific queries coupled with promoted hashtags from the two campaigns represent a comprehensive dataset that captures a sizeable amount of political

\textsuperscript{21} The trending hashtags are included for completeness of the dataset although they do not represent a promoted hashtag.
discourse from the 2012 Presidential Election. Figure 2 presents an overview of the process described in the previous section compiling the complete dataset for analysis.

![Figure 2: Collection and Dataset Construction Overview](image)

Appendix A provides a table of all of the queries that were used to construct the dataset along with the total number of tweets collected and as described earlier, the date that a tweet with query was first seen in the dataset. The higher amount of tweets collected that were related to the Obama queries is in line with other research that finds that social media discourse about the incumbent is greater during elections (Williams & Gulati, 2007; Williams & Gulati, 2008; Williams & Gulati, 2009).

The combination of the time period that collection occurred coupled with the terms used helps in the construction of the corpus of political discourse. Political discourse as conceptualized by Van Dijk (van Dijk, 1997) and others (Fairclough & Fairclough, 2012; Fairclough, 1992) represents discourse about politicians, among politicians and about issues that are of concern to politicians. These criteria are met by the selection criteria used to build this dataset.

It is impossible to collect all of the possible political discourse during a time period, as individuals may not explicitly identify that the tweet is about the politics. There were also some technical limitations to collecting all tweets during that time period even with the identified queries as discussed earlier. Since Twitter has a variety of mechanisms to identify discourse about topics that include hashtags and the ability
to mention individuals this dataset is syntactically rich. The addition of the keyword searches helps to capture any discourse by individuals who use only text as opposed to the Twitter prescribed features.

**Dataset provenance and partitioning**

The complete dataset forms the foundation for the analysis, but there are a number of datasets that were subsets of the complete dataset that were used to address each of the research questions. To help orient the reader, Figure 3 provides an overview of the provenance of the data along with the research questions that each of the datasets address and how these datasets relate to each other.

![Figure 3: Dataset Overview](image)

Table 6 provides a detailed description of each of the datasets along with the number of tweets in each. A similar table with the digital location of the archived datasets is included in Appendix B. Each of the datasets was created as a subset of the complete dataset based on the presence of one of the syntactical

---

22 The datasets used for this study are currently archived by the author for use in future research publications. Requests will be assessed based on the purpose and can be made by emailing cmascaro@gmail.com.

23 The datasets and statistics provided in the next section are derived from the scripts and methods in the “Analytical Methods” section, but are provided here to orient the reader to the dataset.
features and then further sliced to address more specific parts of the related research question identified in the illustration. The criteria and technical methods to do this slicing are discussed in the analytical methods section. For a description of the fields that are included in the Complete Dataset and the subsets of data, please see Appendix C.

### Table 6: Dataset Descriptions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rows</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Dataset</td>
<td>52,487,179</td>
<td>This dataset was the complete dataset of 52,487,179 tweets that represent the elimination of duplicates from the combination of the Northern Virginia and Oregon datasets.</td>
</tr>
<tr>
<td>Hashtag Dataset</td>
<td>26,322,494</td>
<td>This dataset consists of all of the tweets that contained a hashtag. This was used as the foundation for further slicing based on the presence of #election2012, #debates and the promoted hashtags.</td>
</tr>
<tr>
<td>#Election2012 Subset</td>
<td>1,632,995</td>
<td>This dataset consists of all of the tweets that contained the #election2012 hashtag.</td>
</tr>
<tr>
<td>#Debates Subset</td>
<td>1,089,254</td>
<td>This dataset consists of all of the tweets that contained the #debates hashtag.</td>
</tr>
<tr>
<td>Promoted Hashtag Subset</td>
<td>1,421,776</td>
<td>This dataset consists of all of the tweets that contained the #cantafoild4more, #forward, and #forward2012 hashtags.</td>
</tr>
<tr>
<td>Debate Days Subset</td>
<td>10,319,375</td>
<td>This dataset consists of all of the tweets on the day of and day after the four debates during the 2012 election.</td>
</tr>
<tr>
<td>At-mention Dataset</td>
<td>38,213,732</td>
<td>This dataset consists of all of the tweets that contained an at-mention.</td>
</tr>
<tr>
<td>At-reply Dataset</td>
<td>4,606,908</td>
<td>This dataset is a subset of the at-mention dataset where the at-mention is at the first position of the tweet text. There are three subsets of this data that were created. These datasets included individuals that were only a source of an at-reply, those that were only a target and those that were on both sides of the conversation.</td>
</tr>
<tr>
<td>URL Dataset</td>
<td>17,105,877</td>
<td>This dataset consists of all tweets that contained a URL.</td>
</tr>
<tr>
<td>Retweeted URLs</td>
<td>9,000,739</td>
<td>This dataset is a subset of the URL dataset and contains all of tweets with a URL that were a retweet.</td>
</tr>
<tr>
<td>At-reply URLs</td>
<td>523,216</td>
<td>This dataset is a subset of the URL dataset and contains all of tweets with a URL that were an at-reply.</td>
</tr>
<tr>
<td>Retweet Dataset</td>
<td>28,922,377</td>
<td>This dataset consists of all tweets that were a retweet and was further slices by the syntax (&quot;RT @[username]&quot; or &quot;via @[username]&quot;).</td>
</tr>
</tbody>
</table>

24 A row represents one tweet except for the datasets “Retweeted URLs” and “At-Reply URLs”. In these datasets, a row represents a unique shortened URL as opposed to a tweet.
Dataset Description

In total, there were 52,487,179 unique tweets that were captured using the combination of 68 queries (47 hashtags, 15 handles, 6 keyword searches). Of these unique tweets that were collected, there were 28,019,513 that contained unique text. This disparity exists because a tweet that is retweeted 100 times exists 100 times in the dataset, as each retweet is its own unique tweet in the eyes of Twitter, but from an analysis perspective the actual text is not unique. Therefore, the comparison of unique tweet count to overall tweet count highlights that only 53.3% of the total tweets contained original text.

The most frequently occurring tweet in the dataset was a tweet from Barack Obama following his reelection that was a photo of himself and the First Lady hugging with the text “Four More Years.” According to Twitter, this tweet was retweeted over 810,000 times in 48 hours and it appears in the dataset 394,494 times. This is three times the next highest retweeted tweet, which was retweeted 110,997 times. Until early 2014, Obama’s tweet was the most popular tweet and most retweeted tweet of all time. The fact that this tweet is the most frequently occurring in the dataset and was collected approximately 50% of the time that Twitter says it was retweeted provides some external validation to the collection infrastructure as being able to collect a significant amount of the activity even during an acute event.

It is likely that the times that this tweet was retweeted and not collected by the system occurred when the system had already collected the 1,500 maximum tweets per minute. According to Twitter, at 11:19pm ET there were 327,452 tweets per minute happening. This rate is a record for an event in Twitter (Sharp, 2012b). This high number of tweets per minute likely resulted in missed collection for certain queries. Since the dataset is large and most research relies on some form of a sample, the fact that the system collects on a consistent schedule indicates that the sample is representative and mirrors the activity in Twitter.

Syntactical feature overview

The analysis of a dataset of this size is becoming common in social media and big data research. In order to ground the research the conceptualizations of the syntactical features studied are in Table 7.

---

25 As of March 2014, this superlative is held by a photo tweeted by Ellen Degeneres at the Oscars.

26 This table is presented in the literature review, but is also included in this section to further aid the reader.
These definitions are used to slice the dataset and are further explained in the following analytical methods section.

Table 7: Syntactical Feature Conceptualization

<table>
<thead>
<tr>
<th>Syntactical Feature</th>
<th>Common Syntax</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>At-Reply</td>
<td>@[username] at first position of tweet text</td>
<td>To directly address another individual in a public manner.</td>
</tr>
<tr>
<td>At-Mention</td>
<td>@[username] at any point in tweet text</td>
<td>To highlight a tweet to another individual or to talk about someone. Mentioning them will inform them of the tweet.</td>
</tr>
<tr>
<td>Retweet</td>
<td>RT @[username] “tweet text”</td>
<td>To further disseminate another individuals tweet.</td>
</tr>
<tr>
<td>Links</td>
<td>http://[until whitespace]</td>
<td>To include external information in a tweet. Note: Twitter uses a URL shortener, but also accepts other URL shorteners as links too.</td>
</tr>
<tr>
<td>Hashtags</td>
<td>#[alphanumeric text]</td>
<td>To tag a message with a conversational marker or to add a tweet to an existing stream of discourse independent of a follower/followee network.</td>
</tr>
</tbody>
</table>

In previous research, descriptive statistics about a Twitter dataset may be biased as a result of the collection criteria. For example, a dataset collected using one hashtag and one handle would have a significant amount of hashtags and handles, but may lack any instances of an individual not using a hashtag or using an individual’s last name as opposed to their Twitter handle. As a result of the selection criteria, the dataset would be biased to containing a specific type of activity.

As described in the previous section, this study attempted to avoid this type of bias by using more than just hashtags and at-mentions as selection criteria. This broadens the type of data that were collected. This approach has not been identified in any other study of Twitter as other studies have traditionally relied on a set of hashtags or handles to build a dataset.

Table 8 details the raw counts of some of the syntactical features in the dataset. There were over 48 million hashtags used, but only 1 million were unique. A similar disparity exists between the total number of at-mentions and the number of unique at-mentions. Similarly, the number of unique base URLs is less than the total number of URLs collected, since many of the base URLs were used repeatedly.

---

27 A base URL is the domain of the URL without any of the specific directories. For example, the URL www.cnn.com/USA/news/Story, would have a unique base URL of www.cnn.com.
Examining the overall presence of certain syntactical features in the complete dataset illustrates a corpus with a diverse set of characteristics. Table 9 provides an overview of the percentage of tweets that contained certain syntactical features. We see that nearly one-third of all tweets contained a URL and just over one-half contained a hashtag. Nearly three-quarters of all of the tweets contained an at-mention and this includes the nearly nine percent that were constructed as an at-reply and the 55% that were a retweet. The percentage of at-replies, at-mentions and retweets are similar to previous analyses of election data on Twitter, but the percentage of tweets with URLs and hashtags is lower in this dataset (Mascaro, Black, Gallagher, & Goggins, 2012). The reasons for these differences may be the addition of more keyword queries in this dataset construction and also the increased times of acute “bursty” activity where the percentage of URLs is found to be less.

Table 9: Syntactical Feature Presence in the Complete Dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>32.59%</td>
</tr>
<tr>
<td>Hashtag</td>
<td>50.15%</td>
</tr>
<tr>
<td>At-Mention</td>
<td>72.81%</td>
</tr>
<tr>
<td>At-reply</td>
<td>8.78%</td>
</tr>
<tr>
<td>Retweet</td>
<td>55.10%</td>
</tr>
</tbody>
</table>

Table 10 identifies the percentage of tweets that contain an at-mention of the four candidates. President Obama (@BarackObama) was mentioned in almost one of out every six tweets whereas Governor Romney (@MittRomney) was mentioned in nearly one out of every twelve tweets. Both vice-presidential candidates were mentioned less than the Presidential candidates. A temporal analysis of the candidate at-mentions is presented in chapter 4.
Since a significant part of the findings depend on the temporal unit of analysis, the day, a more granular analysis of the range of percentages of the syntactical features is presented in Table 11. The variance in these ranges by day illustrates how behavior does shift on a daily basis and that there is substantial variance as a result of external events. These summary statistics illustrate that in order to understand activity one must take into account the context that the activity is happening since the observable range can be large, but explainable given the context. This is done in this study through the theoretically informed sample and analysis of the activity in the context of the events.

### Table 11: Syntactical Feature Percentage Presence Summary by Day

<table>
<thead>
<tr>
<th>Syntactical Feature</th>
<th>Overall Percentage</th>
<th>Minimum Percentage</th>
<th>Maximum Percentage</th>
<th>Mean Percentage</th>
<th>Median Percentage</th>
<th>Range Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>At-mention</td>
<td>72.81%</td>
<td>53.05%</td>
<td>80.02%</td>
<td>73.97%</td>
<td>74.41%</td>
<td>26.97%</td>
</tr>
<tr>
<td>At-reply</td>
<td>8.78%</td>
<td>6%</td>
<td>15.3%</td>
<td>10.4%</td>
<td>10.7%</td>
<td>9.3%</td>
</tr>
<tr>
<td>Retweet</td>
<td>55.10%</td>
<td>37.74%</td>
<td>60.99%</td>
<td>53.34%</td>
<td>53.9%</td>
<td>23.25%</td>
</tr>
<tr>
<td>URL</td>
<td>32.59%</td>
<td>16.74%</td>
<td>67.28%</td>
<td>40.92%</td>
<td>42.54%</td>
<td>50.54%</td>
</tr>
<tr>
<td>Hashtag</td>
<td>50.15%</td>
<td>29.68%</td>
<td>67.33%</td>
<td>43.1%</td>
<td>41.34%</td>
<td>37.65%</td>
</tr>
</tbody>
</table>

### Analytical Methods

The following study uses a variety of methodological approaches to parse and analyze the datasets described earlier in the chapter. As a result of the number of methods used for the analysis of these datasets, each method will be presented in the context of the research question along with the specific finding or set of findings that it is used to explain. This presentation helps to guide the reader and allows for a better understanding of the approaches used to address each research question and how the findings relate each method and the data examined.
The foundation of most of the analysis detailed in this study is a script written in the statistical programming language R\textsuperscript{28}. This script “TwitterZombieAnalysisScript.r” (herein: TZScript and further detailed in Appendix D) ingests a dataset from the TwitterZombie collection system and produces a series of files that allow for more granular analysis that led to the majority of the findings\textsuperscript{29}. In addition to producing a series of files that are used for analysis, the script produces an annotated version of the dataset that includes additional columns that identify the presence of syntactical features as described earlier in the chapter in Table 7.

These columns contain the values TRUE or FALSE and are used to partition the dataset into the “Hashtag,” “URL,” At-mention,” and “Retweet” datasets based on the presence of syntactical features within a tweet. Once the subsets of data are created they can also be processed using the same analytical script. Specific output files used for each finding are detailed within the respective section below. A basic representation of the process is found in Figure 4.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{TZScript_Analysis_Process.png}
\caption{TZScript Analysis Process}
\end{figure}

\textsuperscript{28} All of the analysis for this study was done using R version 2.15.1.

\textsuperscript{29} An electronic version of this script can be found at https://github.com/cmascaro/Dissertation-Scripts/blob/master/TwitterZombieAnalysisScript.r
Question 1: How does the political discourse related to the 2012 Presidential election manifest itself in a technologically mediated environment?

- How does this discourse in Twitter differ during acute events such as debates compared to long-term discourse that occurs throughout the election?

- How does the use of politically oriented hashtags in Twitter facilitate this discourse?

The first research question addresses how political discourse manifested itself in Twitter during the 2012 Presidential election, how this activity differed depending on if an acute event was happening and how hashtags facilitated this activity. Addressing this research question required the use of the complete dataset, hashtag dataset and debates days subset to identify the presence of syntactical features and examine when activity surrounding acute events occurred. Table 12 provides an overview of the datasets, methods, scripts and output files used for each of the findings related to research question 1.

Table 12: Overview of Research Question 1 Datasets, Methods, Scripts and Output Files

<table>
<thead>
<tr>
<th>Finding</th>
<th>Dataset(s)</th>
<th>Method</th>
<th>Script</th>
<th>Output File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Complete Dataset</td>
<td>Temporal Frequency Analysis of Syntactical Features</td>
<td>TZScript</td>
<td>syntacticfeatureoverview.csv</td>
</tr>
<tr>
<td>2</td>
<td>Hashtag Dataset</td>
<td>Frequency Analysis; Co-occurrence Analysis</td>
<td>TZScript</td>
<td>hashcounts.csv</td>
</tr>
<tr>
<td>3</td>
<td>#Election2012_Subset; #Debates Subset</td>
<td>Temporal Frequency Analysis</td>
<td>TZScript</td>
<td>timeseries.csv</td>
</tr>
<tr>
<td>4</td>
<td>Debate Days Subset</td>
<td>Frequency Analysis</td>
<td>TZScript</td>
<td>hashcounts.csv</td>
</tr>
<tr>
<td>5</td>
<td>Promoted Hashtag Subset</td>
<td>Time Series Analysis; Frequency Analysis</td>
<td>TZScript</td>
<td>timeseries.csv</td>
</tr>
</tbody>
</table>

30 A description of the functionality of the scripts can be found in Appendix D and Appendix E. These scripts can be found at https://github.com/cmascaro/Dissertation-Scripts

31 A description of the specific output files can be found in Appendix D.

32 The complete dataset represents all data collected and is the foundation for all of the subsets of data described in chapter 3.
Temporal Use of Syntactical Features (Finding 1)

An analysis of how users employed the numerous syntactical features was conducted to examine how political discourse manifested itself in Twitter. This research question was addressed using the complete dataset (all of the collected data) and parsing out the percentage of tweets that contained at-mentions, at-replys, URLs, retweets and hashtags using the unit of the analysis of one day (Bruns & Stieglitz, 2012b; Bruns & Stieglitz, 2013). These percentage calculations are calculated during the TZScript run and are included in the syntacticfeatureoverview.csv output file that is further discussed in Appendix D.

Given the nature of political activity, the unit of analysis of the day provides the ability to encapsulate the lead up and aftermath of an event. The time used to calculate each day is UTC. This is done to standardize times across all analysis. Many of the events spanned multiple days since the analysis of nighttime events such as debates often went into the next day on the east coast. Also, since U.S. voters can live anywhere in the world to cast ballots and American politics are often of international interest, the fact that the time used to partition days does not correlate to the east coast does not provide an issue as it is held standard throughout all analysis.

Hashtag Frequency Analysis (Finding 2)

Examining frequently occurring hashtags allows for an understanding of the prevalent discourse markers used by the public. The TZScript output file hashcounts.csv was used to examine the most frequently occurring hashtags in the complete dataset. This file presents a frequency count of all of the hashtags within the dataset.

Given the nature of the collection infrastructure there is an inherent collection bias in the dataset. The 68 terms that were explicitly collected are likely to be the most frequently occurring terms in the dataset.

---

33 The number of times that the specific syntactical feature appeared within each tweet was not material to the analysis as the research question focused on whether the syntactical feature was present and not how many times it was used (Bruns, 2012; Bruns & Stieglitz, 2012a; Bruns & Stieglitz, 2013). The nature of Twitter is that a tweet can contain multiple syntactical features so it is possible that one tweet would contribute to multiple counts in this analysis and as a result the percentages by day do not sum to 100% percent.

34 UTC was between 4-5 hours ahead of Eastern Time throughout the examined time period.

35 This time is also native to Twitter data so it required limited transformation in the analysis, which limits the possibility of introducing errors into the analysis.
complete dataset. In order to account for this type of “collection bias,” the 68 collected terms were eliminated from portions of the analysis. Following this elimination of the collected hashtags, the frequencies of the remaining hashtags were calculated to identify the most frequently occurring hashtags that co-occurred in the dataset with explicitly collected terms.

The elimination of the explicitly collected hashtags from the analysis of the most frequently occurring hashtags allows for a co-occurrence analysis of the hashtags that occurred frequently with the explicitly collected terms. The syntactical features that occur the most frequently in the extracted dataset are those that were coupled with a syntactical feature that was explicitly collected. Given the nature of collection carried out in this manner, it is imperative to conduct analyses such as these to identify terms that were frequently occurring, but not explicitly collected.

**Acute versus Non-Acute Event Analysis (Finding 3)**

The hashtag dataset was sliced into two subsets to examine the frequency and concentration of hashtags related to acute and non-acute events. The two subsets for this analysis included all of the tweets that included the #election2012 hashtag (#Election2012 Subset) and the #debates hashtag (#Debates Subset). These two datasets were processed using TZScript. The output file timeseries.csv, which identifies the frequency of tweets per day, was used to examine the differences in frequency based on whether the hashtag was associated with an acute event such as the four debates (#debates) or a non-acute event such as the election (#election2012). This methodological approach was adapted from previous research of acute events such as sporting events (Gruzd et al., 2011), television shows (Harrington, Highfield, & Bruns, 2012; Wohn & Na, 2011) and debates (Black et al., 2012; Mascaro, Black, & Goggins, 2012; Shamma et al., 2009; Shamma et al., 2010b; Shamma et al., 2010b).

**Debate Days Hashtag Usage (Finding 4)**

The methods used to identify finding 3 examine the specific use of the #debates and #election2012 hashtags as the unit of analysis, whereas finding 4 examines the temporal unit of analysis of the days when the debates occurred. Tweets from the days where the debates occurred were extracted from the complete dataset (Debate Days Subset) and processed using TZScript. The output file hashcounts.csv was then used to examine the most frequency occurring hashtags during each of the days that the debates occurred.
Because of the time used in the dataset construction (UTC), the day following the actual debate day was also included in the analysis as the debate often started at midnight or one o’clock in the morning depending on the debate. This allowed for the time immediately preceding the debate and also the actual debate to be included in the same analysis. Table 13 includes the specific days included in the construction of the “Debate Days” subset.

<table>
<thead>
<tr>
<th>Debate</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presidential Debate #1</td>
<td>10/03/2012 - 10/04/2012</td>
</tr>
<tr>
<td>Vice-Presidential Debate</td>
<td>10/11/2012 - 10/12/2012</td>
</tr>
<tr>
<td>Presidential Debate #1</td>
<td>10/16/2012 - 10/17/2012</td>
</tr>
<tr>
<td>Presidential Debate #1</td>
<td>10/22/2012 - 10/23/2012</td>
</tr>
</tbody>
</table>

**Promoted Hashtag Frequency (Finding 5)**

A subset of the promoted hashtags, #forward2012, #forward and #cantafford4more was created to identify how the frequency of these hashtags fluctuated with events (Promoted Hashtag Subset). This promoted hashtag subset was processed using TZScript and the output file timeseries.csv was used to examine the temporal frequency of these hashtags (in a similar approach to finding 3). This approach builds on previous work that focused on the analysis of discussion topics (Highfield, 2012). In the case of this analysis, the discussion in this case is conceptualized as the hashtags.

**Question 2: How does the political discourse that occurs in Twitter using the syntactical features of the at-mention and at-reply surrounding the 2012 Presidential Election identify an emerging participatory public engaged in political discourse?**

The second research question addresses how the at-mention syntactical feature was used in the context of the election. There are two types of subject-oriented discourse that are operationalized through analysis of at-mention activity. The first type of activity is indirect conversation about individuals where they are the subject of the tweet, but not being directly addressed. This behavior is identified by the inclusion of the at-mention anywhere in the tweet. In tweets with at-mentions, the author of the tweet is
highlighting the tweet to the individual and to those who may be following or searching for that individual, but may not be explicitly reaching out to that individual (as would be the case with an at-reply).

The second usage of the at-mention in Twitter is the use of the at-reply in the first position of the tweet. An at-reply identifies that tweet as conversational as it indicates that the individual who is authoring the tweet is directly addressing the individual whose handle is in the first position in the tweet. There can only be one handle in the first position of the tweet and therefore this allows for a network analytic approach to examining at-reply activity. The analysis of these two usages of the Twitter handle in the tweet allows for an understanding of who is being talked about and who is being talked to.

The dataset used for addressing this research question was created by extracting the tweets that contained an at-mention from all of the collected tweets (At-Mention Dataset). The at-reply dataset is a subset of the at-mention dataset that includes only tweets that had an at-reply. These datasets were previously visualized in Figure 3. Table 14 provides an overview of the datasets, methods, scripts and output files used for each of the findings related to research question 2.

<table>
<thead>
<tr>
<th>Finding</th>
<th>Dataset(s)</th>
<th>Method</th>
<th>Script</th>
<th>Output File</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>At-mention Dataset</td>
<td>Frequency Analysis; Coding of Accounts</td>
<td>TZScript</td>
<td>mentioncounts.csv</td>
</tr>
<tr>
<td>7</td>
<td>At-reply Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
<tr>
<td>8</td>
<td>At-reply Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
<tr>
<td>9</td>
<td>At-reply Dataset</td>
<td>Network Analysis; Frequency Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
</tbody>
</table>

**At-mention Analysis (Finding 6)**

Analysis of at-mention activity was done using the at-mention subset. The at-mention subset was processed using TZScript and the frequency of at-mentions was analyzed using the output file

---

36 A description of the functionality of the two scripts can be found in Appendix D and Appendix E. These scripts can be found at https://github.com/cmascaro/Dissertation-Scripts

37 A description of the specific output files can be found in Appendix D.
The aggregation of handles and analysis of the frequency in the data highlights those individuals who are the subjects of the discourse in the dataset. This type of analysis is akin to traditional forms of communication analysis that identifies how many times a name appears or is talked about in traditional media, but is adapted to the social media environment (Bruns & Burgess, 2011; Bruns & Burgess, 2012).

In an effort to extend on the frequency analysis, the highly occurring accounts in at-mentions were coded to identify the types of accounts that were prominent in the corpus. Accounts were coded based on the profile of the Twitter account with the intent of identifying if the account was associated with a campaign, media or individual. This analysis was conducted through an iterative process where the 128 most frequently occurring accounts were coded based on profile attributes and tweet content to identify if the account was a bot (Chu, Gianvecchio, Wang, & Jajodia, 2010). The top 128 accounts were included in the analysis as many of the accounts after the top 128 were difficult to categorize as being associated with one of the groups. The top 128 accounts represent being mentioned over 27,000 times and accounted for 41.5% of the total at-mention activity in the complete dataset.

In total, there were eight codes that were identified based on the profiles (Campaign, Celebrity, Individual, Interest, Media, Parody, Spam and Tech). The codes that were generated were then applied to a similar analysis conducted on the accounts that had the highest at-reply in-degree to identify the types of accounts that were the most prominent recipients in the conversational activity.

**At-reply Network Analysis (Finding 7)**

The at-reply dataset was created by identifying the at-mentions that appeared in the first position of a tweet and then generating an edgelist where the author of the tweet was the head and the user that was mentioned in the tweet was the tail (Bruns & Stieglitz, 2012b; Honeycutt & Herring, 2009; Huberman et al., 2008; Rossi & Magnani, 2012). For example, the following tweet “@BarackObama, thank you for all you have done #teamobama2012” authored by “@maizeandblue” would create a directed edge: maizeandblue → barackobama. This would be one line in an edgelist that comprised all at-reply activity.

---

38 The handles were converted to lowercase to take into account that some individuals may have capitalized parts of the username incorrectly. Since Twitter usernames are case independent this approach does not affect analysis.
This finding used the output file replytoedge.csv that was generated in the TZScript processing of the at-reply subset. This contained an edgelist used for analysis. The in-degree versus out-degree measures for each individual in the network were analyzed to identify individuals who had high in-degree and high out-degree. An individual who initiated a high number of conversations would be classified as having high out-degree and an individual who was the recipient of a high number of at-replies would have high in-degree. This methodological approach allowed for an examination of those who appeared frequently in the at-reply network. The codes generated in the analysis from finding 6 were then applied to this data.

**Conversational Activity Identification (Finding 8)**

In an effort to understand highly occurring “conversations” the directed edges constructed for the at-reply network analysis were aggregated into pairs using Excel. Therefore, if user @maizeandblue directed 1,000 at-replies towards @BarackObama, the at-reply dataset would have a line that included @maizeandblue | @BarackObama | 1,000. This aggregation allowed for an understanding of the most frequently occurring activity at the conversational pair unit of analysis. This would identify the presence of repeated interaction. The 725 most frequently occurring pairs (representing pairs that occurred more than 100 times) were then examined to identify the most frequently occurring conversations. This approach allowed for a better understanding of repeated interactions and who the recipient of these repeated interactions were.

**Engagement Identification (Finding 9)**

An analysis of those who were on both sides of the conversation was conducted in Excel using the replytoedge.csv file to examine those. This approach took place in three parts. The first analysis was of those individuals who only sent an at-reply, the second was that of those who only received an at-reply and the third was of those who were on both sides of the at-reply activity.

In order to examine the number of times that individuals in each dataset used the at-reply, summary statistics were produced. These statistics illustrated that there were differences in the number of times that individuals used the at-reply depending on if they were present on both sides of the conversation or only present on one side of the conversation. The temporal aspect of this analysis was not conducted as it is unknown the amount of missing data, but this finding does illustrate differences in at-reply use building on previous research (Honeycutt & Herring, 2009).
**Question 3:** To what extent do the URL and retweet syntactical features in Twitter facilitate information exchange surrounding the 2012 Presidential Election?

The third research question examines how information exchange occurs through the use of URLs and retweet behavior. The presence of a URL in a retweet compared to an original message or an at-reply indicates different user behavior. There are two datasets used to address this research question. The first dataset is the subset of the complete dataset that contains all retweets (Retweet Subset) and the second is the subset of the complete dataset of tweets that contain URLs (URL Subset).

There were three datasets (the complete dataset and two subsets) of the URL dataset used to address how URLs facilitated different types of information exchange in the dataset. The first dataset of URLs consists of all of the URLs that were shared in the dataset in every message (referred to as complete URL dataset). The second dataset consists of all of the URLs that were shared in the dataset that were part of a retweet (referred to as retweet URL dataset). The third dataset of URLs consists of all of the URLs that were part of an at-reply message (referred to as at-reply URL dataset).

Analysis of the retweet dataset identified two different types of retweet behavior. As a result, the retweet dataset was split into two datasets, one with tweets that contained the syntax “RT @[username]” and the other that contained tweets with the syntax “via @[username]” (Kooti, Yang, Cha, Gummadi, & Mason, 2012). These two datasets represent different ways of retweeting information and the differences are further explained in the findings section. Table 15 provides an overview of the datasets, methods, scripts and output files used for each of the findings related to research question 3.

<table>
<thead>
<tr>
<th>Finding</th>
<th>Dataset(s)</th>
<th>Method</th>
<th>Script</th>
<th>Output File</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>URL Dataset; Retweeted URLs; At-reply URLs</td>
<td>Frequency Analysis; Coding of URLs</td>
<td>URLDecodeScript</td>
<td>The output of URL Decode Script is one file with the URLs decoded as described in Appendix E.</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>URL Dataset; Retweeted URLs; At-reply URLs</td>
<td>Frequency Analysis; Coding of URLs</td>
<td>URLDecodeScript</td>
<td>The output of URL Decode Script is one file with the URLs decoded as described in Appendix E.</td>
<td></td>
</tr>
</tbody>
</table>

39 A description of the functionality of the two scripts can be found in Appendix D and Appendix E. These scripts can be found at https://github.com/cmascaro/Dissertation-Scripts

40 A description of the specific output files can be found in Appendix D.
Decoding Shortened URLs (Findings 10-12)

Twitter uses shortened URLs for most of the URLs that are shared in the tweets. Since many URLs in Twitter are also shared as shortened URLs from a URL shortening service such as bit.ly, the process must be done multiple times. As a result of this, a process of decoding the URL is required. In order to do this decoding of a shortened URL a script was written (URLDecodeScript.r) that would ingest a list of shortened URLs and expand them\(^{41}\). The script ingests the linkcounts.csv output of the TZScript and goes through an iterative process to decode the links as visualized in Figure 5. In this study, each URL was decoded three times. The functionality of the script is further detailed in Appendix E.

\(^{41}\) An electronic version of this script can be found at https://github.com/cmascaro/Dissertation-Scripts/blob/master/URLDecodeScript.r
Table 16 identifies the number of URLs in each of the three datasets (complete, retweet, at-reply) and the timeframe in which the URLs were decoded. The decoding of the three datasets was done separately and then compared to ensure validity of the script. The compilation of these three datasets helps to highlight significant differences between how URLs were used as information artifacts in the electoral discourse. The complete URL dataset identifies a form of collective agenda setting (McCombs & Shaw, 1972; McCombs & Shaw, 1974; McCombs, 2005; Roberts, Wanta, & Dzwo, 2002) among all Twitter messages during the electoral period. Analysis of the types of URLs that are shared identifies how individuals were using sources external to Twitter to engage with others and share information. The retweet URL dataset allows for an examination of the types of information that were being shared by individuals through the use of the retweet mechanism.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Decoded URLs</th>
<th>Time Period Decoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>17,356,265</td>
<td>2/7/2013-2/21/2013</td>
</tr>
<tr>
<td>Retweet</td>
<td>10,448,022</td>
<td>2/1/2013-2/6/2013</td>
</tr>
<tr>
<td>At-reply</td>
<td>712,402</td>
<td>1/25/2013-1/29/2013</td>
</tr>
</tbody>
</table>

The actual decoding process was run on a set of Amazon Web Services EC2 instances. A micro EC2 instance was loaded with a subset of each of the datasets and was decoded during the specified time. Since the decoding of each URL needs to send an http request to a server three times to ensure that it is completely decoded the process of decoding millions of URLs is not processor intensive, but time intensive. This problem was addressed by using 20 EC2 instances at once to decode the URLs. All of these EC2 instances were located in Northern Virginia.

---

42 An EC2 instance is a virtual machine that is hosted at an Amazon colocation facility.
43 A “micro EC2 instance” is the smallest instance that could be used and was chosen for cost and the low processing requirements of iterating through a list of URLs.
44 It is unknown if certain URLs may have responded differently to infrastructure that was hosted elsewhere, but for transparency, the location is provided.
In total, approximately 6,900 hours of EC2 compute time (representing 287.5 days of total computing time) was used to decode all of the URLs. This represents a significant amount of computing time that was done in a reasonable amount of time (less than a month) because of parallelized cloud computing resources. The combination of the collection of the data, the decoding of the data and use of cloud computing resources to do this in a shortened amount of time illustrates the unique nature of this dataset and the difficulty of collecting, managing and analyzing these type of “big data” datasets.

**Coding of Base URLs (Findings 10-11)**

One of the difficulties with sharing URLs in social media is that the basic URL includes technology specific information to indicate how it was viewed (mobile versus non-mobile website) and can often include a timestamp or other unique features in its subdirectory structure. This makes the aggregation of the full URLs difficult as many different URLs may point to same story. The base URL is used as the unit of analysis for this section, since the analysis of the specific stories that are referenced is difficult to aggregate and disambiguate. The base URL is the part of the URL that goes up to and includes the top-level domain marker (.com, .org, .edu). For this analysis, all parts of the URL after the top-level domain suffix are stripped off.

Examining the base URLs of the information that is being shared affords the opportunity to understand the sources being used in the discourse. There currently exists no accepted way to code URLs in social media. One of the reasons for this difficulty is that over time it is likely that categorizations for one website may change. Huffington Post started out as a blog and has evolved into a mainstream news source. This study adopts an approach from another research group that examined URLs in Twitter as being user-generated or mass media (Hu et al., 2012) and the type of activity that was being discussed (Himelboim, McCreery, & Smith, 2013). This approach allowed for a classification scheme to be developed that had limited ambiguity in identification. Since the dataset for this study was representative of political discourse, another code “campaign” was also generated to classify campaign related websites that were being shared.

There are only a limited number of user-generated websites, so only the top 50 of each of the base URLs were coded. This allowed for a more robust comparison among the popular base URLs in the three datasets, as there were likely to be significant differences beyond the top 50. The presence of a URL in the top 50 represents being shared at least 1,455 times in the at-reply dataset. Extending on previously used
codes, URLs were coded for whether they were related to a campaign. Websites such as URL shorteners that were unable to be decoded were coded as NA and not discarded in this approach since they represent a large part of the dataset and need to be accounted for by examining the presence of Dead URLs (finding 12).

**Dead URLs (Finding 12)**

One of the shortcomings with social media and the Internet more broadly is that shortened URLs, a necessary component of communications with limited text such as Twitter, often no longer point to the original source as the original source has moved or one of the intermediate shortened URLs is no longer valid. This has many consequences when understanding the type of information being shared, but one of the most significant problems is that the ability to archive and revisit the source material to understand the context of the sharing no longer exists (Scifleet, Henninger, & Albright, 2013). Research has identified that URLs that are shared in social media disappear at a rate of .02% per day (SalahEldeen & Nelson, 2012).

An analysis was done on the URLs that no longer pointed to the original source in the dataset. This approach was based on the only published work that has examined these phenomena (SalahEldeen & Nelson, 2012; SalahEldeen & Nelson, 2013a; Salaheldeen & Nelson, 2013b). A tweet that only resolved to a URL shortener after three decoding attempts was considered a dead URL. This analysis differs slightly from previous research that looked for server response to identify a URL as “dead.” After three iterations of decoding a shortened URL, it is likely that the URL is either dead or originally linked to spam (Klien & Strohmaier, 2012). After three decoding attempts the top base URLs (of the URL shorteners) that contained a URL shortener were aggregated and then examined to understand the URL attrition in the dataset.

**Retweet Network Analysis (Findings 13-14)**

A retweet is a tweet in which an individual includes a tweet in his or her own stream that originates from someone else. This is most commonly done with the syntactical structure of “RT @[username]”, but can also be done a number of other ways depending on user preference or the third party application being used (Kooti et al., 2012). The retweet subset was built by identifying all tweets with “RT @[username]” and “via @[username]” and then processing that dataset using TZScript. The construction of the retweet network is found in the TZScript output retweetedge.csv. This file forms the basis for the retweet analysis that is included in findings 13 and 14.
To build a retweet network, the user that is including the tweet in their stream is treated as the source (or head) and the person being retweeted is treated as the target (or the tail). For example, the tweet “RT @BarackObama, 4 more years!” that was retweeted by @maizeandblue would create a directed edge maizeandblue $\rightarrow$ BarackObama. Similarly, the tweet: “New story about Obama and Romney via @huffpostpol” sent by @maizeandblue would create a directed edge maizeandblue $\rightarrow$ huffpostpol.

Generating these edges creates a weighted, directed network that identifies individuals that frequently retweet or are retweeted by others and builds on previous analysis of retweets using the same approach to construct edgelists (Macskassy & Michelson, 2011; Mustafaraj & Metaxas, 2011; Nagarajan, Purohit, & Sheth, 2010; Starbird & Palen, 2012; Stieglitz & Dang-Xuan, 2012; Suh et al., 2010; Wang, Wang, Zhou, & Zhang, 2012).

In the operationalization of the network, an individual with high out degree is someone that retweets others a lot and an individual with a high in degree is someone who is being retweeted often. Through network analysis it is possible to identify those who are information propagators and those whose information is propagated the most.

**Methods Summary**

The previous chapter presents the complex approach to identifying the queries for the dataset, compiling the dataset and the analytical methods used to parse and interpret the dataset. The richness of the dataset coupled with the depth and range of the analytical methods used creates a set of findings that contribute to a diverse set of literature. Table 17 provides an overview of the each of the datasets, methods, script and output file used to arrive at each finding that are previously detailed.
<table>
<thead>
<tr>
<th>Finding</th>
<th>Dataset(s)</th>
<th>Method</th>
<th>Script</th>
<th>Output File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Complete Dataset</td>
<td>Temporal Frequency Analysis of Syntactical Features</td>
<td>TZScript</td>
<td>syntacticfeatureoverview.csv</td>
</tr>
<tr>
<td>2</td>
<td>Hashtag Dataset</td>
<td>Frequency Analysis; Co-occurrence Analysis</td>
<td>TZScript</td>
<td>hashcounts.csv</td>
</tr>
<tr>
<td>3</td>
<td>#Election2012_Subset; #Debates Subset</td>
<td>Temporal Frequency Analysis</td>
<td>TZScript</td>
<td>timeseries.csv</td>
</tr>
<tr>
<td>4</td>
<td>Debate Days Subset</td>
<td>Frequency Analysis</td>
<td>TZScript</td>
<td>hashcounts.csv</td>
</tr>
<tr>
<td>5</td>
<td>Promoted Hashtag Subset</td>
<td>Time Series Analysis; Frequency Analysis</td>
<td>TZScript</td>
<td>timeseries.csv</td>
</tr>
<tr>
<td>6</td>
<td>At-mention Dataset</td>
<td>Frequency Analysis; Coding of Accounts</td>
<td>TZScript</td>
<td>mentioncounts.csv</td>
</tr>
<tr>
<td>7</td>
<td>At-reply Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
<tr>
<td>8</td>
<td>At-reply Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
<tr>
<td>9</td>
<td>At-reply Dataset</td>
<td>Network Analysis; Frequency Analysis</td>
<td>TZScript</td>
<td>replytoedge.csv</td>
</tr>
<tr>
<td>10</td>
<td>URL Dataset; Retweeted URLs; At-reply URLs</td>
<td>Frequency Analysis; Coding of URLs</td>
<td>URLDecodeScript</td>
<td>The output of URL Decode Script is one file with the URLs decoded as described in Appendix E.</td>
</tr>
<tr>
<td>11</td>
<td>URL Dataset; Retweeted URLs; At-reply URLs</td>
<td>Frequency Analysis; Coding of URLs</td>
<td>URLDecodeScript</td>
<td>The output of URL Decode Script is one file with the URLs decoded as described in Appendix E.</td>
</tr>
<tr>
<td>12</td>
<td>URL Dataset; Retweeted URLs; At-reply URLs</td>
<td>Frequency Analysis; Coding of URLs</td>
<td>URLDecodeScript</td>
<td>The output of URL Decode Script is one file with the URLs decoded as described in Appendix E.</td>
</tr>
<tr>
<td>13</td>
<td>Retweet Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>retweetededge.csv</td>
</tr>
<tr>
<td>14</td>
<td>Retweet Dataset</td>
<td>Network Analysis</td>
<td>TZScript</td>
<td>retweetededge.csv</td>
</tr>
</tbody>
</table>

45 A detailed table of the number of tweets in the dataset and a brief description along with the archived digital location can be found in Appendix B.
46 A description of the functionality of the two scripts can be found in Appendix D and Appendix E. These scripts can be found at https://github.com/cmascaro/Dissertation-Scripts
47 A description of the specific output files can be found in Appendix D.
48 The complete dataset represents all data collected and is the foundation for all of the subsets of data described in chapter 3.
The 14 findings presented in the next chapter represent a significant contribution to knowledge. In addition to these contributions, the methodological approach taken throughout this study is a contribution in and of itself to the literature related to collection, management and analysis of theoretically informed “big social data” (Burgess & Bruns, 2012; Goggins et al., 2012a; Goggins et al., 2010; Goggins et al., 2012b). This contribution is illustrated through the variety of datasets that are derived from the collected data along with the different methodological approaches used to examine the data and arrive at the findings.
CHAPTER 4: FINDINGS

The following chapter presents the findings of a study of 53 million tweets from the 85 days that encapsulate the weeks before the 2012 Republican National Convention (August 20, 2012) through one week following the 2012 Presidential Election (November 13, 2012). The methodological approach used to construct the dataset and conduct the analysis allows for a comprehensive understanding of the activity within Twitter that would not have been possible in a small focused study. The criteria used for constructing the dataset consisted of a set of 68 queries related to the candidates, political parties and events specific to the election. Therefore, the dataset can be characterized as one of political discourse (van Dijk, 1997), as it consists of individuals talking about politics amongst each other and towards candidates.

Overview

Table 18 provides a succinct overview of the research questions, dataset used to address the research questions, general methods and the subsequent findings. This table is meant to provide a guide to the reader and not to capture the nuanced detail presented throughout the chapter.

Table 18: Overview of RQs, Datasets, Methods and Findings

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Dataset⁴⁹</th>
<th>Method⁵⁰</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How does the political discourse related to the 2012 Presidential election manifest itself in a technologically mediated environment?</td>
<td>• Complete Dataset • Hashtag Dataset</td>
<td>• Temporal Analysis of syntactical features • Co-occurrence analysis • Frequency Analysis</td>
<td>• Citizens engaged in political discourse differently within the Twitter in response to acute electoral events (Finding 1a). • The syntactical features used during acute events such as the debates differed from the syntactical features used in the overall dataset (Finding 1b). • Citizens combined hashtags associated with salient news items with election related hashtags to inject news events into the political discourse (Finding 2). • User behavior around hashtags</td>
</tr>
<tr>
<td>a. How does this discourse in Twitter differ during acute events such as debates compared to long-term discourse that occurs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

⁴⁹ A detailed overview of the datasets can be found in chapter 3. A table with the datasets and archived locations can be found in Appendix B.

⁵⁰ A detailed description of the methods used in relation to each finding is found in chapter 3. Appendix C provides an overview of the data fields collected by the TwitterZombie infrastructure and Appendices D and E provide an overview of the scripts used for analysis as detailed in chapter 3. For electronic versions of the scripts please see https://github.com/cmascaro/Dissertation-Scripts.
throughout the election?
b. How does the use of politically oriented hashtags in Twitter facilitate this discourse?

d. How does the use of politically oriented hashtags in Twitter facilitate this discourse?

differed based on whether the hashtag was associated with an acute or non-acute event (Finding 3a).
- Political discourse using hashtags associated with acute events is concentrated around the event (Finding 3b).
- The officially sanctioned hashtag of the four debates was not the most popular throughout all four debates indicating emergent behavior related to hashtags where users adopted their own discourse markers (Finding 4).
- Hashtags promoted by the two campaigns persisted throughout the election at varying rates illustrating a difference between the hashtags promoted by the two campaigns (Finding 5).

2. How does the political discourse that occurs in Twitter using the syntactical features of the at-mention and at-reply surrounding the 2012 Presidential Election identify an emerging participatory public engaged in political discourse?

<table>
<thead>
<tr>
<th>At-mention dataset</th>
<th>Frequency Analysis</th>
<th>Media and campaign related accounts are the most mentioned accounts with campaign related accounts comprising the largest percentage of activity in at-mentions (Finding 6a).</th>
</tr>
</thead>
<tbody>
<tr>
<td>At-reply dataset</td>
<td>Network Analysis</td>
<td>At-mentions of the candidates follow a bursty pattern surrounding salient events where the candidates are the subject (Finding 6b).</td>
</tr>
<tr>
<td></td>
<td>Coding of account types</td>
<td>There was limited adoption of the at-reply indicating limited conversational behavior (Finding 7a).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The most frequent initiators of conversation were individuals and the most frequent recipients of these at-replies were candidates (Finding 7b).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Citizens used the at-reply syntactical feature to direct messages towards campaign and media but this engagement was not reciprocated (Finding 7c).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The most active “conversations” were one-sided and directed towards the candidates and media (Finding 8).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Individuals that received an at-reply initiated more conversations, indicating that political engagement may beget political engagement (Finding 9).</td>
</tr>
</tbody>
</table>
3. To what extent do the URL and retweet syntactical features in Twitter facilitate information exchange surrounding the 2012 Presidential Election?

- URL Dataset
- Retweet Dataset
- Frequency Analysis
- Coding of URL types
- Network Analysis
- Citizens primarily shared URLs associated with user-generated content and mass media (Finding 10).
- The retweeting of URLs inflated their overall presence in the dataset as the most frequently occurring URLs in the corpus had a high presence in the retweet URL dataset (Finding 11).
- Approximately 5% of the URLs no longer resolved to the original location of the information demonstrating the ephemeral nature of social media information (Finding 12).
- The most prolific retweeters (high out-degree in the retweet network) were those who do not appear to be overtly tied to a campaign or media organization (Finding 13a).
- Individuals who were retweeted the most (high retweet in-degree) highest were candidates or related to a campaign (Finding 13b).
- Information that originated on websites associated with new media entities was shared using a different retweet syntax (Finding 14).

Question 1: How does the political discourse related to the 2012 Presidential election manifest itself in a technologically mediated environment?

- How does this discourse in Twitter differ during acute events such as debates compared to long-term discourse that occurs throughout the election?

- How does the use of politically oriented hashtags in Twitter facilitate this discourse?

Political discourse in Twitter during the 2012 Presidential election was concentrated around salient events including the conventions, debates and the day of the election. There were relative lulls at other times during the studied time period that illustrates a limited focus of activity around a hashtag or other syntactical feature. During acute events such as debates, individuals engaged in Twitter differently than during non-acute events by relying on hashtags more than at-mentions and URLs. This reliance on hashtags and decrease in URL usage marks a higher level of commenting on the acute event as opposed to sharing
information about the event. In this instance, Twitter functioned as a public forum for users to comment on real-time activity.

Citizens adopted the Twitter promoted and curated hashtag for the most prominent set of events during the election period, the debates, but other hashtags also emerged to mark the discourse. In the case of acute events such as the debates, most of the activity using #debates was concentrated around the actual day of the event with limited activity occurring beyond those days. This time box of hashtag usage represents the ephemeral nature of event oriented political discourse using event specific hashtags. The extent of how long a hashtag stays active may be a proxy measure for interest in an event and helps to identify the duration of discourse related to a specific topic.

Hashtags purchased by the campaigns for Twitter to highlight on the main page were also frequently used on the day of the promotion and usage persisted throughout other events such as the debates, press conferences and the times immediately preceding and following these electoral events. This illustrates the difference between event specific hashtags and hashtags used to promote a campaign issue. The combination of these findings contributes to the understanding of how political discourse manifests itself temporally during an election. Syntactical features such as hashtags play a significant role in this activity as they allow individuals to include a discourse marker to highlight their contributions to others in the context of the event.

**Finding 1a: Citizens engaged in political discourse differently within Twitter in response to acute electoral events.**

The variance of tweets per day correlates to the occurrence of prominent electoral events. The most popular days in the dataset based on raw tweets count were Election Day, followed by the four debates and the two days during the respective party conventions where President Obama and Governor Romney accepted the party nomination\(^5\). Tweet counts spike immediately preceding, during and following these events (Figure 6). These bursts illustrate that prominent electoral events contributed to a high amount of the activity on Twitter during the election.

---

\(^5\) As discussed in the methods section, the time used to calculate each day is UTC. This is done to standardize times across all analysis. Analysis of the graphs also illustrates distinct activity based on the aggregation by day indicating validity to the use of these times as cut offs.
Analysis of tweets per day only provides a high level view of activity. Analysis of the syntactical features within the dataset at the day unit of analysis allows for a further understanding of how technologically mediated interaction manifested itself during the 2012 Election. Campaign events are usually contained within one day, which reflects and helps to drive the 24-hour news cycle. Focusing on the dataset at the daily unit of analysis mirrors the phenomena of this news cycle and creates an appropriate temporal frame for analysis.

The percentage that a syntactical feature occurs on a given day represents an observation that can be measured and compared over time. When plotting the occurrence of syntactical features by day, relationships between specific features and electoral events begin to emerge. Figure 7 demonstrates the fluctuation in activity on a daily basis for each of the percentages of syntactical features.
There are a number of relationships between syntactical features that are reflected in Figure 7. When the number of retweets drops, the number of at-mentions drops. This is most drastic on 8/20/2012, 9/1/2012, 9/17/2012 and 9/27/2012. These are the four lowest days of retweets in the dataset. These days are notable since there are no conventions, debates, press conferences or other important news events that occur within two days on either side of the days with the lowest percentage of retweets.

The absence of these events coupled with the fact that those who are the most retweeted are candidates and campaign related accounts (finding 13b) suggests that events that are directly related to the election drive retweet activity in this dataset. This correlation is also identified with at-replys and at-mentions, but the variation is not as significant, because the overall percentage of at-replys is not as high as the other syntactical features.

There is also an inverse relationship between the occurrence of hashtags and URLs. During acute events there is a spike in hashtag usage that correlates to a decrease in URL percentage. The four days where this is most salient are the four debate days (discussed in finding 1b). This demonstrates that the usage of hashtags that are commonly associated with acute events may lead to an increase in commenting about the event and in turn led to limited sharing of information external to the stream of discourse.

In the case of the debates, individuals use hashtags as a way to participate in a conversation about an event and remark about what is happening as opposed to sharing URLs or information external to
Twitter about the event. The real-time nature of these types of events may also limit the ability for individuals to share information in the form of a URL and it is possible that information about the event does not yet exist in an external source. Further examining the syntactical feature usage on the days of the debates helps to identify the differences in activity.

**Finding 1b: The syntactical features used during acute events such as the debates differed from the syntactical features used in the overall dataset.**

The usage of syntactical features on the four debate days were similar to each other, but differed from the syntactical feature usage of the complete dataset (Table 19). This indicates that the debates facilitated different activity than the activity that occurred throughout the rest of the dataset. The number of at-mentions, at-replys and retweets were all within two percentage points of the overall dataset average by day on each of the debate days, suggesting that usage of these syntactical features during the debates was similar.

<table>
<thead>
<tr>
<th>Date</th>
<th>Tweets</th>
<th>URL %</th>
<th>Hash %</th>
<th>At-mention %</th>
<th>At-reply %</th>
<th>Retweet %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Dataset</td>
<td>52,487,179</td>
<td>32.59%</td>
<td>50.15%</td>
<td>72.81%</td>
<td>8.78%</td>
<td>55.10%</td>
</tr>
<tr>
<td>10/4/12</td>
<td>2,383,943</td>
<td>16.74%</td>
<td>64.29%</td>
<td>70.04%</td>
<td>7.41%</td>
<td>54.89%</td>
</tr>
<tr>
<td>10/12/12</td>
<td>1,780,518</td>
<td>22.41%</td>
<td>61.85%</td>
<td>71.69%</td>
<td>7.83%</td>
<td>56.40%</td>
</tr>
<tr>
<td>10/17/12</td>
<td>2,177,364</td>
<td>19.73%</td>
<td>66.35%</td>
<td>71.27%</td>
<td>8.09%</td>
<td>55.01%</td>
</tr>
<tr>
<td>10/23/12</td>
<td>2,098,484</td>
<td>20.48%</td>
<td>67.33%</td>
<td>71.82%</td>
<td>8.23%</td>
<td>55.95%</td>
</tr>
</tbody>
</table>

The difference in syntactical feature usage during the four debate days is most prominently represented by a lower percentage of URLs being shared and a higher percentage of tweets that contained hashtags. The lower percentage of URLs suggests that individuals were not sharing external information during the debates. Instead of sharing external information individuals are commenting on the event using the event specific hashtag (#debates). This specific behavior suggests the reason for the inverse relationship between hashtags and URLs on debate days. Further analysis of the usage of #debates is presented in finding 3b.
Finding 2: Citizens combined hashtags related to salient news items with election related hashtags to inject news events into the political discourse.

The most popular hashtags in the corpus are those directly associated with the candidates and the campaign (Table 20). Event specific hashtags such as those associated with debates and conventions also occurred frequently. Only one of the campaign promoted hashtags (#forward) was one of the 15 most frequently occurring hashtags. Promoted hashtags are those paid for by an organization and featured on the front page of Twitter for one day.\(^5\) The lack of other promoted hashtags in the most frequently occurring hashtags illustrates that other hashtags were more widely adopted.\(^3\)

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>4,127,748</td>
<td>8.58%</td>
</tr>
<tr>
<td>obama2012</td>
<td>2,674,218</td>
<td>5.56%</td>
</tr>
<tr>
<td>Romney</td>
<td>2,126,806</td>
<td>4.42%</td>
</tr>
<tr>
<td>election2012</td>
<td>1,737,448</td>
<td>3.61%</td>
</tr>
<tr>
<td>romneyryan2012</td>
<td>1,722,473</td>
<td>3.58%</td>
</tr>
<tr>
<td>Tcot</td>
<td>1,469,682</td>
<td>3.06%</td>
</tr>
<tr>
<td>Debates</td>
<td>1,051,814</td>
<td>2.19%</td>
</tr>
<tr>
<td>debate</td>
<td>1,043,494</td>
<td>2.17%</td>
</tr>
<tr>
<td>dnc2012</td>
<td>1,025,728</td>
<td>2.13%</td>
</tr>
<tr>
<td>voteobama</td>
<td>883,246</td>
<td>1.84%</td>
</tr>
<tr>
<td>vote</td>
<td>817,921</td>
<td>1.70%</td>
</tr>
<tr>
<td>p2</td>
<td>799,018</td>
<td>1.66%</td>
</tr>
<tr>
<td>gop2012</td>
<td>708,099</td>
<td>1.32%</td>
</tr>
<tr>
<td>forward</td>
<td>574,973</td>
<td>1.20%</td>
</tr>
<tr>
<td>debate2012</td>
<td>495,264</td>
<td>1.03%</td>
</tr>
</tbody>
</table>

\(^5\) This is examined further in finding 5.

\(^3\) As identified in the methods section, all hashtags and at-mentions were converted to lowercase for analysis, as Twitter is case independent.
There are two hashtags (in bold in Table 20) that were not explicitly collected, #tcot and #p2. These two hashtags are associated with two Twitter specific, political campaigns, Top Conservatives on Twitter (#TCOT) and the progressive response to that campaign (#p2). These two hashtags are two of the most widely used political hashtags in the corpus and are used to demonstrate an ideological affiliation even though there has been evidence of the coopting of these hashtags to promote ideologies counter to the intended purpose. Their salience in the dataset illustrates that they co-occur frequently with hashtags and other syntactical features related to the 2012 Presidential Election as these hashtags are inherently political.

There are many hashtags that occurred frequently in the dataset that were not explicitly collected. These were present in the dataset as a result of their co-occurrence with other syntactical features that were explicitly collected. The fact that some hashtags were explicitly collected and others were not may introduce a collection bias since a hashtag is more likely to occur at a higher frequency if it was explicitly collected. As described in the methods in chapter 3, extracting the terms that were explicitly collected and then examining the most frequently occurring hashtags in the new dataset can control for some of this bias. This extraction leaves the hashtags that most frequently co-occur with syntactical features that were explicitly collected.

The most frequently occurring hashtags in the extracted dataset illustrate the presence of salient news events and other hashtags used for partisan affiliation purposes such as #p2, #tcot, #lyhnbty54 (Table 21). The top 15 hashtags that were not explicitly collected represent 10.04% of the total hashtag usage in the dataset.

**Table 21: Most Frequently Occurring Hashtags Not Explicitly Collected**

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tcot</td>
<td>1,469,682</td>
<td>3.06%</td>
</tr>
<tr>
<td>p2</td>
<td>799,018</td>
<td>1.66%</td>
</tr>
<tr>
<td>teamobama</td>
<td>476,640</td>
<td>0.99%</td>
</tr>
<tr>
<td>Gop</td>
<td>438,410</td>
<td>0.91%</td>
</tr>
<tr>
<td>Teparty</td>
<td>261,380</td>
<td>0.54%</td>
</tr>
<tr>
<td>Tlot</td>
<td>179,005</td>
<td>0.37%</td>
</tr>
<tr>
<td>Rnc</td>
<td>165,185</td>
<td>0.34%</td>
</tr>
</tbody>
</table>

54 This hashtag stands for, “Let Your Heart Not Be Troubled” and is a biblical reference that Sean Hannity popularized in hashtag form. An analysis of this hashtag is outside the scope of this study.
One of the most frequently occurring hashtags that was not collected explicitly is #sandy. This hashtag is associated with Hurricane Sandy, which hit the Northeastern United States right before the election in late October. This event was important to the election for two reasons. First, the event caused widespread damage in New York and New Jersey and triggered a Federal emergency response with President Obama visiting the site the week before the election. Second, the conservative governor of New Jersey Chris Christie, a key Romney supporter, voiced support for President Obama’s disaster response efforts.

Another one of the hashtags that was prominent in the corpus was #benghazi. This hashtag was associated with the September 11, 2012 terrorist attack on the United States consulate in Benghazi, where the United States Ambassador to Libya and three other State Department Officers were killed. This was a prominent campaign issue for the conservative media because they suggested that the campaign and White House were covering up what actually happened. Further, the hashtags associated with #ohio, a key battleground state, and #4moreyears, a hashtag that emerged following Obama’s victory as frequently occurring.

The high occurrence of these hashtags even though they were not explicitly collected illustrates that citizens used them in combination with electorally specific hashtags. This illustrates the electoral significance of these news events even though they were not major campaign policy issues or planned events.

**Finding 3a: User behavior around hashtags related to acute events was more concentrated around the event compared to hashtags related to non-acute events whose use was distributed over time.**

Temporal analysis of hashtag usage allows for the examination of the difference between hashtags associated with acute events as opposed to hashtags associated with non-acute events. In the course of this
analysis, we see that two types of hashtags were identified, those that persist over the whole time period and those that are mainly used during specific events.

The hashtag #election2012 was not officially promoted as a hashtag until the day before Election Day when Twitter aggregated tweets with the hashtag on the website. Though not promoted until the end of the campaign, #election2012 was used throughout the time period (Figure 8). The #election2012 hashtag represents a hashtag that was not unique in its construction (most Americans are used to the elections being referred to as election2012 or decision2012), but even without the promotion by Twitter or a campaign, this hashtag had a steady use with bursts during campaign events with the most significant burst occurring on Election Day (when it was promoted).

![tweets with #election2012](image)

Figure 8: Frequency of #election2012 Tweets by Day

These minor bursts of activity occurred during the party conventions and during the three Presidential debates. There is only a limited spike of activity on the day of the VP debate, but it is still noticeable compared to the other days during the time period. The #election2012 represents a hashtag that emerged organically early in the electoral period and was used more frequently through promotion online.
Finding 3b: Political discourse using hashtags associated with acute events is concentrated around the event.

Analysis of the official hashtag of the most salient acute events, the four debates, represents a concentration of the hashtag on the actual days of the events with limited use beyond those days. The days with the highest usage of the #debates hashtag are the days of the debate. Even though the four debates represented only 4 of 85 days of activity, the total number of tweets on those four days represented 16.1% of the total tweets in the dataset. This illustrates a concentration of the overall electoral discourse on those days beyond just the use of the hashtag. Figure 9 illustrates the number of tweets per day over time for the hashtag #debates.

Figure 9: Frequency of #debates by Day

On the four debate days, the percentage of tweets with the officially sanctioned hashtag ranged from 7.6% on the day of the Vice-Presidential Debate to 14.60% on the day of the second Presidential debate (Table 22). 93% of the total usage of #debates, occurred on the days of the debates. Because of the time used in Twitter, most of the debates occurred at 1am UTC on the reflected date in the dataset. Taking this into account and incorporating the number of times that #debates was used the day before the debate in the dataset, over 98% of the hashtag usage was accounted for within the 48 hours of the debate. This

---

55 Assuming a normal distribution of tweets, approximately 3.85% of tweets would have occurred on those four days.
highlights that debate discourse was concentrated around the debate and did not persist through other parts of the time period.

### Table 22: Percentage of Overall Tweets with #debates

<table>
<thead>
<tr>
<th>Event</th>
<th>Percentage of Overall Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Debate 1</td>
<td>11.60%</td>
</tr>
<tr>
<td>Debate 2</td>
<td>14.60%</td>
</tr>
<tr>
<td>Debate 3</td>
<td>11.90%</td>
</tr>
<tr>
<td>VP Debate</td>
<td>7.60%</td>
</tr>
</tbody>
</table>

Finding 4: The officially sanctioned hashtag of the four debates was not the most popular hashtag throughout the four debates indicating emergent behavior related to hashtags where users adopted their own discourse markers.

The officially sanctioned hashtag for the debates, #debates, was only the most popular hashtag in the dataset during 2 of the 4 debates. In addition to #debates a number of venue specific hashtags beyond those that were promoted emerged. Analysis of the most popular hashtags used during the days of the Presidential debates highlights an interesting splinter in hashtag usage.

The top 10 most frequently occurring hashtags for each of the days of the four debates varies by debate (Table 23). As discussed in the data collection section, some of the debate venues and sponsors promoted their own debate specific hashtag and this promotion may have led to some fragmentation of hashtag use.

### Table 23: Top Hashtags by Debate

<table>
<thead>
<tr>
<th>Debate 1</th>
<th>Debate 2</th>
<th>Debate 3</th>
<th>VP Debate</th>
</tr>
</thead>
<tbody>
<tr>
<td>debates</td>
<td>debates</td>
<td>debate</td>
<td>vpdebate</td>
</tr>
<tr>
<td>debate</td>
<td>debate</td>
<td>debates</td>
<td>debates</td>
</tr>
<tr>
<td>debate2012</td>
<td>romney</td>
<td>obama</td>
<td>biden</td>
</tr>
<tr>
<td>obama</td>
<td>obama</td>
<td>debate2012</td>
<td>debate</td>
</tr>
<tr>
<td>romney</td>
<td>debate2012</td>
<td>romney</td>
<td>romneyryan2012</td>
</tr>
<tr>
<td>denverdebate</td>
<td>obama2012</td>
<td>lynndebate</td>
<td>teamjoe</td>
</tr>
<tr>
<td>obama2012</td>
<td>presidentialdebate</td>
<td>obama2012</td>
<td>tcot</td>
</tr>
</tbody>
</table>
In the case of the first debate, the official press release by The University of Denver identified the official hashtag as #dudebate, but this hashtag was only used 850 times. In contrast, the hashtag #denverdebate, which was promoted by some media outlets was the sixth most popular hashtag during the day of the debate. This coupled with the late announcement by Twitter that #debates would be the official hashtag (it was announced the night before the first debate) led many individuals to use a variety of hashtags to mark their discourse during the nights of the debate.

Hashtag usage during acute events is a mixture of event and subject focused hashtags. In addition to the hashtags that users used to engage in the context of the debate they also used candidate-specific hashtags to comment about the candidates and affiliate with the campaign. During the third debate, the promoted hashtag #cantafford4more was the ninth most popular hashtag on that day. This is examined further in the next section, but illustrates how a promoted hashtag becomes popular in the context of one event.

On the day of the vice-presidential debate, 412,909 tweets (23.2%) contained the hashtag #vpdebate whereas 133,073 tweets contained #debates even though Twitter promoted #debates as the official hashtag. The hashtag #vpdebate emerged organically during the vice-presidential debate. This suggests that users may have viewed #debates as a hashtag associated only with the Presidential debates even though it was still promoted as being associated with the Vice-Presidential Debates.

The Vice-Presidential Debate is often viewed as less popular, but the number of tweets that occurred on the day of the debate was not less than on the other three debate days. The day of the vice-presidential debate was the fifth most popular day of activity, behind the three Presidential debates and Election Day. This illustrates that all of the acute events were popular compared to other days without prominent electoral events. The varying usage of hashtags for major events illustrates the need to track multiple streams of discourse and also illustrates the difficulty of promoting one hashtag for aggregating discourse even by the technological platform hosting the activity.

---

Organic emergence is determined by the lack of a media or other coordinated campaign to promote the hashtag.
Finding 5: Hashtags promoted by the two campaigns persisted throughout the election at varying rates illustrating a difference between the hashtags promoted by the two campaigns.

Each morning at 6am Twitter was checked to see if one of the campaigns chose to promote a hashtag on that day. These hashtags were added to collection upon identification and collected through the end of the election period. Five of the hashtags promoted by the campaigns were included in dataset: #romneyshambles, #forwardnotback, #forward, #forward2012, #cantafford4more. Promoted hashtags are those that can be paid for by an individual or a group. This promotion is intended to encourage engagement or highlight an event. The paid promotion places the hashtag at the top of the trending list that a user sees when navigating Twitter.

Twitter users used promoted hashtags by the two campaigns differently. President Obama’s two most popular promoted hashtags were used throughout the election and on Election Day (#forward, #forward2012). Governor Romney’s most popular promoted hashtag (#cantafford4more) was concentrated around the debates, which suggests that the hashtag did not have the same level of persistence that those used by President Obama did.

Each promoted hashtag has a tweet associated with it and that tweet is featured at the top of the search results for that hashtag. In the case of the election, there was also a hashtag promoted by Twitter, specifically for the debates (#debates) that allowed for a specific page that curated all debate activity on Twitter. Examining the temporal frequency of the promoted hashtags illustrates how they emerge and disappear relative to their promotion and electoral events.

The hashtag #forward2012 was promoted by the Obama campaign following his acceptance of the nomination at the Democratic National Convention (specifically on 9/7/2012) and had some minor bursts of usage, specifically on Election Day. A similar hashtag #forward was promoted the day before the election. Figure 10 illustrates that #forward was used throughout the campaign even though it was not promoted until the end of the time period. The two hashtags peaked during similar timeframes with #forward and #forward2012 both popular during the Democratic National Convention even though #forward2012 was the only hashtag being promoted.\(^{57}\)

---

\(^{57}\) This is likely the result of #forward being a shortened version of #forward2012
The hashtag #forward was used approximately six times as much on Election Day and the day immediately following President Obama’s victory. The minor peaks of #forward during the month of October occurred on Presidential debate days. The co-occurrence of these with the debates illustrates the adoption of the hashtag as an affiliation during the debate. There was also a moderate amount of activity each day following the DNC where the hashtag was initially promoted illustrating a slight persistence of activity beyond the day of promotion.

Governor Romney’s campaign also utilized the promoted hashtag feature of Twitter. Figure 11 displays the frequency of the promoted hashtag #cantafford4more. This hashtag was used less than 10 times from August 20th until Governor Romney’s campaign promoted it on the night of the first debate. Governor Romney’s tweet from October 3, that was used in combination with the promotion of the hashtag read “Another term for @BarackObama will bring more taxes, regulations, and debt that have ground our recovery to a halt. #CantAfford4More”. This tweet summarized a key Romney campaign talking point and illustrates how the platform can be condensed and represented using a hashtag.

---

58 This hashtag was not collected until the promotion was identified on the first night of the debate, but its lack of presence in the dataset suggests that it did not co-occur with other hashtags or syntactical features before the day of the first debate.
Figure 11: Temporal Distribution of #cantafford4more

There is an immediate peak of activity of #cantafford4more following the initial promotion of the hashtag. The usage of the hashtag fluctuated with the most significant peaks occurring on the four debate days and in the days leading up to that. Unlike the promoted hashtags by President Obama, there was not a significant peak on Election Day. This lack of a burst on Election Day and relative limited use beyond the last debate may signify a lack of enthusiasm surrounding the hashtag and may illustrate a proxy measure for campaign enthusiasm or it may reflect campaign staff efforts to promote it.

The difference in how the promoted hashtags from the two campaigns were used illustrates the differences between the two campaigns effects on the user base. President Obama’s promoted hashtags persisted throughout the election up until Election Day with bursty activity surrounding the debates. Governor Romney’s promoted hashtags (in particular #cantafford4more) was only used heavily during the time period of the debates (and at a much lower rate than President Obama’s) and was not used heavily in the lead up to the election. This may illustrate a dwindling enthusiasm as time passed for Governor Romney. The bursts of activity on days where the promoted hashtags were not even promoted does show the ability for a hashtag to “stick” and continue to be used even at a smaller frequency illustrating that the discourse markers that were promoted persisted in the activity.
**Question 2:** How does the political discourse that occurs in Twitter using the syntactical features of the at-mention and at-reply surrounding the 2012 Presidential Election identify an emerging participatory public engaged in political discourse?

The at-mention in Twitter allows an individual to highlight a tweet to another individual by including their handle in the text of the tweet. This alerts the other individual through Twitter that they have been mentioned in another tweet. Citizens can include the at-mention in the first position of the tweet to indicate that the tweet was directed at another account. In the 2012 Election, citizens primarily used the at-mention to talk about media and campaign related accounts. At-mentions of the candidates correlated to their presence in salient events such as debates or when they formally accepted the party nomination. This demonstrates that individuals were using the at-mention to talk about the primary subjects of the election (the candidates). Citizens used the at-reply syntactical feature to reach out to the candidates, campaign related accounts and the media, but this outreach did not garner any response.

The combination of talking about the candidates using at-mentions and talking to them using the at-reply illustrates that citizens used the technology to technologically engage with candidates. This attempted engagement using the at-reply and talking about the candidates and campaigns reflects a participatory public interested in political discourse. The most frequent conversations were one sided and involved candidates as the recipient of the message. The technology afforded a perceived technological access to the candidates and campaigns, but in the case of the 2012 Presidential Election, the outreach from citizens to candidates was not reciprocated, as there was no evidence of candidates replying to or mentioning citizens in their tweets.

**Finding 6a:** Media and campaign related accounts are the most mentioned accounts with campaign related accounts comprising the largest percentage of activity in at-mentions

In total, 72.81% of tweets contained an at-mention making it the most used syntactical feature in all of the collected data highlighting the fact that Twitter is used to talk to and about other individuals. Media and campaign related accounts are the most mentioned accounts with campaign related accounts comprising the largest percentage of at-mention activity (Table 24). The 128 most frequently mentioned accounts represent 41.5% of the total at-mention activity (and were mentioned at least 27,000 times). Media accounts represent 36.72% of these accounts, 22.66% percent are campaign-related and 17.19% belong to a mix of online and offline celebrities. Celebrity accounts include actors or musicians who had
publicly supported one of the candidates or participated in a campaign event. The celebrity account presence in the discourse highlights the fact that they were related to the election in some way and not injecting themselves as a way of self-promotion in a popular event such as the election.

<table>
<thead>
<tr>
<th>Type</th>
<th>Account Occurrences</th>
<th>Percentage</th>
<th>Tweets</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campaign</td>
<td>29</td>
<td>22.66%</td>
<td>16,688,375</td>
<td>73.08%</td>
</tr>
<tr>
<td>Celebrity</td>
<td>22</td>
<td>17.19%</td>
<td>1,419,409</td>
<td>6.22%</td>
</tr>
<tr>
<td>Individual</td>
<td>2</td>
<td>1.56%</td>
<td>63,916</td>
<td>0.28%</td>
</tr>
<tr>
<td>Interest</td>
<td>7</td>
<td>5.47%</td>
<td>329,512</td>
<td>1.44%</td>
</tr>
<tr>
<td>Media</td>
<td>47</td>
<td>36.72%</td>
<td>2,887,573</td>
<td>12.64%</td>
</tr>
<tr>
<td>Parody</td>
<td>17</td>
<td>13.28%</td>
<td>1,298,168</td>
<td>5.68%</td>
</tr>
<tr>
<td>Spam</td>
<td>1</td>
<td>0.78%</td>
<td>38,208</td>
<td>0.17%</td>
</tr>
<tr>
<td>Tech</td>
<td>3</td>
<td>2.34%</td>
<td>111,128</td>
<td>0.49%</td>
</tr>
</tbody>
</table>

Campaign accounts only received 22% of the total mentions by aggregating by code, but they were responsible for 73% of the total tweets with mentions. By aggregating based on the number of tweets for each code as opposed to the number of accounts, it was identified that 73.08% of the total tweets with an at-mention, mentioned a campaign related account and 12.64% mentioned a media organization (Table 24). This illustrates the high visibility of campaign related accounts in the dataset. This also indicates that campaign accounts are disproportionately represented based on the percentage of their total at-mentions. This gives a more accurate perspective of the number of mentions of each account type.

Though the media account types are the second most frequently occurring in the dataset, the total percentage is approximately 1/6 of that of the campaign tweets. This illustrates that there are a number of media accounts that are mentioned frequently, but there is a limited concentration of activity compared to the candidate accounts. This skew illustrates the ability for a small subset of accounts, ones associated with the campaigns, to dominate activity at the raw tweet count level.
Finding 6b: At-mentions of the candidates follow a bursty pattern surrounding prominent electoral events where the candidates are the subjects of the event.

The percentage of tweets that mention a candidate on a daily basis fluctuates in relation to electoral events (Figure 13). The most notable fluctuations occur when the candidate is the most prominent figure in the event. In political discourse during an election season, the most prominent events are the debates. Given the nature of Presidential Elections in the United States, the Presidential Candidate is often the subject of the most discussion and activity in the news. This phenomenon is reflected in the dataset.

The Presidential candidates comprise a significant portion of the activity on a daily basis with President Obama being mentioned in approximately 17% of all tweets on a daily basis. Outside of the vice-presidential debates, the mentions of the vice-presidential candidates are limited. The Vice-Presidential candidates were mentioned in less than 5% of tweets on all days except for the Vice-Presidential debate and the day that Vice President Biden made a speech at the DNC. This illustrates the salience of the top of the ticket in the electoral discourse even though the vice-presidential candidates had the same access to the technology to engage and share information and individuals had the same ability to mention them in the dataset.

Figure 12: Candidate At-Mention Percentage by Day
President Obama is mentioned in 8.5% more of tweets on a daily basis than Mitt Romney during the election period (mean 17.38% compared to mean 8.84%) (Table 25). The vice-presidential candidates are seldom mentioned except for the day of the vice-presidential debate. This higher percentage reflects the salience of the President in daily discourse as a result of his incumbent status. The frequency of at-mentions may also be a result of the fact that President Obama tweeted 1,917 times compared to Mitt Romney’s 293 times. This higher activity ensured that President Obama was more consistently in the timelines of the followers and led to more retweets and exposure to the ideas of the President.

Table 25: Candidate At-Mention Summary Statistics

<table>
<thead>
<tr>
<th>Syntactical Feature</th>
<th>Minimum Percentage</th>
<th>Maximum Percentage</th>
<th>Mean</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>@BarackObama</td>
<td>9.03%</td>
<td>30.92%</td>
<td>17.38%</td>
<td>16.86%</td>
<td>21.89%</td>
</tr>
<tr>
<td>@MittRomney</td>
<td>2.69%</td>
<td>13.68%</td>
<td>8.84%</td>
<td>8.99%</td>
<td>10.99%</td>
</tr>
<tr>
<td>@JoeBiden</td>
<td>.1%</td>
<td>8.86%</td>
<td>.83%</td>
<td>.56%</td>
<td>8.76%</td>
</tr>
<tr>
<td>@PaulRyanVP</td>
<td>.18%</td>
<td>8.26%</td>
<td>1.59%</td>
<td>1.26%</td>
<td>8.08%</td>
</tr>
</tbody>
</table>

Another explanation for the higher presence of @BarackObama compared to the presence of @MittRomney may be that individuals did not use Mitt Romney’s official Twitter handle and instead referred to him by name in their tweets. @BarackObama was collected 8,592,370 times, whereas @MittRomney was collected 3,908,195 times. The number of times that Mitt and Romney (which would also include the handle) was collected was 10,400,391 whereas “Barack AND Obama” was collected 4,114,765 and “President and Obama” was collected 6,630,349.

This difference illustrates that those talking about each of the candidates used the technology differently to do so. Those talking about President Obama used the technologically prescribed method (his Twitter handle) to highlight the tweets to him, whereas those talking about Governor Romney referred to him by name and not his Twitter handle meaning that searches for his handle would not bring up these tweets and would also not specifically highlight these tweets to anyone from his staff following the discourse. This could be the result of the fact that President Obama tweeted more than Governor Romney and that President Obama’s presence on Twitter dated back to the 2008 election, but it also could reflect a difference in the way that the electorate perceived the candidates.
Finding 7a: There was limited adoption of the at-reply indicating limited conversational behavior.

In total there were 4,606,922 conversational pairs (8.78% of all tweets) in the dataset. This represents the least used syntactical feature of all of the syntactical features studied. 93.2% of the individuals that used the at-reply used it five times or less and 67.4% used it only once. There were 1,547,474 unique source handles and 719,772 unique target handles. This illustrates that there is a small set of individuals who are provoking conversation with others who never respond. The most frequent conversation initiators (those with high out-degree in the at-reply network) were individuals not tied to an organization or campaign. The most frequent conversation recipients (those with high in-degree in the at-reply network) were those directly tied to the campaign. None of the campaign related accounts replied to an at-reply message (finding 7c).

The percentage of tweets with an at-reply on a daily basis varied from 6.00% on November 7, 2013, the day after the election to 15.30% on November 11, 2013, five days after the election (Figure 13). It is notable that the day of the election (November 6) is the second lowest day for conversational behavior, whereas the second highest day in conversational behavior is August 24th. This is a noteworthy news day as there was a shooting at the Empire State building, which sparked discourse related to gun control. Gun control is traditionally a lightning rod topic that leads to increased activity in the news and this is reflected in the data. This is also in the run up to the first convention of the election season and a short time after Paul Ryan was announced as the VP candidate, which also may have led to an increase in activity.

![At-reply % By Day](image)

Figure 13: At-reply Percentages By Day
There are also distinct patterns of conversational activity that relate to acute events. Similar to the inverse relationship noted in finding 1a between URLs and hashtags, there appears to be a relationship between at-replys and acute events. There are distinct dips in at-reply activity during the two party conventions, the four debates and on Election Day. These days all had significant amounts of activity that were facilitated by the use of a hashtag.

The data suggests that the acute events led to less conversational behavior and more broadcast commenting on the activity. This is likely the result of the high volume of activity on those days. Given the significant number of messages on each of the days of the acute events, the ability to read a message and respond to it using an at-reply is lower. Although there are some dips during acute events, there are distinct patterns of bursty activity immediately before and immediately following the debate. This further suggests that the acute events brought out different behavior of individuals as illustrated in finding 1b.

**Finding 7b:** The most frequent initiators of conversation were individuals and the most frequent recipients of these at-replys were candidates and the media.

There were 3,071,676 unique pairs of conversational activity. This indicates repetitive activity between some pairs of individuals as the statistics illustrate that this behavior was one-way. The source or head for the conversational pair is the individual who initiated the interaction by including the target or the tail’s Twitter handle in the first position of the tweet text. This indicates that the individual was directing the tweet to the handle. The most applicable approach to analyzing this data from a network perspective is examining the in-degree and out-degree since it is a disconnected network and other measures such as betweenness and closeness would not be applicable.

The concentration of activity of those who initiated conversation (out-degree) was more diffuse. Table 26 represents the top 10 individuals in at-reply out-degree. The most frequent conversation initiator was an account that was tweeting and as of May 2013 is still tweeting random messages often in Japanese directed towards @BarackObama. This individual had over 186,000 tweets as of May 2013 and based on the profile and frequency of activity (every minute for 24 hours a day) it is likely that the account is an automated bot.
The most notable characteristic about this account is that it was responsible for two and a half times the amount of at-replies initiated than the second most prolific at-reply initiator. The second most frequently occurring account also appears to be a bot. This account also has more tweets than the third most frequently occurring account. This illustrates that outliers in this data are automated Twitter bots that are programmed to tweet at specified intervals. The next accounts that were high in at-reply out-degree were prolific Twitter users, but analysis of their profiles and other tweets suggests that they were real individuals and not bots.

Those that received at-replies were much more concentrated. The two Presidential candidates had the highest in-degree in the at-reply network. In total, the four candidates received 37.6% of the at-replies (Table 27). Barack Obama and Mitt Romney were significant outliers in the context of the at-reply activity. This concentration illustrates the prominence of these accounts in the political discourse. The candidates were not only tweeting a significant amount, but were objects of the media and as a result many individuals used Twitter as a way to attempt to interact with them.

<table>
<thead>
<tr>
<th>Handle</th>
<th>Out-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>peace_full</td>
<td>20727</td>
</tr>
<tr>
<td>denshasamurai</td>
<td>7946</td>
</tr>
<tr>
<td>salsonthejob</td>
<td>4677</td>
</tr>
<tr>
<td>railgirl1952</td>
<td>4672</td>
</tr>
<tr>
<td>ykhalim</td>
<td>3595</td>
</tr>
<tr>
<td>sasha2000</td>
<td>3151</td>
</tr>
<tr>
<td>theusawire</td>
<td>2924</td>
</tr>
<tr>
<td>theintlwire</td>
<td>2912</td>
</tr>
<tr>
<td>dialogkr</td>
<td>2862</td>
</tr>
<tr>
<td>fedupwithgovern</td>
<td>2754</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Handle</th>
<th>In-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>barackobama</td>
<td>1,054,197</td>
</tr>
<tr>
<td>mittromney</td>
<td>544,362</td>
</tr>
<tr>
<td>michelleobama</td>
<td>94,692</td>
</tr>
<tr>
<td>paulryanyvp</td>
<td>80,588</td>
</tr>
</tbody>
</table>
In the context of the election, those high in at-reply in-degree could be broken into two categories, campaign related accounts and provocateurs. The account associated with the Obama campaign (@obama2012) and the official White House account (@whitehouse) were also common recipients of at-reply’s illustrating the prominence of campaign related accounts beyond those of just the candidates. Media accounts also had high in-degree in the at-reply network with MSNBC commentator Ed Schultz (@edshow) having the tenth highest in-degree and media personality Donald Trump (@realdonaldtrump) being the sixth highest in at-reply in-degree.

Finding 7c: Citizens used the at-reply syntactical feature to direct messages towards campaign and media accounts, but this engagement was not reciprocated.

Campaign and media accounts comprise the majority of accounts that were the most targeted of the at-reply with over 42% of the at-reply messages being directed towards the most salient campaign and media accounts (Table 28). The top 87 accounts were coded to identify the types of accounts that were the recipient of messages. This statistic correlates to the high number of at-mentions of these accounts illustrating that citizens use Twitter to talk about and talk to public figures.

---

<table>
<thead>
<tr>
<th>Account</th>
<th>In-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>joebiden</td>
<td>55,850</td>
</tr>
<tr>
<td>realdonaldtrump</td>
<td>41,098</td>
</tr>
<tr>
<td>obama2012</td>
<td>33,081</td>
</tr>
<tr>
<td>anndromney</td>
<td>22,689</td>
</tr>
<tr>
<td>whitehouse</td>
<td>21,419</td>
</tr>
<tr>
<td>edshow</td>
<td>19,812</td>
</tr>
</tbody>
</table>

---

59 The top 87 individuals represent 43.8 percent of the total conversation pair data and 69 percent of the targets that were not included in the source info. This accounts for every account that was a target of an at-reply more than 1,000 times and yet did not respond.
Table 28: Types of Individuals Who Received At-replys

<table>
<thead>
<tr>
<th>Code</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Frequency</th>
<th>Percentage of These Codes</th>
<th>Percentage of Overall Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campaign</td>
<td>25</td>
<td>28.74%</td>
<td>1860158</td>
<td>92.09%</td>
<td>40.38%</td>
</tr>
<tr>
<td>Celebrity</td>
<td>12</td>
<td>13.79%</td>
<td>45904</td>
<td>2.27%</td>
<td>1.00%</td>
</tr>
<tr>
<td>Individual</td>
<td>1</td>
<td>1.15%</td>
<td>1293</td>
<td>0.06%</td>
<td>0.03%</td>
</tr>
<tr>
<td>Interest</td>
<td>2</td>
<td>2.30%</td>
<td>3751</td>
<td>0.19%</td>
<td>0.08%</td>
</tr>
<tr>
<td>Media</td>
<td>44</td>
<td>50.57%</td>
<td>102634</td>
<td>5.08%</td>
<td>2.23%</td>
</tr>
<tr>
<td>Parody</td>
<td>1</td>
<td>1.15%</td>
<td>2352</td>
<td>0.12%</td>
<td>0.05%</td>
</tr>
<tr>
<td>Political</td>
<td>1</td>
<td>1.15%</td>
<td>1666</td>
<td>0.08%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
<td>1.15%</td>
<td>2184</td>
<td>0.11%</td>
<td>0.05%</td>
</tr>
</tbody>
</table>

Over 75% of the coded accounts were campaign or media related and the next highest percentage of coded accounts were those related to celebrities. These celebrities were those involved in some form of campaign activity such as promoting one of the candidates on Twitter or headlining a campaign rally. The six accounts that were not associated with the campaign, media or a celebrity were a mix of individuals and parody accounts.

The number of at-mentions received by the campaign related accounts in this subset of data was 92%. In the context of the complete dataset, these coded campaign related accounts accounted for over 40% of the total number of tweets in the complete dataset. This illustrates the dominance of these accounts in the at-mention data. Media accounts were the second most frequent recipient of at-replys with just over 5% of the tweets in the dataset of the top 87 accounts and just over 2% in the complete dataset.

None of the campaign related accounts responded to an at-reply. This analysis was conducted through both analysis of the collected data and also by examining the timelines of the candidates to ensure nothing was missed in collection. This lack of response illustrates that the at-reply was used for a one-way conversation where individuals would direct tweets towards candidates or campaigns about policy issues or other activity and would also direct messages towards media personalities to voice their opinion on certain activity. The lack of response from candidates and a limited response from media personalities situates the at-reply as a way for candidates and media personalities to give the perception that they were interacting when they were only using it as a way to receive information.
Finding 8: The most active “conversations” were one-sided and directed towards the candidates and media.

Expanding the analysis beyond the individual account unit of analysis helps to highlight the frequently occurring “conversational pairs.” A conversational pair is the combination of a source and target handle. There were 490,563 conversation pairs that appeared multiple times meaning that 2,581,113 conversation pairs appeared only once. This accounts for 84% of all of the conversational pairs collected indicating a significant number of “conversations” only had one interaction.

There were 725 total pairs of conversation (of 3,071,676 total) that occurred more than 100 times (Table 29). The accounts that were the targets of these conversations (in-degree) were coded to identify the most frequently occurring activity at the conversational pair unit of analysis. This differs from the individual unit of analysis since these pairs represent discrete pairs of activity and illustrate repeated directed interaction from one handle to another.

Table 29: Top recipients in Conversational Pairs

<table>
<thead>
<tr>
<th>Handle</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>barackobama</td>
<td>361</td>
</tr>
<tr>
<td>mittromney</td>
<td>159</td>
</tr>
<tr>
<td>paulryanvp</td>
<td>13</td>
</tr>
<tr>
<td>whitehouse</td>
<td>11</td>
</tr>
<tr>
<td>cspanwj</td>
<td>7</td>
</tr>
<tr>
<td>joebiden</td>
<td>7</td>
</tr>
<tr>
<td>obama2012</td>
<td>6</td>
</tr>
<tr>
<td>realdonaldtrump</td>
<td>6</td>
</tr>
<tr>
<td>edshow</td>
<td>5</td>
</tr>
<tr>
<td>katyperry</td>
<td>5</td>
</tr>
</tbody>
</table>

The accounts reflected in Table 29 illustrates that the most frequent use of the at-reply is directed at individuals who played highly visible roles in the election. The candidates and press are the most visible figures in an election and as a result were the frequent recipients of a significant amount of activity. The top three recipients of the at-reply are the candidates and the fourth is the official White House Account. The
fifth most frequently occurring account is the C-Span Washington Journal account, which occurs highly as a result of promoting the Twitter handle on its daily television program. Following these accounts we see Vice President Biden, the Obama campaign followed by two celebrities (Donald Trump and Katy Perry) and progressive television personality Ed Schultz.

**Finding 9: Individuals that received an at-reply initiated more conversations indicating that political engagement may beget political engagement.**

There are differences in the number of at-replies that individuals used depending on whether they sent and received an at-reply. Of the 719,772 individuals that received an at-reply, 265,359 also initiated a conversation (36.7%). These 265,359 individuals represent a subset of those that were active on both sides of conversational activity (herein: “active conversationalists”). As identified in finding 7a, 1,547,474 individuals initiated a conversation in the complete dataset. This means that only 17.1% of individuals who started a conversation also received a message. Analysis of this subset of individuals on both sides of the conversation allows for a better understanding of both sides of engagement.

The median number of times that an active conversationalist sent a message was 2 and the median number of times that an active conversationalist received a message was 1 (Table 30). The data suggests that a majority of the individuals that are present in both sides of the conversational data are not actively engaging using the at-reply, since the median number of times they are the source and target is 2 and 1. The mean number of times that an active conversationalist used an at-reply is higher, suggesting that there are a number of highly active individuals on both sides of the conversation as identified in previous findings. This two to one ratio also illustrates that individuals in the dataset are more likely to send a message and it also further explains the earlier finding that there are twice as many individuals who sent a message as those that received one.

**Table 30: Summary Statistics for those that both sent and received an at-reply**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>8.57</td>
<td>2</td>
</tr>
<tr>
<td>Target</td>
<td>6.33</td>
<td>1</td>
</tr>
</tbody>
</table>

Donald Trump and Katy Perry are frequent recipients of the at-reply since these individuals played a role in the latter part of the campaign. Katy Perry wore a dress with “Forward” on it at one of President Obama’s final campaign rally’s, which resulted in a significant number of tweets about her. Donald Trump was vocally opposed to Obama and as a result was vocal on Twitter.
Separating those that either sent or received an at-reply and those that sent and received a message illustrates that individuals who were on both sides of a conversation were more likely to send a message. Table 31 illustrates that the mean and median number of messages for those who were only present on one side of the conversation. Those that only sent an at-reply did so approximately 1.82 times whereas those who only received a message did so on average 6.4 times.

<table>
<thead>
<tr>
<th>Source Only</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.82</td>
<td>1</td>
</tr>
<tr>
<td>Target Only</td>
<td>6.4</td>
<td>1</td>
</tr>
</tbody>
</table>

In comparison to the statistics for those who were present on both sides of the conversation, the number of initiated conversations was lower by using both the mean and median as a measure for those that were only present on one side of the conversation. This illustrates that a high number of individuals initiated a small number of conversations.

Although the causal direction is difficult to identify, given the data, those present on both sides of the conversation used the at-reply to initiate conversations more. Therefore, the data suggests that engagement begets engagement. Those that receive an at-reply are more likely to be sending them and as a result are on both sides of the conversation.

Analysis of at-mentions and conversational behavior illustrate how individuals highlighted information to others and engaged with others in an open technologically mediated forum. A similar examination of how individuals used the retweet mechanism illustrates how individual shared other Twitter information internally and the examination of what URLs were shared across all three types of datasets presents a unique analysis of the type of information that was being shared.

**Question 3: To what extent do the URL and retweet syntactical features in Twitter facilitate information exchange surrounding the 2012 Presidential Election?**

In the 2012 election, citizens primarily shared URLs that referenced user-generated content from websites such as Twitter, YouTube and Instagram. The types of URLs that were shared were similar across all types of tweets including retweets and at-replys. One of the reasons for this similarity was that retweets contributed to the presence of a significant percentage of the URLs in the complete dataset. This illustrates
that the URLs that occurred the most frequently during the 2012 Presidential Election were the result of information sharing using the retweet mechanism.

The individuals that retweeted others the most were not overtly associated with a campaign, candidate or political party. This signifies that those who were sharing information using the retweet mechanism were citizens and not campaigns or candidates. The most frequently retweeted accounts were candidates and accounts associated with the campaigns. This illustrates the wide propagation of candidate and campaign originated information by citizens during the election.

**Finding 10: Citizens primarily shared URLs associated with user-generated content and mass media outlets.**

During the 2012 Presidential Election, the two most popular base URLs that were shared in the complete dataset, retweet dataset and at-reply dataset were twitter.com and youtube.com (Table 32). A URL of twitter.com is representative of an individual sharing another user’s status or photo that had been previously shared in Twitter. The inclusion of youtube.com illustrates the sharing of a video hosted on YouTube. The prominence of twitter.com and youtube.com suggests that the “Twittersphere” is self-referential and even with the ability to share external information internal sharing is predominant. In addition to Twitter and YouTube, campaign websites are popular in all types of tweets.

<table>
<thead>
<tr>
<th>Table 32: Most Frequently Occurring URLs by Tweet Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complete</strong></td>
</tr>
<tr>
<td>twitter.com</td>
</tr>
<tr>
<td>youtube.com</td>
</tr>
<tr>
<td>instagram.com</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
</tr>
<tr>
<td>barackobama.com</td>
</tr>
<tr>
<td>t.co</td>
</tr>
<tr>
<td>mi.tt</td>
</tr>
<tr>
<td>trib.al</td>
</tr>
<tr>
<td>twitpic.com</td>
</tr>
<tr>
<td>washingtonpost.com</td>
</tr>
</tbody>
</table>
The most frequently occurring base URLs are similar among all three datasets with only minor differences. The mi.tt URL represents links to the mittromney.com website. This was an official URL shortener that was used during the campaign to share information from Mitt Romney’s official website. Barack Obama’s website is more popular than Mitt Romney’s website in all three of the datasets, the complete dataset, retweet dataset and at-reply dataset.

Campaign related information from President Obama was the fourth most retweeted base URL and campaign related information from Governor Romney was the seventh most retweeted URL. This suggests that individuals used the retweet mechanism to proliferate information that originated from official campaign websites. This finding correlates to Barack Obama’s popularity in other measures such as hashtags and at-mentions and also correlates to the finding that the candidates are the most highly retweeted individuals. This finding also relates to the finding 11 that identifies that retweets contribute to the presence of many URLs in the dataset. In the case of Barack Obama, approximately 75% of URLs that pointed to his campaign website in the dataset were a result of a retweet.

In order to better understand the type of URLs shared, the top 50 base URLs were coded as being associated with the Campaign, Mass Media or from User Generated content or NA for those that were unable to be included in one of those categories. Mass media websites dominated the type of URLs that were shared (Table 28)61. There is a significant number of mass media websites that were shared across all three of the datasets. This high number reflects the fact that the election is a prominent event for mass media outlet. As identified in Table 33, the most popular URLs that were shared were those for both newspapers such as the Washington Post and New York Times and online only news sources such as Huffington Post and Breitbart.com.

61 The sheer number of mass media websites that exist on the Internet helps to contribute to the significant number of them being present in the data. Compared to mass-media websites, the number of websites that host user-generated content is more limited.
Table 33: Top 50 Base URL Codes by Dataset

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>Retweet</th>
<th>At-reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campaign</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Mass Media</td>
<td>36</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>User Generated</td>
<td>7</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>NA</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

The most significant difference between the complete, retweet and at-reply datasets is that there are more user-generated websites being shared in the conversational dataset and less mass media websites being shared compared to the other two datasets. Though mass media websites made up a significant number of codes (based on raw number), URLs coded as user-generated make up a majority of the URLs that were shared. The percentage across the three datasets ranges from 54.35% in the complete dataset to 65.07% in the at-reply dataset (Table 34). Additionally, there were more URL shorteners that were ranked high in the conversational dataset (coded as NA). This illustrates that the type of information being shared in the conversational datasets was different than the other two datasets and that some of this information was no longer accessible.

Table 34: Percentage of Overall Tweets with URLs by Code

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>Retweet</th>
<th>At-reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campaign</td>
<td>7.01%</td>
<td>11.09%</td>
<td>5.53%</td>
</tr>
<tr>
<td>Mass Media</td>
<td>36.32%</td>
<td>31.48%</td>
<td>22.73%</td>
</tr>
<tr>
<td>User Generated</td>
<td>54.35%</td>
<td>56.25%</td>
<td>65.07%</td>
</tr>
<tr>
<td>NA</td>
<td>2.33%</td>
<td>1.19%</td>
<td>6.67%</td>
</tr>
</tbody>
</table>

Finding 11: The retweeting of URLs inflated their overall presence in the dataset as the most frequently occurring URLs in the corpus had a high presence in the retweet URL dataset demonstrating the importance of retweets as an information sharing mechanism.

The similar ranking of all of the popular URLs in the complete, retweet and at-reply datasets illustrates that the most prevalent base URLs occurred at a similar rate in the different types of tweets. The main difference is the percentage of each dataset that each of the URLs represent and how this differs among all three datasets. The top 20 complete URLs that were able to be decoded (excluding the shortened
Retweets and at-replys play a significant role in the sharing of the most frequently occurring URLs (Table 35). The most frequently occurring URL in the dataset, twitter.com had over 85% of its presence in the dataset. This was mostly as a result of its presence in retweets and at-replys. Retweeting a URL inflates the number of times it is present in the dataset as a retweet is not an original piece of information. Retweet behavior is a form of sharing original information with one’s followers whereas the use of an at-reply to share a URL is a directed information exchange that is highlighted to one individual, but seen by others. Therefore, using a URL in an at-reply or retweet constitutes a different type of behavior than using it independent of one of these syntactical features.

<table>
<thead>
<tr>
<th>URL</th>
<th>Retweet %</th>
<th>At-reply %</th>
<th>Total Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>twitter.com</td>
<td>80.74%</td>
<td>4.54%</td>
<td>85.28%</td>
</tr>
<tr>
<td>youtube.com</td>
<td>64.60%</td>
<td>7.86%</td>
<td>72.46%</td>
</tr>
<tr>
<td>instagram.com</td>
<td>43.46%</td>
<td>1.44%</td>
<td>44.90%</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>81.35%</td>
<td>3.12%</td>
<td>84.47%</td>
</tr>
<tr>
<td>barackobama.com</td>
<td>75.34%</td>
<td>2.14%</td>
<td>77.48%</td>
</tr>
<tr>
<td>twitpic.com</td>
<td>77.71%</td>
<td>4.44%</td>
<td>82.15%</td>
</tr>
<tr>
<td>washingtonpost.com</td>
<td>57.27%</td>
<td>3.59%</td>
<td>60.86%</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>47.95%</td>
<td>3.18%</td>
<td>51.12%</td>
</tr>
<tr>
<td>breitbart.com</td>
<td>65.76%</td>
<td>4.14%</td>
<td>69.91%</td>
</tr>
<tr>
<td>thinkprogress.org</td>
<td>75.09%</td>
<td>3.93%</td>
<td>79.01%</td>
</tr>
<tr>
<td>facebook.com</td>
<td>26.20%</td>
<td>4.86%</td>
<td>31.05%</td>
</tr>
<tr>
<td>politifact.com</td>
<td>82.68%</td>
<td>4.02%</td>
<td>86.70%</td>
</tr>
<tr>
<td>tumblr.com</td>
<td>13.53%</td>
<td>0.91%</td>
<td>14.44%</td>
</tr>
<tr>
<td>news.yahoo.com</td>
<td>27.53%</td>
<td>2.29%</td>
<td>29.82%</td>
</tr>
<tr>
<td>bbc.co.uk</td>
<td>52.16%</td>
<td>1.42%</td>
<td>53.58%</td>
</tr>
<tr>
<td>foxnews.com</td>
<td>65.64%</td>
<td>3.89%</td>
<td>69.52%</td>
</tr>
<tr>
<td>politico.com</td>
<td>70.73%</td>
<td>3.20%</td>
<td>73.93%</td>
</tr>
<tr>
<td>dailykos.com</td>
<td>72.83%</td>
<td>5.54%</td>
<td>78.37%</td>
</tr>
<tr>
<td>motherjones.com</td>
<td>79.42%</td>
<td>2.98%</td>
<td>82.40%</td>
</tr>
<tr>
<td>abcnews.go.com</td>
<td>67.38%</td>
<td>3.99%</td>
<td>71.36%</td>
</tr>
</tbody>
</table>
A large number of other popular URLs had only a slightly less reliance on these two syntactical features to account for their high numbers. The second most popular base URL, youtube.com, had just over 72% of its presence in the dataset as a result of retweets and at-replys. This illustrates that a significant percentage of the sharing of the two most popular URLs were a result of sharing information via the retweet mechanism and also through at-repillys.

The range of base URL percentages that exists in retweets varies from 13.53% for tumblr.com to 82.68% for politifact.com. The percentage range for the presence of URLs in the at-reply dataset ranges from .91% for tumblr.com to 7.86% for youtube.com. In total, the range for the combined percentages is from 14.44% to 86.70%. This is a wide range, but the mean is 64.94% and the median is 71.97%.

This skew towards the higher percentages indicates syntactical features such as retweets and at-repillys contribute to higher counts of certain URLs. This presence is most significant for retweets and is likely the result of the technologically defined ability to click on a tweet and retweet it. This is significant because the fact that the presence is dependent on retweets illustrates that the type of content being shared is not original and is the result of individuals retweeting others content.

Finding 12: Approximately 5% of the URLs no longer resolved to the original location of the information demonstrating the ephemeral nature of social media information.

Approximately 5% of the collected shortened URLs were unable to be decoded. The decoding of the shortened URLs occurred from January 25th – February 21st, 2013, approximately three months after the conclusion of the election. This illustrates that the information that was being shared in the network was no longer present at the original URL in a relatively short amount of time after it was shared. The most common URL shorteners and the number of URLs that were unable to be decoded from them (defined as being used greater than 5000 and being in the top 315 base URLs in the dataset) and the associated counts are identified in Table 36.
Table 36: Percentage of URLs unable to be Decoded By URL Shortener

<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>Retweet</th>
<th>Conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total URLs</td>
<td>17,356,265</td>
<td>10,448,022</td>
<td>712,402</td>
</tr>
<tr>
<td>t.co</td>
<td>490,365</td>
<td>480,256</td>
<td>15,924</td>
</tr>
<tr>
<td>trib.al</td>
<td>202,519</td>
<td>175,149</td>
<td>4,467</td>
</tr>
<tr>
<td>adf.ly</td>
<td>57,743</td>
<td>777</td>
<td>50</td>
</tr>
<tr>
<td>bit.ly</td>
<td>31,558</td>
<td>8146</td>
<td>821</td>
</tr>
<tr>
<td>ow.ly</td>
<td>28,060</td>
<td>14,691</td>
<td>1,496</td>
</tr>
<tr>
<td>q.gs</td>
<td>12,352</td>
<td>856</td>
<td>27</td>
</tr>
<tr>
<td>goo.gl</td>
<td>7818</td>
<td>3,950</td>
<td>3,273</td>
</tr>
<tr>
<td>is.gd</td>
<td>7146</td>
<td>549</td>
<td>31</td>
</tr>
<tr>
<td>Subtotal</td>
<td>837,561</td>
<td>684,374</td>
<td>26,089</td>
</tr>
<tr>
<td>Percentage</td>
<td>4.77%</td>
<td>6.55%</td>
<td>3.66%</td>
</tr>
</tbody>
</table>

The percentage of URLs that were unable to be decoded was higher in the retweet dataset than in the at-reply dataset, which suggests that retweeted URLs may be more ephemeral. The overall percentage of URLs that could not be resolved for the complete dataset was 4.77%. This number was higher for the retweet dataset (6.55%) and lower for the conversational dataset (3.66%). The higher prevalence of dead URLs in the retweet dataset may also indicate that URLs that are part of retweets were part of short-lived campaigns where information aged off in a quicker manner.

Finding 13a: Those with high out-degree in the retweet network (those who retweeted the most tweets) were those who do not appear to be overtly tied to a campaign or media organization.

In total, there were 28,922,377 retweets, amounting to 55.10% of the dataset. As described in chapter 3, a network analytic approach is used to examine the retweet data. The individual who is retweeting another individual is the head of the network pair and the person being retweeted is the tail of the pair. There were 6,888,620 unique individuals that retweeted someone and 1,742,635 unique individuals that were retweeted. There were 21,939,189 unique pairs of retweeters and those who were retweeted. This amounts to approximately 24% of total pairs occurring more than once and is indicative of a concentration of activity in a small number of individuals who were retweeted more than once.

The most prolific retweeters (high out-degree in the retweet network) are individuals who do not appear to be overtly tied to a campaign or media organization. Analysis of the top 414 accounts in retweet
out-degree (everyone who retweeted another tweet at least 1,500 times) does not identify any account that is overtly connected to a campaign, politician, or other mass media outlet. This suggests that the frequent retweeting was done by individuals that were unaffiliated with a campaign. Further, the lack of campaign accounts illustrates the limited use of the retweet functionality for large-scale information sharing on behalf of campaign related individuals and other mass media organizations.

One notable account that was a frequent retweeter was @GOPPrimary. This account was created with the explicit intent to retweet candidates seeking out the GOP nomination and to also retweet certain information that contained the hashtags #2012GOP and #GOP2012. The owner of the account is unknown, but based on profile information and the tweets it does not appear to be overtly tied to the party or a candidate.

**Finding 13b: Individuals who were retweeted the most (high retweet in-degree) were candidates or related to the campaign.**

In contrast to those that were prolific retweeters, those who were the most retweeted (high in-degree in the retweet network) were associated with a campaign, media outlet or celebrity (Table 37). The two most retweeted accounts in the dataset are associated with the two Presidential candidates. President Obama’s in-degree is eight times greater than that of Governor Romney’s. Retweets of @BarackObama were the top four most retweeted tweets in the dataset. The most retweeted tweet in the corpus was a photo of President Obama and the First Lady Michelle Obama hugging with the caption “Four More Years.” This tweet was tweeted by President Obama on election night after the election was called. Of the top 2,359 tweet texts that represent any text that appeared more than 1000 times, 98% of them were tweets that began with “RT @”

<table>
<thead>
<tr>
<th>Handle</th>
<th>In-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>barackobama</td>
<td>3,399,011</td>
</tr>
<tr>
<td>mittromney</td>
<td>417,893</td>
</tr>
<tr>
<td>michelleobama</td>
<td>362,942</td>
</tr>
<tr>
<td>obama2012</td>
<td>314,038</td>
</tr>
<tr>
<td>chrisrockoz</td>
<td>173,255</td>
</tr>
<tr>
<td>paulryanyvp</td>
<td>154,602</td>
</tr>
<tr>
<td>Joebiden</td>
<td>141,249</td>
</tr>
</tbody>
</table>
The other accounts that were highly retweeted were First Lady Michelle Obama followed by President Obama’s Campaign account, @obama2012. Following this is the parody account @chrisrockoz and then the two vice Presidential Candidates, @paulryanvp and @joebiden. Representative Ryan’s Congressional Campaign Account (@reppaulryan2012) was the 38th most retweeted account and was retweeted a total of 41,714 times.

There is a distinct cutoff of the top 4 accounts and the account that is the fifth most retweeted as the First Lady’s account is retweeted almost twice as much as @chrisrockoz. This prominence in the network of accounts associated with Obama illustrates a strong social media strategy on behalf of the Obama campaign. Obama’s campaign controlled three of the four most retweeted accounts. In contrast, Governor Romney’s wife’s account was only retweeted 19,937 times and was the 97th most retweeted account in total. This put her behind most major news outlets such as the Washington Post, New York Times and AP and major media personalities such as Bill Maher, Dick Morris and Donna Brazile.

One of the notable accounts that was created as part of one of the memes that emerged out of the campaign was that of Big Bird (@BigBird). This account was a parody account that was created following a moment in a debate where Mitt Romney discussed cutting funding to PBS and Sesame Street. The @BigBird account was the 24th most retweeted account. This illustrates the proliferation of memes throughout the Twitter network.

Finding 14: Information that originated on websites associated with new media entities was shared using a different syntax.

The presence of “via @[username]” at the end of a tweet is recognized as a form of a retweet and is often technologically facilitated by some applications and websites. Analysis of this subset of retweets indicates that the most frequently retweeted accounts using that mechanism are Internet media outlets and other content sharing websites. This suggests that the type of retweeting may be specific to the type of information being shared as this uncovers the presence of distinctly different types of accounts.
There were 36,500 unique Twitter handles that had information shared using the “via @” syntax. The top 10 accounts make up 37.2% of the total sharing using “via @” (Table 38). The most frequently occurring handle is @youtube, which is indicative of individuals sharing videos from YouTube using the Twitter button on the website. The other popular accounts that are high in retweet in-degree belong to new media outlets that first started online, notably the Huffington Post, Daily Kos and Breitbart News. This illustrates that new media websites such as these are the most prevalent in contributing information that is shared using this one specific type of retweet feature.

**Table 38: Via @ In-degree**

<table>
<thead>
<tr>
<th>Handle</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>youtube</td>
<td>161,147</td>
</tr>
<tr>
<td>huffpostpol</td>
<td>85,544</td>
</tr>
<tr>
<td>sharethis</td>
<td>46,537</td>
</tr>
<tr>
<td>moveon</td>
<td>31,864</td>
</tr>
<tr>
<td>dailykos</td>
<td>25,700</td>
</tr>
<tr>
<td>breitbartnews</td>
<td>20,608</td>
</tr>
<tr>
<td>huffingtonpost</td>
<td>19,703</td>
</tr>
<tr>
<td>politicususa</td>
<td>19,294</td>
</tr>
<tr>
<td>rollingstone</td>
<td>17,760</td>
</tr>
<tr>
<td>thinkprogress</td>
<td>17,228</td>
</tr>
</tbody>
</table>

The presence of certain types of accounts as being highly visible using one specific form of retweet syntax represents a technical difference being manifested through the interaction between Twitter and other websites. Information from these new media sites is shared differently in Twitter and is done in the context of a video or an article from the website. This differs from the other retweet data that is often the result of sharing information from a specific person.

**Looking forward**

The theoretically informed sample, longitudinal analysis and diverse analytic approach to this study uncovered a number of findings that contribute to the understanding of technologically mediated political discourse. The temporal examination of syntactical feature usage identifies the impact of electoral...
events on social media and how these syntactical features, specifically hashtags, facilitate political
discourse. The slicing of the data by syntactical feature allows for a more complete understanding of how
individuals are using Twitter to “talk about” and “talk to” candidates, although this activity is not
reciprocated from the candidates.

Further examination of information exchange illustrates that the URLs that are shared are
predominantly from user-generated and campaign related websites and that the presence of these URLs
relies on their inclusion in retweets and at-replies. Finally, it was identified that individuals are those
responsible for retweeting campaign related information, specifically the information from the Presidential
candidates. These findings have numerous implications for political science theory, technologically
mediated discourse, social media research and big data collection and analysis that will now be discussed.
CHAPTER 5: DISCUSSION

Twitter provides a technological platform that reduces geographic constraints and allows citizens to seek out and engage with others. It also allows candidates to engage with citizens directly. Compared to other social media technologies such as Facebook, Twitter allows unidirectional relationships to be created with relative ease and provides more native syntactical features (hashtags, at-mentions, at-replies, shortened URLs, and retweets) that facilitate different forms of communication.

The findings discussed in the chapter 4 illustrate that even with the availability of the numerous technological and syntactical features in Twitter that help to facilitate interactions and information sharing, there is still a limited realization of the promise that the technology affords. Instead of fundamentally changing political discourse by having individuals use it for two-way communication, Twitter amplifies the existing political environment where communication is one-way and discourse is disconnected.

During the studied time period, Twitter attempted to provide a centralized place for users to consume information by aggregating tweets during acute events that contained specific hashtags (#debates, #electionday, #gop2012, #dnc2012). The shear amount of activity related to these events (findings 3-5) resulted in tweets getting lost in the noise. Even with the technological ability to filter activity based on the numerous syntactical features such as hashtags, Twitter was unable to aid users in digesting the significant amount of activity that occurred within the environment. This set of findings suggests that more content (in this case tweets) in an open forum is not necessarily better. In order to better facilitate political discourse in a constructive and digestible manner, technological platforms may need to focus on better ways to enable users to filter activity.

The rest of this chapter enumerates the implications of the findings in the context of technologically mediated political discourse and social media research. First, the theoretical implications are discussed followed by a discussion of the implications for social media research. Next, the practical implications for campaigns, citizens and the technologies used for civic engagement are presented. Finally, the chapter concludes by discussing the next steps for research using these datasets for large scale Twitter research.
Theoretical Implications

Examining previous theories of political engagement in the context of this dataset helps bring to light how the 2012 Presidential Election on Twitter can be situated in decades old theories of political engagement. Previous research has found that social media use is a proxy for excitement about candidates and issues in elections (Williams & Gulati, 2007; Williams & Gulati, 2008; Williams & Gulati, 2009). The number and timing of tweets that were collected during the 2012 Presidential election suggests that Twitter was a proxy for excitement related to the 2012 election. It is difficult to compare the excitement for candidates in this election to previous elections since the environment (both technological and political) differs, but the fact that over 25 million Twitter accounts participated in the political discourse during the 2012 past election illustrates a substantial interest in online political discourse.

As Putnam (2001) argued in his seminal book, Bowling Alone, social capital in the United States has declined. His research points to a decline in voter turnout along with a decrease in civic participation as the reason for the decline in social capital. Putnam blames technology for some of this decline as he claims that individuals are more interested in television and other entertainment activities instead of civic activity. The number of participants and overall activity in this dataset reflects that individuals are interested in politics, but that this interest and engagement has shifted from traditional involvement to more contemporary mechanisms such as social media to comment on political events and share information. It is possible that individuals who are heavy social media users consider participation in hashtag discourse on Twitter or in a Facebook group as being part of a civic organization. Although excitement about politics in social media exists, there is still a significant lack of actual conversation among individuals. This argument is normative, but it is possible that the concept of conversation has changed along with the introduction of newer technology.

Democratic theory

Temporal analysis of the data identifies that the activity is concentrated in a subset of acute events (specifically the conventions, debates and the actual day of the election) (findings 1, 3 and 4). These findings echo Lippmann’s argument that the public is “phantom” but that certain events may lead to greater excitement and in turn lead to an increase in discourse (Lippmann, 1925). Lippmann was likely referring to an election as the “event,” but what we see in this dataset is that an election is a complex event that has
numerous related events that make up the larger context. The longitudinal approach used in this study allows for the analysis of a series of events that when combined contribute to the understanding of the greater context – the election.

The acute events occur on different days and address different topics (each debate had its own format and topic), but in the social media environment they are linked through the use of a set of shared discourse markers (hashtags and to some extent the candidate’s Twitter handles). Each debate is a discrete event as identified in the concentration of activity using debate specific hashtags on the days of the debates, but the persistence of the hashtag across time ties these events together into a series.

These markers allow individuals to be able to participate in the conversation between events. The longitudinal approach allows for further insight that could not be garnered from isolated analysis of these events. The different units of analysis that can be analyzed using data collected in this manner allow for the contrast of activity between debates in both activity and how individuals use different hashtags with varying frequency among the different events.

In Lippmann’s discussion he was more focused on day-to-day political activity outside of an election, but in the dataset we see that the electoral time period is a microcosm of political activity. Certain events drew significant participation followed by a lull in activity. The days where the activity was less than average still had significant activity compared to other topics on Twitter. This pattern of activity likely exists outside the context of an electoral time period in that certain issues may bring about significant discourse, but may be preceded or followed by a lull in activity that may not be indicative of the true discourse related to the topic.

One of the issues with studying political discourse outside the context of the election is that it might be difficult to collect a defined set of activity as there is so much going on since many of the issues may be emergent and not previously known. Depending on the snapshot of data collected, this can significantly alter the findings of any research or analysis. As evidenced in the findings to research question 1, the collection of only the #debates hashtag would have missed out on collecting a majority of the activity related to the Vice-Presidential Debate. Therefore, it is important to take a broad, yet theoretically informed approach to collecting this type of data.
There is a significant increase in activity around the debates and the actual Election Day that is similar to most studies on Twitter and current events (Hu, John, Seligmann, & Wang, 2012b; Shamma et al., 2010b; Shamma et al., 2010b). These increases in activity represent greater interest, as there is an event to comment on. This activity has also been found by similar research on the time period (Lin et al., 2013).

The days surrounding the debates accounted for 16.1% of the total activity in the dataset and Election Day accounted for approximately 15% of the total activity in all of the collected data (finding 4). This means that the other 80 days of the electoral period studied represented less than 70% of the total activity collected. Therefore, we see that individual contributions are concentrated around acute events. As discussed in the findings each of these events have some relationship to other events in the shared discourse markers (hashtags in this case). This further illustrates Lippmann’s observations about political engagement from over 80 years ago that news events draw higher interest in political discourse providing further empirical evidence that the public may be “phantom” and may need events to have an interest in politics.

In Dewey’s counterargument to Lippmann’s “Phantom Public” argument he claims that the public lacks cohesion and is distracted by the societal improvements of entertainment that were occurring in the time period preceding the Great Depression (Dewey, 1927). Dewey argued that this distraction could be fixed with technology. These technological improvements would facilitate the connection of multiple disconnected publics and would lead to cohesion between previously disconnected groups and as a result would create a renewed interest in politics.

Twitter is one such technology that has the ability to help address this lack of cohesion and interest in politics as it affords individuals the ability to find others and engage with them using hashtags. The significant amount of activity noted during the studied time period (as identified in findings 1-3) illustrates an interest in politics, but the limited amount of repeated interactions that were identified in the at-reply data (as identified in finding 7) illustrate that there are still a number of disconnected groups and individuals. The data reflect that Twitter is being used to contribute to the larger discourse, but that this discursive behavior is not occurring between individuals, but instead is primarily individuals voicing their opinions or ideas to a public forum.

One of the greatest gaps in political social media research is identifying whether online activity translates into votes on Election Day. It is unclear if the activity reflected in the data led to greater
engagement and votes in the election, because there is no way to tie the collected social media data with offline behavior. One thing that is reflected in the findings (findings 3-5) is that Twitter users participated at a higher rate during acute events than at other times during the election. Although activity was greater during acute events, a majority of users (51.34%) only participated once, which represents a limited amount of repeated activity.

Therefore, even with the introduction of technology to bring together large groups of individuals in an open forum, the public is still phantom in many ways, as they only use the technology to comment in a large public forum. Instead of actually allowing the public to engage, the introduction of an open technology such as Twitter may be an even greater distraction as there is so much going on and the technology is still limited in the way that individuals can filter the activity. This limited filtering along with the ability of anyone to introduce a hashtag may harm cohesion, as the number of hashtags is relatively unlimited. The asymmetry of Twitter’s friend relationships also limits this cohesion because of the ease of exposure to numerous individuals and the context collapse that occurs with what individuals are saying (Marwick & boyd, 2011). The ease of technology ends up being its greatest downfall in facilitating activity beyond commenting.

**The Internet’s effect on political processes**

There is evidence to suggest that that the Internet is continuing to facilitate the amplification model of politics (Agre, 2002). Agre’s amplification model states that the Internet amplifies existing political processes as opposed to the Internet addressing a fundamental shortcoming in the process and leading to change (reinforcement model). Agre’s survey of the literature identifies that most authors identify the presence of the reinforcement model in situations where broader participation is facilitated. In contrast, the amplification model echoes that the Internet will bring no change and yet will amplify processes that already exist.

This study provides evidence that suggests that the 2012 Presidential Election on Twitter fits the amplification model. We see that the broader participation by individuals does not seem to be organized even with syntactical features that make the identification of other interested individuals easier. This is most evident in the findings related to research question two (findings 6-9) as they illustrate that although there is a lot of activity, most of this activity is not cohesive around a specific activity. In the case of the
analysis of the at-reply (findings 7-8), individuals used it to talk to and about candidates and the media as opposed to engaging with each other.

Although, the Internet overcomes limitations of the physical world by allowing millions of citizens to “talk” to each other, with this facilitation comes the difficulty of people being able to manage the multitude of interaction possibilities with the others that are co-present in the technology. Finding 8 illustrates that most of the use of the at-reply was one-sided and that at-replys were mostly directed towards candidates and the press. Finding 9 indicates that individuals who received at-replys were more likely to initiate a conversation. These findings suggest that engagement from others brought on engagement from a specific person. In this case, technology is helpful by breaking down barriers between individuals allowing for more engagement, but these barriers may have existed not from a lack technology, but a fundamental lack of individuals being able to manage interactions with more than a defined number of people (Dunbar, 1993).

As Davis (Davis, 1998) argued more than 15 years ago, the Internet is “destined to become dominated by the same actors in American politics who currently utilize other mediums.” In this study, those who were most discussed in the data were the candidates and media as they were the subject of the at-replys, at-mentions and the information exchange that occurred (findings 6-7). This is a similar finding that would exist in traditional media sources as newspapers and television news channels talk about what is newsworthy and those topics are the candidates and the issues. This provides further evidence that the amplification model is characteristic of this activity.

News outlets would talk about and carry information put out by the candidates and we see similar activity in Twitter. Citizens used the retweet function to share information from the candidates with others in the network. The actual deliberative possibility that could exist with this technology is not fully realized. Instead, the technology is used as a feedback mechanism for popular figures and in essence amplifies existing processes as Agre identified.

**Twitter as a new public sphere**

There is some evidence to suggest that Twitter represents a public sphere in the context of the 2012 election based upon the criteria set forth by Dahlberg (Dahlberg, 2001) that built upon Habermas (Habermas et al., 1974; Habermas, 1984) and that has been discussed in the context of Facebook
The criteria identified by Dahlberg (2001) as being required for a technological public sphere include:

- “Exchange and critique of reasoned moral-practical validity claims”
- Reflexivity
- Ideal role taking
- Sincerity
- Discursive Inclusion and equality
- Autonomy from state and economic power

In this study, there was no state or economic power as traditionally conceptualized, but the prescription of certain hashtags (#debates) by Twitter for discourse may indicate the presence of a technological power that may be likened to an economic power (finding 4). The fact that individuals used the prescribed hashtag for acute events along with other hashtags that were not prescribed suggests a level of autonomy from economic (technology) and state (campaign) power. The ability to create hashtags also suggests some form of autonomy from these powers, but as previously mentioned, this autonomy may have led to a lack of cohesion and set of disconnected activity.

In the case of promoted hashtags (finding 5), it was in the interest of Twitter as a technology platform and business to help the campaigns realize as much gain as possible from the promotion of these hashtags. The autonomy required as part of the public sphere may have been limited since individuals may have been influenced to use the hashtag by the advertising of it on the first page of Twitter and the fact that Twitter created some of the official hashtags for the events. This behavior is still different from what Habermas (1984) and Dahlberg (2001) implied by an economic power, but the conceptual evolution in the concept of the technology has significant implications. It is now possible for the technological platform to control how activity is carried out with the promotion of certain syntactical features and also the access to certain campaigns to promote their own. It is also widely documented that Twitter filters some tweets, but the specifics of this filtering are unknown.

---

62 Twitter claims in their API documentation that they use filtering to facilitate a better experience by eliminating spam and other objectionable material. This filtering is not well-documented and it is possible that “one man’s spam is another man’s gold.”
The extensive use of URLs in the dataset (findings 10 and 11) illustrates that individuals were using some form of reasoning when making their comments. This would be evidence of the “exchange and critique of reasoned moral-practical validity claims”. As evidenced by the type of URLs shared, many of them linked to user generated content and many were to other tweets (URLs of twitter.com). This is still a form of reasoning as set forth by Dahlberg (2001) and further described by Robertson (2010). The lack of content analysis on such a large number of tweets limits the ability to identify reflexivity, but the inclusion of hashtags and other syntactical features such as issue specific ones (#sandy, #benghazi, #ohio) that were identified as co-occurring with event specific hashtags may be evidence to suggest that individuals are being reflexive in their activity (finding 3).

It is further noted that sincerity is evidenced by the openness and freedom that individuals had to participate. Individuals were not forced to participate in the political discourse in Twitter and there is no evidence to suggest that there was disingenuous activity en masse. The ability to post asynchronously and without any limitation on contribution number suggests that there is some evidence of equality and discursive inclusion along with role taking. Although there was limited conversation that was identified, just under half of the individuals participated more than once. The inclusion of a syntactical feature such as a hashtag in a tweet is evidence that the participants wanted to be heard and Twitter afforded the ability for their ideas to be found by surfacing their tweets if a hashtag was searched for and to some extent by aggregating them on one page for certain hashtags such as #debates.

Moving to the private sphere

Although there is some evidence to suggest that Twitter represented a public sphere, there is further evidence to suggest that it enabled a “private sphere” as described by Papacharissi (Papacharissi, 2004; Papacharissi, 2010). The private sphere is one where individuals are “alone, but not lonely or isolated (117).” There is some evidence for what Papacharissi identifies as five new civic habits that individuals have developed to engage with others in the new technological landscape.

In Twitter, individuals use hashtags to participate in a larger discourse, but are also aware that they are sharing their tweets with their followers. This is evidence of context collapse (Marwick & boyd, 2011), but these “multiplied” social audiences allow for individuals to present themselves in the context of a larger conversation. This illustrates that individuals are focused both on a mix of private and public sphere
activity. Although individuals used hashtags to join a public conversation and engage in a public sphere of activity, the sharing of URLs focused on other individual’s tweets, which is likely to include one’s own tweets and the tweet of their friends, is evidence of a private or limited public sphere.

The fact that there was a significant number of tweets that were not at-replys and were not a retweet suggest evidence of “narcissism” as described by Papacharissi (finding 1). Those who tweeted used Twitter to have their voice be heard. A user’s timeline is essentially a blog that is written 140 characters at a time. The lack of engagement and limited information sharing, but significant amount of hashtag and at-mention use illustrates that individuals wanted to use the technologically prescribed syntactical features to be heard, but not to contribute in a deliberative manner.

Although the notions of the public sphere and private sphere may seem at odds with each other, they are somewhat complementary in that individuals who create a private sphere in their online worlds may use it to extend their actions from the private sphere of the virtual world to the public sphere of the physical world. Without a mixed methods approach of studying online and offline behavior this type of analysis is not possible, but it is something that could be done on a much smaller scale and would be most effective if done in the context of a localized election so that variables such as campaign issues, user population and technology used could be controlled for.

**Deliberative discourse not realized**

The lack of at-replys in the corpus illustrates that Freelon’s notion of deliberative democratic discourse (Freelon, 2010b) is not realized in this dataset. There is some evidence of deliberative behavior since there is a public issue focus, equality, and discussion topic focus, but there is limited repeated interaction among individuals (finding 7). As a result of this limited interaction there is a lack of questioning and reciprocity, which are aspects of deliberative behavior.

There is no evidence that a significant number of conversations occurred in the public between individuals as most of the at-reply activity was directed towards candidates and media (finding 7). This does not mean that conversations did not occur online; it only suggests that Twitter was not the chosen venue for these interactions. The most common “conversation pairs” were those that were one individual directing at-replys towards a candidate (finding 8). Although the technology was there to facilitate this activity, and individuals had the ability to engage with each other, there was limited adoption. This would
suggest that the most apparent model of democratic communication that is represented within this dataset is somewhere between liberal individualist and communitarian with some limited attributes of the deliberative nature.

Twitter is not facilitating traditional one-on-one conversation en masse. Individuals may feel as though they are conversing in a large group, but not to one person. This is a combination of the public and private sphere. Individuals use the technology to reach out to public figures that they have no other access to. In the case of media personalities, many encourage this reaching out, but yet they do not respond. This lack of conversation is something that individuals may not mind, but it could also lead to a feeling of not being heard and may have negative implications in future election cycles. It is possible that technological discourse will evolve in future elections to become more deliberative, but the ideal of deliberative communication may not be achieved given the large scale of Twitter even with the existence of syntactical and technological features to facilitate these types of communication.

**Socio-technical Implications**

There are numerous socio-technical implications for social media research that are highlighted by the findings. These implications relate to the presence of syntactical features and how they were used to talk about candidates and how information was exchanged. This dataset also provides a foundation for understanding the presence of certain syntactical features such as retweets in comparison to other research that has been done during previous political cycles.

**Syntactical Feature Usage**

The dataset reflects a significant percentage of individuals who use syntactical features and more specifically the official Twitter handles of the candidates to highlight them in the discourse. Campaigns and the candidates use Twitter to reach out to the public as another broadcast mechanism in combination with their websites (Aragón, Kappler, & Kaltenbrunner, 2013). The differing uses by the two parties illustrates that the technology has not yet been fully utilized for the engagement opportunities that it affords.

Although candidate mentions comprised a large amount of the actual discourse, it was still limited in total adoption with the maximum percentage on any given day accounting for only 30% of the total tweets. Individuals used Barack Obama’s Twitter handle to talk about him, whereas they used Mitt Romney’s name (not Twitter handle) to talk about him. This is identified by the higher percentage of
Barack Obama’s handle as opposed to Mitt Romney’s and the higher percentage of Mitt Romney’s name without his handle. Using only the name without the technological feature makes the tracking and identification of the discourse more difficult since it requires the tracking of the keywords instead of just the tracking of the mentions of the handle.

This finding differs from previous political research in which most of the discourse was overtly candidate or party centric (Bruns & Burgess, 2011; Jurgens et al., 2011). The differences in findings between this study and previous work may be a result of the dataset construction or the fact that the election was more than just about the candidates and also about the issues. Further, the ability to use promoted hashtags and other affiliation hashtags such as #romney or #obama may have allowed individuals to indirectly discuss the candidate or the campaign without using the candidate handle. These other syntactical features allowed for an affiliation for each tweet, but not an overt calling out of the candidate. The difference in usage of the at-mention versus the hashtag is an area for future exploration.

The overall syntactical feature distribution of this dataset is similar in some ways to other political Twitter datasets and also differs in some regard. The percentage of retweets (55.10%) is similar to previous political datasets of United States politics (Mustafaraj & Metaxas, 2011), but differs from similar analysis on European elections (Tumasjan et al., 2010). This could be a result of the different environment and dataset construction, but also could speak to a difference in user population (Americans versus Europeans). This finding suggests that culture may have an effect on the activity in the technology.

The percentage of at-replys was higher than other datasets representing a higher adoption of the syntactical feature, albeit for something not necessarily conversational in nature as evidenced by the lack of back and forth activity (Honeycutt & Herring, 2009). As most research on Twitter uses the hashtag for creation of datasets, the percentage of tweets with a hashtag is not comparable (since a dataset with a hashtag as the selection criteria means 100% of tweets will have a hashtag). Therefore, these high level statistics are presented in chapter 3 as a way to establish a basis for further research into mixed syntactical feature datasets.

**Information Exchange**

There were a large number of shortened URLs that no longer resolved to the original URL that was shared. This accounted for approximately 5% of the total URLs that were shared in the dataset. This
small percentage is likely to be immaterial to the overall findings given the large amount of data analyzed, but the decoding of the URLs was done soon after the event. It is highly likely that the number of shortened URLs that no longer resolve increases daily as has been identified in other research (Salah Eldeen & Nelson, 2012). Therefore, it is important to resolve shortened URLs as quickly as possible in doing this type of research to attempt to gather as much of the original information as possible.

There are a significant number of retweets in the dataset and the URLs that are included are different than in other types of tweets in the dataset. Research on what a retweet means has been limited thus far, but the use of the URLs and types of individuals that are retweeted (campaign related) may suggest that individuals are using retweets and URLs to propagate information. It is also likely that individuals are using retweets to show an affiliation to a certain candidate or ideology as opposed to using it for conversational purposes.

**Practical Implications**

This study highlights a number of practical implications for campaigns and candidates and also for the citizenry. The study also highlights a number of other implications for understanding the influence of technological platforms in the facilitation of civic and political discourse. These implications are based on this study and the technological environment of the 2012 election with Twitter being a focal point of activity. With the increasing reliance on technology to facilitate technological discourse, the overall implications that are identified in this study will have some application to future elections, but as identified in the literature review the technological location of this discourse may change. As a result of this change in technology used in future elections, the syntactical and technological features that are available to the users and candidates may afford different types of interactions.

**Campaigns**

Microtargeting, the use of large datasets to target specific groups of individuals, has been a strategy of campaigns for decades, but only recently with the advent of the Internet have campaigns been able to do it on a large scale (Sosnik, Dowd, & Fournier, 2006). The large number of individuals that are using public channels to communicate, in the context of this study Twitter, means that the efforts to understand what this communication means and how actions can be taken to influence this communication is increasing. The Obama campaign had a team of dozens working for 18 months using data mining to
identify undecided voters (Rutenberg, 2012). The data used for this identification came from traditional microtargeting datasets, but as data analytics and communication technologies evolve, it is likely that these techniques will evolve and use other data feeds to further augment the existing targeting efforts.

Campaigns could use social media data to identify individuals who may be on the fence about a specific issue related to a candidate and marry that up with other demographic and behavioral data to craft a specific message for them. It is possible that this specially crafted message could be delivered through social media such as microtargeted, promoted hashtags that would only be delivered to a subset of individuals for them to use and propagate throughout the network. Using the technology, the campaign would then be able to track the response to the message and see if an individual changes their commenting in public and the extent of the proliferation of the message that was promoted to the individual. As identified in the findings (finding 5), promoted hashtags have a differing effect on the different user bases, but there is still a significant research gap in understanding why specific promoted hashtags were used more than others and how this use varied among individuals in the dataset.

This type of analysis would require the use of natural language processing coupled with targeted social network analysis to identify those who were active in certain aspects of the activity. It is possible that this identification could also be used to recruit individuals who are influential in the network and have them help promote a message or a hashtag organically. This “organic” hashtag promotion is something that has been identified in the past as “astro-turfing,” as a result of its perceived grassroots nature, but this type of activity does not have to have the negative connotation since this type of activity already occurs in physical campaigns (Ratkiewicz et al., 2011). The only difference is that in physical campaigns it is often much harder to identify those who are most likely to be influential. The barrier that may exist to identifying astro-turfing in traditional campaigns is overcome with the collection and analysis of activity in Twitter as all of the activity is recorded.

There was no evidence in the dataset or in the collected timelines that any of the official accounts responded to anyone publicly. The overwhelming evidence of individuals using the at-reply to reach out to campaigns suggests that the campaigns should also publicly reply to some of these individuals to give the perception that they are listening to what is being said to them. This connection may help build confidence or support by making the campaign seem in touch with the people. The ability to use platforms such as
Tweetdeck make this activity obtainable by allowing multiple campaign staffers login to one account and respond on behalf of the campaign. This type of behavior is similar to how administration offices currently respond to traditional mail and engagement from citizens.

Citizenry

Citizens now have the ability for the candidates to deliver information directly to their web browsers or mobile phones. This type of information acquisition allows for individuals to do more research on candidates and share information with other citizens that may be interested. In this study, the public used the syntactical features to reach out to the candidates and the media with no response (finding 7). The concentration of at-replys targeted towards public figures suggests that individuals wanted to reach out to these figures to engage in a dialogue or to feel as though they are being heard.

In the case of the 2012 election, the technology for interaction exists between the citizenry and public figures, but it has not been fully utilized. Candidates use it to disseminate information to the citizens and when a response is garnered they do not continue the interaction. Therefore, the technology is still only being used for one-way communication. Given the scale of the electorate, a continued interaction between all of the individuals who reached out directly to the candidates may have been difficult given the scale, but it would have been possible to attempt to carry on some interaction.

In addition to the limited interaction between candidates and citizens, there was also a lack of repeated interaction among the citizens (finding 8). The limited amount of deliberation among citizens may be the result of a lack of unifying force in the technology. Although hashtags exist to identify others to engage with they may not facilitate the discovery of other individuals easily. If an individual contributed to the discourse using a popular hashtag (such as #debates), their message was likely lost in the noise given the significant amount of activity that surrounded these acute events (findings 3 and 4). It is possible that being lost in the noise and not garnering any response may have lead to engagement attrition where engagement that was hoped for was never realized.

In the case of Twitter, the technology facilitated an ideological and issue dimension that traversed geographic boundaries (Agre, 2004). This allowed individuals to reach out to others to engage, but even though the technology existed, it was not adopted in such a way as identified by the lack of conversational activity. The ability to coalesce around issues as identified by Agre (2004) was limited since the most
active times had tens of thousands of tweets per minute with limited ability for users to filter the data from the significant amounts of noise.

This is something that could be addressed by providing better filtering for the user or providing a moderated hashtag where individuals would be able to construct a place to filter out what they identified as noise. This can be done using some third party Twitter applications, but the most commonly used applications in the dataset were Twitter for iOS, Blackberry, Android and the Twitter website. These applications lack extensive filtering beyond the basic functionality of the technology. The addition of collaborative filtering within more commonly used applications could allow citizens to better weed out noise or messages they may not want to see. Although this runs counter to some notions of the public sphere and deliberative discourse, it might help to facilitate further engagement by allowing the users to construct their own contextual environment.

Further, it is possible that similar to how physical political activity requires a set of civic skills (Agre, 2004; Papacharissi, 2004; Papacharissi, 2010) that political engagement and activity on the Internet may require a specific set of civic skills. The combination of technology skills coupled with political prowess helped President Obama build a grassroots movement in 2008 to win as a relatively unknown first term Senator. These types of skills need to be promoted among the citizenry. The belief that citizens may be able to translate physical civic skills into virtual civic skills is lacking since the environment is different and is not geographically bound. It is possible that with the aging of the millennial generation that these skills will continue evolve and political engagement on the Internet will continue to get stronger.

**The influence of the technological platform**

Twitter (as a technology) has a significant influence in the usage of social media for the election. The official hashtag of “#debates” was not one that was decided by the public or by the campaigns, it was specifically chosen by Twitter (the corporate entity). The tweets that used this hashtag were aggregated on a specially crafted page so that individuals could easily see all of the tweets with that hashtag aggregated in one place during the debates.

The adoption of the hashtag by individuals along with the disparate pockets of individuals that used other hashtags for the debates illustrates how the dictated hashtag is not the one used the most (finding 63). On election day alone, these four applications accounted for over 75% of all of the collected activity.
4). Further, the fact that #vpdebate emerged as more popular than #debates during the Vice-Presidential debate and was not promoted anywhere illustrates how the culture of Twitter has evolved into one where users coalesce around a hashtag without any specific coordination. Additionally, we see that hashtags promoted by the event organizers in the case of the first debate at Denver University were relatively unused by participants. This is in contrast to some of the hashtags used in primary debates that do gain traction relatively easily (Black et al., 2012), but demonstrates the impact that a technological platform can have on the discourse for an event.

The ability for a campaign to promote a hashtag puts new media technologies in a similar place as traditional technologies such as television and newspapers. Just as a television advertisement may lead to discussion or help to persuade someone, a tweet and associated promoted hashtag could spark discourse around a specific issue. The adoption of the promoted hashtags suggests that a conversation was sparked by the use of these. In the case of the Obama campaign, this conversation persisted beyond the day where the hashtag was promoted throughout other parts of the electoral time period and led to the promotion of similar hashtags weeks later.

This type of activity demonstrates the significant role that the technological features and the technological platform (Twitter specifically, but corporate interests more broadly) play in promoting public discourse. There was no evidence that any of the third party candidates promoted a hashtag. It is unknown as to whether this was financial decision by the campaign or that Twitter chose to not allow them to. The influence on the decision making process of voters is unknown for technologies such as Twitter, but as this technology continues to gain a greater place in elections it is possible that some regulations may be created similar to those that exist in television that afford candidates equal opportunity to “air time” based on certain criteria and standing in the polls. The ability for any organization to purchase a promoted hashtag may also lead to greater influence of the role of money in politics.

**Methodological Implications for Social Media Research**

The study highlights some methodological implications that have not been identified in other research that examines Twitter and social media. One of the implications of this research is the manner in which the construction of a theoretically informed dataset that mixes different types of collection terms allows for a richer analysis. In the context of this study, the collection of keywords, hashtags and at-
mentions allows for the temporal analysis of syntactical feature adoption and how this varies over time. These differences highlight interesting characteristics of how individuals construct discourse in a technologically mediated environment. The difference in feature utilization identifies interesting behavioral characteristics that are tied to acute events. Further the unit of analysis (the day) choice and the ability to slice the dataset multiple ways based on clearly identified syntactical features uncovers unique findings.

**Dataset Construction**

The aggregation of the dataset that was constantly being refined and expanded as new events emerged allowed for a deeper analysis of the 2012 election while it unfolded. In this study, 68 different terms were used to construct the dataset and these were from different categories (candidates, events, promoted hashtags) to attempt to get as near a complete dataset of political discourse as possible that occurred during the studied timeframe.

The elimination of terms that were explicitly collected from analysis helps to identify the most common hashtags and at-mentions (findings 2 and 6) that emerged organically within the discourse beyond just the collected terms. This co-occurrence analysis could be done in real-time and fed into a collection system to collect syntactical features that were not tasked for collection originally, but that were related based on their co-occurrence.

The collection of more than just one hashtag or a small subset of hashtags as has been done for previous studies on political discourse on Twitter provides a mechanism to collect tweets where individuals did not use a hashtag or a candidate’s handle. This broader collection approach that was detailed in chapter 3 provides a more comprehensive dataset representative of political discourse during the 2012 Presidential Election.

The ability to slice this information by the different collection terms and also by other units of analysis such as day and event helps to further identify unique characteristics that emerge in an event that has considerable news visibility as identified in chapter 4. Further, the limited number of tweets that were concentrated in one hashtag or handle such as was the case in finding 14, where less than 15% of the tweets on debate days were tagged with the officially sanctioned hashtag illustrates the need to develop theoretically informed datasets that incorporate more than just one hashtag or at-mention. These datasets
would help to ensure that a collected dataset was more representative of an event and not just a small snapshot.

**Unit of Analysis Choice**

Without the temporal component that the longitudinal perspective provided, it would have been impossible to identify how syntactical feature usage varied by the events. Most social media research uses a specific word (hashtag) as the primary selection criteria and unit of analysis. The temporal component is often a secondary unit of analysis that is often muddled in the analysis. In this study, the dataset was constructed to be as inclusive as possible with a focus on time as the unit of analysis as electoral events occur in a temporal environment.

Using a temporal unit of analysis of a day allows for the understanding of an event in the context of a news cycle. The use of UTC time, although a limitation in some regards may have helped the analysis by diffusing the time of the event activity across multiple days since many of the electoral events spanned multiple time zones. This may have allowed for a more complete analysis of the activity, but this is an area of future study given the unknown demographics and location of Twitter users.

If one were to study only one of the debates during the 2012 election cycle it is possible that the findings such as how the official hashtag was adopted would have been misleading given the difference in activity among all four debates. Using a collection apparatus to collect all of the data surrounding the larger event (the election), it is possible to see the difference in hashtag adoption across all events of a similar type and how the prominence of certain hashtags evolved and how certain hashtags such as #vpdebate, while intuitive, emerged organically. By only collecting on the #debates hashtag, it is possible that it would not have been identified that #vpdebate played such an important role in facilitating discourse. Further, the inclusion of selection terms other than hashtags allows for the study of this activity across multiple syntactical feature units of analysis.

The presence of issue specific hashtags, identified through co-occurrence analysis, illustrates that elections do not happen in a vacuum where the discourse is identified by the election specific hashtags. Therefore, collecting only a small subset of hashtags for larger events may lead to the overlooking of important activity that might provide material insight into the activity that was happening in the technological platforms.
Conceptualizing the Approach Clearly

It is imperative when studying technologies to understand the context and atmosphere along with how individuals use the technology and what the findings may mean from a technological and social perspective. Understanding the technology and the syntactical features along with clearly operationalizing them is a necessary aspect of any research in social media. Further, there are a number of ways that have organically emerged to use syntactical features alternatively than the ones that are commonly accepted. For example, there are at least seven ways to retweet something (Kooti et al., 2012) and at least one other way to use an at-reply (with a period preceding the @ to broadcast it beyond just the two people and their followers, if the privacy setting is chosen). In this study, separating one of the retweet methods (via @) identifies a collection of usernames that are fundamentally different than the usernames retweeted using the other methods. Although the overall percentage of these tweets was small, important findings were derived once the dataset was extracted. This suggests differences in information behavior for the retweet mechanism chosen.

In the URL analysis, we see that URLs appear at different rates depending on if they are part of at-replies or retweets or normal tweets. The presence of these URLs in the other datasets may represent a different purpose. The retweeting of a tweet with a URL is likely be a show of support or information propagation and this could be one of the reasons that mass media URLs showed up less in the retweet dataset compared to the overall dataset. Further, it is imperative to be able to identify the differences in these types of messages and construct different datasets for study to identify differences.

Some research has been done to identify what users intend by the use of specific syntactical features, but this is still in its infancy. There has been some work for retweets as identified in the literature review, but there is a lack of understanding of user intent for the inclusion of URLs, at-replies, at-mentions and other syntactical features. This is an area of future research that will require a combination of online and offline research to understand user intent.

Conclusion and Next Steps

The combination of datasets used for this study has allowed for an examination of how technology specific syntactical feature usage varies by day and event and how these syntactical features can be used by individuals to engage with each other, with candidates and share information. The 14 findings from this
study contribute to a number of existing literatures from the theoretical, practical, substantive and methodological perspectives. The implications for these findings are broad and extend beyond just politics as more communication is occurring online.

Although these findings stand on their own as contributions to knowledge, there are a number of other areas of research that this dataset would be valuable for. The application of natural language processing to these datasets may allow for the identification of valence in the tweets to further the understanding of how the syntactical features were used. This methodological approach would also allow for a deeper understanding of the responses to candidate activity given an event such as a debate or speech.

The unique dataset construction allows for more granular approaches to examining the data. In the case of the acute events that were studied, it is possible to examine activity using a variety of units of analysis, such as time or the syntactical features used in the context of an event. In this study, the debates are examined in the context of the whole dataset, but further examination of how a message was retweeted throughout a debate or speech given by the candidates would help to uncover how candidate information diffuses through a network in real time. This type of analysis is possible given the broad collection terms that were used to create the complete dataset. This type of method could also be applied more broadly to the time period in general and understand how individuals respond and share candidate-specific information over the course of an election.

Further, the construction of a “user \(\rightarrow\) syntactical feature” dataset would allow for an understanding of how individuals use different hashtags, at-mentions and keywords. Analysis of those who tweeted more than a given number of times would allow for an understanding of how they used their syntactical feature of choice and it might be possible to identify the presence of issue entrepreneurship given their dependence or focus on one specific set of hashtags. It is also possible to combine this type of network analysis with positional analysis (where a syntactical feature exists in a tweet) to see how the syntactical feature usage varied over time. The examination of user device information that was collected and specific geographic information would likely yield interesting insights to how different technology was used to participate and where these individuals actually were located.

Beyond the scope of this dataset, it is possible to apply this methodological approach and findings to a smaller dataset for a regional or local election and design a study to gather attitudes of voters that use
social media to compare their offline characteristics with their online activity. This type of study design would allow for a more thorough understanding of the intent of the online actions and would rival seminal studies on voter behavior and attitudes in complexity and impact (Huckfeldt & Sprague, 1987). This type of study would also allow for greater insight into how social media is used in the context of political activity and would address one of the most significant shortcomings of all social media research – the lack of offline corroboration of activity.

There are a number of possibilities for research into the datasets that were constructed for this study and it is unknown how much insight these findings could provide for campaigns in the future given that the technological environment is shifting. The findings presented in this study present a narrative of how the features of one technology facilitated different forms of interaction over the course of an election period and who the most salient figures and information artifacts were.

An ongoing examination of how social media is used in the political context is needed. The methodological approaches of dataset construction, collection and analysis need to be documented and shared with the broader community of political and socio-technical researchers. As detailed in chapter three of this study, identifying the specific dataset, method, script and output is imperative to illustrate the provenance of data and how collection approaches carry through to the actual findings.

In addition to detailed methodological documentation, researchers need to be willing to share the datasets along with processing and analysis scripts used to arrive at the findings along with clarity surrounding how data was collected as detailed in chapter three and the appendices included in this document. This transparency will allow for the field to continue to advance and allow for researchers with different perspectives to derive insights from the same datasets. This will narrow the gap that currently exists between those that can collect interesting datasets and those with interesting research questions that can be answered using datasets that they may not have technical capabilities to collect.
List of References


Kooti, F., Yang, H., Cha, M., Gummadi, K., & Mason, W. A. (2012). The Emergence of Conventions in Online Social Networks.


Zhang, X., Fuehres, H., & Gloor, P. A. (2010). *Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"*. Proceedings from Collaborative Innovation Networks Conference, Savannah, GA.

Appendix A: Number of Tweets Collected by Query

Table 39 illustrates the number of tweets collected for each of the queries in the dataset. In total, 47 hashtags, 15 Twitter handles and 6 keywords were collected. These queries are also case independent although they are presented in the way that they were entered in the collection system.

Table 39: Tweets by Job

<table>
<thead>
<tr>
<th>Query</th>
<th>Frequency</th>
<th>First Seen</th>
<th>Query</th>
<th>Frequency</th>
<th>First Seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>mitt and romney</td>
<td>10,400,391</td>
<td>8/20/12</td>
<td>#ivoted</td>
<td>192,295</td>
<td>11/5/12</td>
</tr>
<tr>
<td>@BarackObama</td>
<td>8,592,370</td>
<td>8/20/12</td>
<td>#BarackObama</td>
<td>166,818</td>
<td>8/20/12</td>
</tr>
<tr>
<td>President AND obama</td>
<td>6,630,349</td>
<td>8/20/12</td>
<td>@whitehouse</td>
<td>160,102</td>
<td>8/20/12</td>
</tr>
<tr>
<td>barack and obama</td>
<td>4,114,765</td>
<td>8/20/12</td>
<td>#enndebate</td>
<td>145,266</td>
<td>8/21/12</td>
</tr>
<tr>
<td>#Obama</td>
<td>3,911,494</td>
<td>8/20/12</td>
<td>@AnnDRomney</td>
<td>142,960</td>
<td>8/20/12</td>
</tr>
<tr>
<td>@MittRomney</td>
<td>3,908,195</td>
<td>8/20/12</td>
<td>#47percent</td>
<td>139,469</td>
<td>6/18/06</td>
</tr>
<tr>
<td>#obama2012</td>
<td>2,495,489</td>
<td>8/20/12</td>
<td>#Biden</td>
<td>135,036</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#Romney</td>
<td>2,225,148</td>
<td>8/20/12</td>
<td>#denverdebate</td>
<td>130,822</td>
<td>9/26/12</td>
</tr>
<tr>
<td>paul AND ryan</td>
<td>2,140,171</td>
<td>8/20/12</td>
<td>@lynndebate</td>
<td>105,460</td>
<td>10/22/12</td>
</tr>
<tr>
<td>#romneyryan2012</td>
<td>1,647,359</td>
<td>8/20/12</td>
<td>#govote</td>
<td>95,841</td>
<td>11/6/12</td>
</tr>
<tr>
<td>#election2012</td>
<td>1,632,921</td>
<td>8/20/12</td>
<td>#decision2012</td>
<td>90,936</td>
<td>8/20/12</td>
</tr>
<tr>
<td>voted and for</td>
<td>1,476,786</td>
<td>11/5/12</td>
<td>@RepPaulRyan</td>
<td>89,646</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#dnc2012</td>
<td>948,250</td>
<td>8/20/12</td>
<td>#JoeBiden</td>
<td>49,800</td>
<td>8/20/12</td>
</tr>
<tr>
<td>joe and biden</td>
<td>939,037</td>
<td>8/20/12</td>
<td>@demconvention</td>
<td>47,533</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#debates</td>
<td>886,074</td>
<td>9/23/12</td>
<td>#barack</td>
<td>36,975</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#debate</td>
<td>838,989</td>
<td>9/18/12</td>
<td>@gopconvention</td>
<td>36,307</td>
<td>8/20/12</td>
</tr>
<tr>
<td>@Michelleobama</td>
<td>742,788</td>
<td>8/20/12</td>
<td>#forwardnotback</td>
<td>29,718</td>
<td>7/28/06</td>
</tr>
<tr>
<td>#voteobama</td>
<td>722,628</td>
<td>8/20/12</td>
<td>@hofdebate</td>
<td>27,093</td>
<td>10/16/12</td>
</tr>
<tr>
<td>#gop2012</td>
<td>683,439</td>
<td>8/20/12</td>
<td>#2012dnc</td>
<td>26,930</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#gop2012</td>
<td>671,031</td>
<td>8/20/12</td>
<td>@VP</td>
<td>20,839</td>
<td>8/20/12</td>
</tr>
<tr>
<td>@PaulRyanVP</td>
<td>643,234</td>
<td>8/20/12</td>
<td>#debatedenver</td>
<td>20,532</td>
<td>9/29/12</td>
</tr>
<tr>
<td>#vote</td>
<td>562,937</td>
<td>11/5/12</td>
<td>#romneyshambles</td>
<td>17,330</td>
<td>9/17/12</td>
</tr>
<tr>
<td>#forward</td>
<td>519,601</td>
<td>8/21/12</td>
<td>#presdebate</td>
<td>15,745</td>
<td>9/29/12</td>
</tr>
<tr>
<td>@Obama2012</td>
<td>457,739</td>
<td>8/20/12</td>
<td>@PaulRyan</td>
<td>11,957</td>
<td>8/20/12</td>
</tr>
<tr>
<td>@JoeBiden</td>
<td>438,508</td>
<td>8/20/12</td>
<td>#centrepvdebate</td>
<td>9,519</td>
<td>10/6/12</td>
</tr>
<tr>
<td>#debate2012</td>
<td>424,333</td>
<td>9/18/12</td>
<td>#16trillionfail</td>
<td>8,901</td>
<td>9/5/12</td>
</tr>
<tr>
<td>Hashtag</td>
<td>Count</td>
<td>Date</td>
<td>Hashtag</td>
<td>Count</td>
<td>Date</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>#MittRomney</td>
<td>393,808</td>
<td>8/20/12</td>
<td>#letgaryjohnsondebate</td>
<td>5,904</td>
<td>9/29/12</td>
</tr>
<tr>
<td>#romney2012</td>
<td>353,334</td>
<td>8/20/12</td>
<td>#octobersurprise</td>
<td>4,733</td>
<td>9/23/12</td>
</tr>
<tr>
<td>#vpdebate</td>
<td>336,392</td>
<td>10/6/12</td>
<td>#2012rnc</td>
<td>2,570</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#cantafford4more</td>
<td>247,210</td>
<td>10/3/12</td>
<td>#demconvention</td>
<td>2,137</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#presidentialdebate</td>
<td>227,601</td>
<td>9/29/12</td>
<td>@Joe_Biden</td>
<td>1,749</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#electionday</td>
<td>221,610</td>
<td>11/5/12</td>
<td>#dudebate</td>
<td>850</td>
<td>9/29/12</td>
</tr>
<tr>
<td>#obamabiden2012</td>
<td>202,737</td>
<td>8/20/12</td>
<td>@PaulRyanPress</td>
<td>251</td>
<td>8/20/12</td>
</tr>
<tr>
<td>#forward2012</td>
<td>197,930</td>
<td>8/21/12</td>
<td>#romneyvp</td>
<td>10</td>
<td>8/20/12</td>
</tr>
</tbody>
</table>
# Appendix B: Datasets Used for Analysis

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rows</th>
<th>Description</th>
<th>Archived Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Dataset</td>
<td>52,487,179</td>
<td>This dataset was the complete dataset of 52,487,179 tweets that represent the elimination of duplicates from the combination of the Northern Virginia and Oregon datasets.</td>
<td><a href="http://goo.gl/7Cm5dU">http://goo.gl/7Cm5dU</a></td>
</tr>
<tr>
<td>Hashtag Dataset</td>
<td>26,322,494</td>
<td>This dataset consists of all of the tweets that contained a hashtag. This was used as the foundation for further slicing based on the presence of #election2012, #debates and the promoted hashtags.</td>
<td><a href="http://goo.gl/Xmdrr1">http://goo.gl/Xmdrr1</a></td>
</tr>
<tr>
<td>#Election2012 Subset</td>
<td>1,632,995</td>
<td>This dataset consists of all of the tweets that contained the #election2012 hashtag.</td>
<td><a href="http://goo.gl/3VBjrT">http://goo.gl/3VBjrT</a></td>
</tr>
<tr>
<td>#Debates Subset</td>
<td>1,089,254</td>
<td>This dataset consists of all of the tweets that contained the #debates hashtag.</td>
<td><a href="http://goo.gl/QkowWM">http://goo.gl/QkowWM</a></td>
</tr>
<tr>
<td>Promoted Hashtag Subset</td>
<td>1,421,776</td>
<td>This dataset consists of all of the tweets that contained the #cantafford4more, #forward, and #forward2012 hashtags.</td>
<td><a href="http://goo.gl/vhxHmh">http://goo.gl/vhxHmh</a></td>
</tr>
<tr>
<td>Debate Days Subset</td>
<td>10,319,375</td>
<td>This dataset consists of all of the tweets on the day of and day after the four debates during the 2012 election.</td>
<td><a href="http://goo.gl/z4QNZ9">http://goo.gl/z4QNZ9</a></td>
</tr>
<tr>
<td>At-mention Dataset</td>
<td>38,213,732</td>
<td>This dataset consists of all of the tweets that contained an at-mention.</td>
<td><a href="http://goo.gl/t4KZuK">http://goo.gl/t4KZuK</a></td>
</tr>
<tr>
<td>At-reply Dataset</td>
<td>4,606,908</td>
<td>This dataset is a subset of the at-mention dataset where the at-mention is at the first position of the tweet text. There are three subsets of this data that were created. These datasets included individuals that were only a source of an at-reply, those that were only a target and those that were on both sides of the conversation.</td>
<td><a href="http://goo.gl/aIEN1X">http://goo.gl/aIEN1X</a></td>
</tr>
<tr>
<td>URL Dataset</td>
<td>17,105,877</td>
<td>This dataset consists of all tweets that contained a URL.</td>
<td><a href="http://goo.gl/K88VPT">http://goo.gl/K88VPT</a></td>
</tr>
<tr>
<td>Retweeted URLs</td>
<td>9,000,739</td>
<td>This dataset is a subset of the URL dataset and contains all of tweets with a URL that were a retweet.</td>
<td><a href="http://goo.gl/228kLz">http://goo.gl/228kLz</a></td>
</tr>
<tr>
<td>At-reply URLs</td>
<td>523,216</td>
<td>This dataset is a subset of the URL dataset and contains all of tweets with a URL that were an at-reply.</td>
<td><a href="http://goo.gl/0MzlC8">http://goo.gl/0MzlC8</a></td>
</tr>
<tr>
<td>Retweet Dataset</td>
<td>28,922,377</td>
<td>This dataset consists of all tweets that were a retweet and was further slices by the syntax (&quot;RT @[username]&quot; or &quot;via @[username]).</td>
<td><a href="http://goo.gl/jNeSt1">http://goo.gl/jNeSt1</a></td>
</tr>
</tbody>
</table>
Appendix C: TwitterZombie Data Fields

TwitterZombie collects the following columns of data:

- **Tweet_id_str**: The unique Twitter assigned Tweet identifier
- **Job_id**: The TwitterZombie job
- **Created_at**: The date/time of the tweet as provided by the Twitter API
- **Text**: The actual tweet text.
- **From_user**: The Twitter handle of the tweet author
- **From_user_id_str**: The author of the tweets unique Twitter numerical identifier
- **From_user_name**: The author of the tweets display name in Twitter
- **To_user**: The Twitter handle that the tweet is directed to (if it is an at-reply)
- **To_user_id_str**: The unique Twitter numerical identifier of the Twitter user that the tweet is directed to
- **To_user_name**: The Twitter display name of the user that the tweet is directed to
- **Source**: The device or application used to author the tweet
- **Location_geo**: The geographic coordinates of where the tweet originated
- **Location_geo_0**: The latitude of where the tweet originated
- **Location_geo_1**: The longitude of where the tweet originated
- **Iso_language**: The native language of the Twitter device/application used to author the tweet.
- **Analysis_state**: A TwitterZombie assigned value to manage synchronization of data
Appendix D: TwitterZombieAnalysisScript.r Detailed

The TwitterZombieAnalysisScript ingests data from TwitterZombie using the format specified in Appendix C and identifies the presence of a syntactical feature using regular expressions. The script employs a series of string processing found in the R library “stringr” and arithmetic processes to calculate percentages and value counts as detailed in the below output descriptions. There is no error handling built into the script, but the file “syntacticalfeatureoverview” displays a list of all of the values calculated throughout the script. If an error occurs during processing this file will be blank, as all of the values will not have been calculated. Therefore, this is a mode of implicit error handling that the user must address.

Output: The analysis script exports a series of files that are used for further analysis. The script is modular so it allows for the commenting out of the output if a file is not needed for analysis. In the case of this study many of the files were not used as they provided analytics surrounding device and language usage. These files can be easily used for further analysis. The full list of possible files includes:

- **Devicecounts**: Contains a table of the applications used and the frequency of each.
- **Deviceedgelist**: Contains an edgelist with the source as the “from_user” and the target as the application used to create the tweet.
- **Fauxtoedge**: Contains an edgelist with the source as the “from_user” and the target as the to_user if the construct “@” is used.
- **Hashcounts**: Contains a table with the hashtags used in the dataset and the frequency of each that occurs. It is important to note that these calculations are done by first converting all of the hashtags to lowercase. Therefore, the counts are case-independent.
- **Hashnumber**: This file displays the distribution of the number of tweets that contain “n” hashtags each.
- **Languagedistribution**: This is a frequency count of the languages represented in the dataset.
- **Languageedgelist**: Contains an edgelist with “from_user” as the source and “language” as the destination
- **Linkcounts**: Contains a table with the shortened links that were extracted from the dataset and the frequency of each. It is important to note that many of the links may be expanded to the same link even though they do not appear as such. Therefore, this file is used as input to the LinkAnalysis script that expands the links and aggregates based on the expansion.
- **Linknumber**: This file displays the distribution of the number of tweets and links in each tweet.
- **Mentioncounts**: Provides a table with the mentions used in the dataset and the frequency of each that occurs. It is important to note that these calculations are done by first converting all of the mentions to lowercase. Therefore, the counts are case-independent.
- **mentionnumber**: This file displays the distribution of the number of tweets and mentions in each tweet.
- **Multiplelanguageusers**: This is a list of the individuals who used multiple languages in the dataset and the number of languages they used.
- **Replytoedge**: This is an edgelist of the at-reply’s that were identified in the dataset. The source column is the originating individual (from_user) and the target is the individual that is being directly mentioned in the tweet.
- **Retweetcounts**: This is a file that identifies the individuals who were the highest retweeted in the whole dataset.
- **Retweetedge**: This is an edgelist of the retweets in the complete dataset. The source column of the edgelist is the “from_user” and the target is the retweeted user.
- **Syntacticalfeatureoverview**: This file provides a number of measures in a table format that allows a researcher to get a quick sense of the file. These measures include:
  - **Numberoftweets**: This is the raw count of tweets in the dataset
  - **Linkpresencepercentage**: This is the percentage of tweets that contain a URL
  - **Hashpresencepercentage**: This is the percentage of tweets that contain a hashtag
  - **Mentionpresencepercentage**: This is the percentage of tweets that contain an at-mention
  - **Replytopresencepercentage**: This is the percentage of tweets that contain an at-reply
  - **Rtpresencepercentage**: This is the percentage of tweets that contain a retweet
  - **Viapresencepercentage**: This is percentage of tweets that contain “via @.” This is a form of retweeting using another syntax
  - **Mtpresencepercentage**: This is percentage of tweets that contain a modified tweets which has the syntax “MT @[username].” This is similar to a retweet, but the original text is modified in some way.
  - **Device duplicatepercentage**: This is the percentage of individuals who use more than one application to tweet in the dataset.
  - **Numberofdevices**: This is the number of total unique applications used in the dataset.
  - **Multipledeviceusers**: This is the raw count of the number of individuals who use more than one application to tweet in the dataset.
  - **Numberofpeopleretweeted**: This is the raw unique count of the number of people who were retweeted.
  - **Numberofuniquetweeters**: This is the raw unique count of individuals who tweeted in the dataset.
  - **Numberofuniquetags**: This is the raw unique count of hashtags
  - **Numberofuniquementions**: This is the raw unique count of @-mentions
  - **Numberofuniquelinks**: This is the raw unique count of shortened links in the dataset.
  - **Uniquetweets**: This is the raw unique count of unique text of tweets in the dataset
  - **Numberofsingleposters**: This is the raw count of individuals who posted only once in the dataset
  - **Singletonpercentages**: This is the percentage of individuals who only contributed once to the dataset.
  - **Percentageofenglishtweets**: This is the percentage of tweets that are in English as determined by the language_iso field in the dataset.
  - **Geopresence**: This is the percentage of tweets that have a specified geolocation as collected by Twitter Zombie.
  - **Fauxtopresencepresence**: This is the percentage of tweets that have the structure “.@[username]” at the first position of the tweet text.
  - **Meantweetcharacterlength**: This is the mean tweet text character length of the dataset.
  - **Internalpersonretweetinfluence**: This is a percentage of individuals who were retweeted that were internal to the dataset.
  - **Internalretweettextinfluence**: This is the percentage of retweet text that was internal to the dataset.
  - **Externalteretweets**: This is a raw count of the number of retweets that were external to the dataset.
  - **Obamapresencepercentage**: The percentage of tweets that contain President Obama’s Twitter handle
  - **Bidenpresencepercentage**: The percentage of tweets that contain Vice President Joe Biden’s Twitter handle.
  - **Ryanpresencepercentage**: The percentage of tweets that contain Paul Ryan’s VP candidate Twitter handle.
- **Mittpresencepercentage**: The percentage of tweets that contain Mitt Romney’s campaign Twitter handle.
- **Singleton**: This is a list of individuals who only tweeted once in the dataset.
- **Timeseries**: This file contains the number of tweets in the dataset from each day.
- **tweetcounts**: This file is a frequency count of all of the tweet text. This allows for the identification of common tweets. Higher frequency tweets tend to be retweets or short common phrases.
- **Ticketscreennamecounts**: This is a frequency table of the individuals that originated tweets in the dataset.
- **Uniquedeviceedgecountusers**: This provides a list of users that used more than one application to contribute a tweet to the dataset along with the number of applications that the individual used.
- **Uniquelanguageedgecounts**: This is a file that contains the number of languages that were used multiple times by individuals.
- **Ryansubset_filename**: A subset of all tweets that contain Paul Ryan’s VP candidate Twitter handle.
- **Bidensubset_filename**: A subset of all tweets that contain Vice President Joe Biden’s Twitter handle.
- **Obamasubset_filename**: A subset of all tweets that contain President Obama’s Twitter handle.
- **Mittsubset_filename**: A subset of all tweets that contain Mitt Romney’s campaign Twitter handle.

The script also outs an annotated version of the original dataset (**MARKEDUP_original_filename**). This file is an annotated version of the initial ingested data with the notation of whether each tweet contains a syntactical feature using TRUE/FALSE. The new columns in this file include:

- **To**: The user that the tweet is directed to (at-reply)
- **Fauxto**: The user that the tweet is publicly directed to (at-reply with a period preceding the @)
- **Fauxtopresence**: The binary presence of the public at-reply construct.
- **Replytopresence**: The binary presence of an at-reply
- **Retweet**: The Twitter handle that was retweeted
- **Via**: The Twitter handle that was retweeted using the “via @” construct.
- **Modifiedtweet**: The Twitter handle that was retweeted using the modified tweet retweet construct “MT @[username]”
- **Linkpresence**: The binary presence of a URL
- **Hashpresence**: The binary presence of a hashtag
- **Mentionpresence**: The binary presence of an at-mention
- **Rtppresence**: The binary presence of a retweet
- **Viapresence**: The binary presence of a retweet using “via @”
- **Mtpresence**: The binary presence of a modified tweet
- **Linknumber**: The number of URLs contained in the tweet
- **Hashnumber**: The number of hashtags contained in the tweet
- **Mentionnumber**: The number of at-mentions in a tweet.
Appendix E: Shortened URL Decoding Script Detailed

**Input:** The input to this script is the linkcounts.csv file that is output from the TwitterZombieAnalysisScript.r.

**Process:** This script ingests the linkcounts.csv file from the large analysis script and trims punctuation from the end of the link. First, the script decodes the original shortened link in the “decode” column. Second, it takes the link from “decode” and attempts to expand it one more time into the column “decodemore.” Finally, it identifies the base of the URL by concatenating the link after the ending slash. It is possible that the decoded link did not have a / at the end of the link and if this is the case, the column “base” will have NA. It is also possible that the link in “decodemore” is still shortened as a result of numerous shortening actions taken by the user. Tests have identified this to occur less than 1 percent of the time and it is likely immaterial to analysis.

**Output:** The output file lists the original link, the original frequency, the first level of decoding and the second level of decoding along with the base URL of the link.
Vita

Christopher Mascaro was born in July 1982 in Holyoke, Massachusetts. He graduated from The University of Michigan with a BA in Political Science in 2004. He went on to obtain an MS in Information Assurance from Capitol College in 2009 and a MA in Government from Johns Hopkins University in 2010. He currently has over 25 academic publications spread between journals and conferences that focus on the analysis of electronic trace data in social media, discussion forums and bespoke communication systems.