Information Sharing in Political Discourse: An Examination of URL Sharing in Twitter During a Presidential Election

Christopher Mascaro, Ph.D.
Drexel University

Sean Goggins, Ph.D.
University of Missouri

Keywords: *Politics, social media, information sharing, twitter, archiving*

ABSTRACT

Sharing information external to a social media platform affords an individual the ability to augment their participation. Information sharing is one of the foundations of political discourse and allows individuals to augment their participation and allows for the justification of positions using information external to a technology as evidence. Analysis of information sharing in the context of political discourse on Twitter during the 2012 United States Presidential Election shared through URLs reveals that that information shared through URLs was predominantly user-generated content from Twitter and mass media information. Further, approximately 5% of the URLs that were linked to within the tweets no longer had the original information available. These findings illustrate that Twitter is a reflexive sharing environment and demonstrates the need for social media archiving to capture ephemeral content that may exist outside of the platform.

INTRODUCTION

Political discourse is one of the most common Internet activities. This discourse occurs across a variety of online platforms, but is heavily concentrated in social media platforms as they provide features to enrich and facilitate dialogue. Social media provides a number of syntactical features that allow for individuals to engage and share information in ways not previously possible. These syntactical features vary by technology, but have similar capabilities in that they allow users to identify specific individuals with whom they want to share information. This information is often hosted on websites external to the medium of sharing. The analysis of websites that are shared within a social media platform can illustrate the type of discourse and information sharing that is occurring.

Twitter gives users the ability to direct messages at others (at-replys), share content originally posted by others (retweet) and highlight content to individuals through the inclusion of their Twitter handle in the message (at-mention). These three distinct syntactical structures allow individuals to carry out different activity within the social media platform. The type of activity and meaning of this activity has been studied at length [2; 3; 5; 6; 11; 13; 32; 36], but there has been limited analysis of any differences or similarities in the type of information individuals share when using specific syntactical features.

In the 2012 Presidential election, citizens primarily shared URLs that referenced user-generated content from platforms 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 a significant percentage of the presence of 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 syntactical feature. This suggests that information sharing was actually the recycling of other information as opposed to unique contributions to information sharing. The findings have numerous implications for understanding how campaigns and the citizenry utilize social media and how findings from previous elections can contribute to a better understanding of activity in future elections.

LITERATURE

Twitter is unlike any other social media platform. Within 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 because the company provides three APIs that make large amounts of data accessible to the research community. As a result of access to this data, Twitter research has evolved and increased from Twitter’s debut in 2006.

Early research on Twitter activity attempted to understand user behavior [19] and user intent when using certain syntactical features [7; 14; 19]. This research established a baseline for understanding how individuals behaved in a new socio-technical environment. Findings from these early studies have limited application to understanding today’s Twitter environment as both the technology and user base have evolved.

The evolution of syntactical feature usage is likely attributed to the changing 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 [14]. In the following 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 addressed in traditional media and allows for coverage in traditional media to be augmented when Twitter users actively spread breaking news [37; 38].

Twitter syntactical features

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

Twitter has several syntactical features that allow for a richer social media experience, including: at-replys, retweets, URLs, at-mentions, hashtags and direct messages. Most of these features such as the hashtag have emerged organically from the user base and have evolved through varied usage by different sets of users. This means that each syntactical feature requires specific conceptualization in different contexts to properly research and compare findings across different datasets. Table illustrates the commonly used Twitter syntactical features with a description of each based on an existing review of the literature. Following Table 1, I review existing literature on the use of each feature. As illustrated by the variety of studies, the findings of studies examining Twitter activity in the context of syntactical features are highly dependent upon the topical content and therefore it is imperative that the context be understood.

[INSERT TABLE 1 HERE]

Information Exchange Through URLS

URLs play an important role in sharing information online. URLs have been studied extensively [21]. In the political domain, analysis of URLs has been used to gauge the interactivity of campaign websites [33]. Other research has identified URLs as indicators of campaign positions and links to sites that are external to a candidate indicate specific relationships between the candidate and organizations responsible for the content [9]. In the United States and other countries, links between websites of candidates and elected individuals can illustrate shared political ideology and partisan connections [16; 21-23].

In journalism, linking through the use of shared URLs between newspaper websites has been shown to facilitate the sharing of information and has lead to an increase in page views as audience size increases [34]. Analysis of linking 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 [25].

In the context of social media there has been limited analysis of URLs and how they relate to interactions within the technology. Robertson et al. (2009) found that on Facebook there is a small number of domains that account for a majority of the URLs that are shared [24]. Further, these URLs were used for a number of purposes including: 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 Twitter used 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 websites hosting user-generated content [12]. Other research has found that tweets are seen as more credible if the base URL included in the tweet is from a known source [4].

Shortened URLs are important in socio-technical systems that privilege the number of characters. Research on 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 [17]. Other research has found that shortened URLs tended to point to one of two categories of information, high quality information and spam [15]. Research to develop systems to identify these different types of URLs has had some success [35]. 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 [8].

One of the issues 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 [27]. This study echoes earlier research, which found that web pages change locations online or disappear as time passes [18]. 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 the types of tweets that the different URLs are embedded in and this analysis is done in the following study. 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 [30].

DATASET

The data collection for the study occurred over an 85 day time period beginning before the Republican National Convention in August 2012 and continuing through the week following Election Day in November 2012 (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, which mark 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 emerging events. These queries include candidate handles, hashtags and keywords that attempted to produce a representative sample of Twitter data during the time period of interest. The data collection for this study was done using a modified version of the TwitterZombie [1] built using cloud infrastructure to provide collection elasticity for scalability based on collection requirements such as the occurrence of an acute event such as a debate.

The initial set of queries was established by identifying all Twitter handles associated with the candidates and the campaigns, including the wives of the candidates because they 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 instances when individual references a candidate without using their Twitter handle. Additionally, specific hashtags that were used in the candidates' Twitter feeds were collected as these represented topics of discourse. Throughout the data collection period, new hashtags were identified as associated with electoral events and promoted by candidates.

In total, 52,487,179 unique tweets were captured using the combination of 68 queries (47 hashtags, 15 handles, 6 keyword searches). Analysis of the larger dataset is the scope of additional studies. The following study presents an analysis of tweets that contained at least one URL. In total, 32.59% of these tweets contained a URL and a number of these were retweets or at-replys. An analysis of the different types of tweets is presented. These tweets comprise the dataset used for the following study.

[INSERT TABLE 2 HERE]

RESEARCH QUESTIONS

The analysis of URLs within tweets is currently a gap in the research related to social media. The following research questions are used to guide the analysis of the previously described dataset.

1. What types of websites were shared within tweets during the 2012 Presidential Election?
2. To what extent do the type of URLs shared in retweets and at-replys differ from those shared in tweets with other syntactical features?
3. To what extent are sources of information that are linked to within the tweet still available?

METHODS

In order to address the research questions three datasets were created. The first dataset of consists of all 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 URLs that were part of an at-reply message (referred to as at-reply URL dataset). The retweet and at-reply datasets are subsets of the complete URL dataset.

Twitter uses shortened URLs for most 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. As a result of this, a process of decoding the collected URL is required to study the website that the URL is pointing to. In order to find the original URL from shortened URLS a script was written that would ingest a list of shortened URLs and expand them[1]. In this study, each URL was decoded three times. The actual decoding process was run on a set of Amazon Web Services EC2 instances to allow for parallel decoding processes[2]. 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 in the dataset.

[INSERT TABLE 3 HERE]

[1] An electronic version of this script can be provided for review.

[2] An EC2 instance is a virtual machine that is hosted at an Amazon colocation facility.

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.

[INSERT TABLE 3 HERE]

Coding Base URLs

One of the difficulties of sharing URLs in social media is that URLs can include extraneous information that may include the technology used to access them or other information that is not relevant to the specific story. To lessen these ambiguities, the authors conducted the analysis using only the base URL. 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 marker were removed.

Examining the base URLs of the information that is being shared affords the opportunity to understand the information sources being used in the discourse. There currently exists no accepted way to code URLs in social media given the variety of studies and domains. One of the reasons for this difficulty is that over time it is likely that the categorizations of a website may change. Huffington Post began as a blog and has since 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 as belonging to mass media [12] and the type of activity that was being discussed [10]. 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. These websites were controlled by the candidate or party and did not have a relationship to an external media organization or company.

There are only a limited number of user-generated websites, so only the top 50 most frequently occurring 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. A URL in the top 50 has been 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 URLs that are no longer active (Dead URLs).

Classifying and Analyzing Dead URLs

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 [27-29]. 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 [17]. 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.

FINDINGS

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 3). 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.

[INSERT TABLE 4 HERE]

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 datasets, the complete dataset, retweet dataset and at-reply dataset.

President Obama's official campaign website was the fourth most retweeted base URL and campaign information related to 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 more than other sources.

Categories of URLs

In order to better understand the types of URLs shared, the top 50 base URLs were coded as being associated with the Campaign, Mass Media, User Generated content or NA for those that were unable to be included in the other categories. Mass media websites dominated the type of URLs that were shared (Table 4). There is a significant number of mass media websites that were shared across all three datasets. This high number reflects that the election is a prominent event for mass media outlets.

[INSERT TABLE 5 HERE]

The most significant difference between the complete, retweet and at-reply datasets is that more user-generated websites are shared in the at-reply dataset and fewer mass media websites are 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 of tweets with user-generated URLs across the three datasets ranges from 54.35% in the complete dataset to 65.07% in the at-reply dataset (Table). Additionally, more URL shorteners 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 information shared in the other two datasets and that some of this information was no longer accessible.

[INSERT TABLE 6 HERE]

URLs Presence by Syntactical Feature

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 top 50 URLs in the complete, retweet and at-reply datasets illustrates that the most prevalent base URLs occurred at a similar frequency 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 links and aggregators as their base URL may be ambiguous), are examined to provide context for how the presence of the base URLs differ in the dataset.

Retweets and at-replys play a significant role in the sharing of the most frequently occurring URLs (Table 6). The most frequently occurring URL in the dataset, twitter.com had over 85% of its presence in the dataset as a result of being part of an at-reply or retweet. Retweeting a URL inflates the number of times it is present in the dataset as a retweet is not an original piece of information. Therefore, using a URL in an at-reply or retweet constitutes a different type of behavior than using it independently of one of these syntactical features as this constitutes an individual sharing someone else's information (retweet) or highlighting information to someone else (at-reply).

[INSERT TABLE 7 HERE]

A large number of other popular URLs relied only slightly less 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-replys.

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% (mean = 64.94%; median = 71.97%).

This skew towards the higher percentages indicates syntactical features such as retweets and at-replys contribute to higher counts of certain types of 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.

Dead URLs

Approximately 5% of the collected shortened URLs were unable to be decoded. Shortened URLs were decoded during a time period lasting 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 Twitter 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.

[INSERT TABLE 8 HERE]

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.

DISCUSSION

The findings presented here make practical and theoretical contributions to a better understanding of how information is shared in a political campaign on social media. The differences in the type of URLs shared and how these URLs were shared within the context of different syntactical features has significant practical implications for campaigns and citizens. Individuals shared over 17 million URLs in the course of the 2012 election cycle on Twitter. This sharing occurred in numerous distinct contexts within Twitter. Those contexts span candidate, issue and event focused frames. Considering the specific type of contextual framing, and how those frames influence the use of URL's will contribute to understanding the framing of political discourse on Twitter during future election cycles. For example, concrete experiments in information sharing during election cycles could include the injection of information focused hashtags instead of political party, candidate or issue focused hashtags. Similarly, journalists could employ a range of new syntactic features aimed at a more egalitarian pattern of information sharing.

The potential of reframing the use of syntactic features around democratic ideals underscores how social media is a focal point for political discourse and in each election candidates continue to evolve strategy and focus on engaging the public through the numerous technologies. The prevalence of this activity has numerous practical implications for socio-technical and political research and for those involved in campaign and sharing information within these technologies. Understanding the type of information exchange and how the information is exchanged is integral to building better systems and further study of such activity.

A significant majority of URL sharing occurred as a result of being retweeted or as part of an at-reply message. The fact that the most shared content was content that was internal to Twitter suggests that there is some insularity of the activity. The presence of a significant number of user-generated URLs in at-reply messages demonstrates that individuals are using the at-reply syntactical feature to share information from other social, user-generated mediums with each other. This activity may represent some aspects of a filter bubble in that the exposure and sharing of certain activity is dictated by the technology and choices individuals make within the technology [20]. Information diffusion and journalistic hashtags, combined with the potential for introducing new syntactic features aimed at democratic electoral ideals could all explicitly influence future technology choices.

The significant percentage of URLs that were included in the dataset as a result of retweets illustrates that certain types of information are more likely to be shared through syntactical features native to the technology. Campaign related websites were the category of URLs most likely to be shared through retweets. This illustrates a unique form of information sharing and the possibility that the URLs were being shared as part of a larger campaign. The prevalence of official campaign material shared via retweets is likely the result of campaign tweets first introducing the information within Twitter. The two most retweeted accounts in the dataset were @barackobama and @mittromney with 3,399,011 and 417,893 retweets respectively. These numbers do not account for all of the campaign related retweet sharing, but it may help to explain the significant prevalence of campaign related information in retweets.

The knowledge of who is more likely to be the subject of retweets and as a result have their information proliferated further can allow for new types of microtargeting of messages and individuals within social media. 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 microtarget on a large scale [31]. The large number of individuals that are using public channels to communicate, in the context of this research Twitter, means that the efforts to understand what this communication means and how actions can be taken to influence this communication is increasing. In the 2012 Presidential campaign, the Obama campaign had a team of dozens working for 18 months using data mining to identify undecided voters [26]. 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 augment the existing targeting efforts and understanding how different information is being shared is one such area.

Twitter as a new public sphere

One of the questions of social media activity is whether constructive discourse is occurring or if the activity represents noise. Dahlberg (2001) specified six criteria based on Habermas (1984) that must be met in order for online discourse to be classified as rational-critical discourse. One of these criteria included "exchange and critique of reasoned moral-practical validity claims." The extensive use of URLs in the discourse illustrates that individuals used some form of reasoning when making their comments as they were linking to other information found internal or external to the technology.

As evidenced by the type of URLs shared, many were 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 other criteria set forth by Dahlberg at scale, but the presence of significant amounts of specific types of tweets (at-replys, retweets) identified in the overview of the dataset suggests that other criteria may be met after further study.

Implications for Information Consumption

A large number of the shortened URLs 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 [27; 28]. 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.

The resolution of URLs is something that is not possible for many individuals who may be consuming the information within the technology. The absence of the original data source alters the consumption of the information, especially when significant time passes between the posting and consumption of the data. Social media platforms such as Twitter should provide caching mechanisms for the URLs that are shared within the technology at the moment that they are shared. This type of service would mimic similar features of search engines such as Google that provide access to cached versions of websites. Without this type of caching mechanism the meaning of many tweets may eventually be lost to the detriment of both future research and personal value.

[1] Black, A., Mascaro, C., Gallagher, M., and Goggins, S. 2012. TwitterZombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere. ACM Group.

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

[3] Bruns, A. and Burgess, J. E. 2011. #Ausvotes: how Twitter covered the 2010 Australian federal election. Communication, Politics and Culture. 44, 2, 37-56.

[4] Castillo, C., Mendoza, M., and Poblete, B. 2011. Information credibility on twitter. Proceedings of the 20th international conference on World wide web, 675-684.

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

[6] Chang, H.-C. 2010. A New Perspective on Twitter Hashtag Use: Diffusion of Innovation Theory. ASIS&T.

[7] De Choudhury, M., Diakopoulos, N. A., and Naaman, M. 2012. Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories. CSCW.

[8] Devesa, J., Cantero, X., Alvarez, G., and Bringas, P. G. 2011. An efficient Security Solution for Dealing with Shortened URL Analysis. WOSIS.

[9] Foot, K., Schneider, S. M., Dougherty, M., Xenos, M., and Larsen, E. 2003. Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere. Journal of Computer-Mediated Communication. 8, 4, 0-0.

[10] Himelboim, I., McCreery, S., and Smith, M. 2013. Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. Journal of Computer-Mediated Communication.

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

[12] Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., and Ma, K. L. 2012. Breaking news on twitter. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, 2751-2754.

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

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

[15] Kandylas, V. and Dasdan, A. 2010. The utility of tweeted URLs for web search. Proceedings of the 19th international conference on World wide web, 1127-1128.

[16] Kim, J. H., Barnett, G. A., and Park, H. W. 2010. A hyperlink and issue network analysis of the United States Senate: A rediscovery of the Web as a relational and topical medium. Journal of the American Society for Information Science and Technology. 61, 8, 1598-1611.

[17] Klien, F. and Strohmaier, M. 2012. Short links under attack: geographical analysis of spam in a URL shortener network. Proceedings of the 23rd ACM conference on Hypertext and social media, 83-88.

[18] Koehler, W. 1999. An analysis of web page and web site constancy and permanence. Journal of the American Society for Information Science. 50, 2, 162-180.

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

[20] Pariser, E. 2011 The filter bubble: What the Internet is hiding from you. Penguin UK.

[21] Park, H. W. 2003. Hyperlink network analysis: A new method for the study of social structure on the web. Connections. 25, 1, 49-61.

[22] Park, H. W. and Thelwall, M. 2008. Link analysis: Hyperlink patterns and social structure on politicians' web sites in South Korea. Quality & Quantity. 42, 5, 687-697.

[23] Park, H. W., Thelwall, M., and Kluver, R. 2005. Political hyperlinking in South Korea: Technical indicators of ideology and content. Sociological Research Online. 10, 3.

[24] Robertson, S. P., Vatrapu, R. K., and Medina, R. 2009. The Social Life of Social Networks: Facebook Linkage Patterns in the 2008 U.S. Presidential Election. International Digital Government Research. 6-15.

[25] Romero-Frías, E. and Vaughan, L. 2012. Exploring the relationships between media and political parties through web hyperlink analysis: The case of Spain. Journal of the American Society

[26] Rutenberg, J. 2012. Data You Can Believe In: The Obama Campaign's Digital Masterminds Cash In. New York Times. Sunday Magazine, MM22.

[27] SalahEldeen, H. and Nelson, M. L. 2012. Losing my revolution: How much social media content has been lost. International Conference on Theory and Practice of Digital Libraries TPDL,(Submitted for publication), TPDL 12, Submitted for publication.

[28] SalahEldeen, H. M. and Nelson, M. L. 2013. Carbon dating the web: estimating the age of web resources. Proceedings of the 22nd international conference on World Wide Web companion, 1075-1082.

[29] Salaheldeen, H. M. and Nelson, M. L. 2013 Resurrecting My Revolution. In Research and Advanced Technology for Digital Libraries, Springer.

[30] Scifleet, P., Henninger, M., and Albright, K. H. 2013. When social media are your source. informationr.net.

[31] Sosnik, D. B., Dowd, M. J., and Fournier, R. 2006 Applebee's America: How Successful Political, Business and Religious Leaders Connect with the New American Community. Simon & Schuster Paperbacks.

[32] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. International Conference on Social Computing. 177-184.

[33] Trammell, K. D., Williams, A. P., Postelnicu, M., and Landreville, K. D. 2006. Evolution of Online Campaigning: Increasing Interactivity in Candidate Web Sites and Blogs Through Text and Technical Features. Mass Communication and Society. 9, 1, 21-44.

[34] Weber, M. S. 2012. Newspapers and the Long-Term Implications of Hyperlinking. Journal of Computer-Mediated Communication. 17, 2, 187-201.

[35] Yang, S., Chitturi, K., Wilson, G., Magdy, M., and Fox, E. A. 2012. A study of automation from seed URL generation to focused web archive development: the CTRnet context. Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, 341-342.

[36] Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., and Su, Z. 2010. Understanding retweeting behaviors in social networks. Proceedings of the 19th ACM international conference on Information and knowledge management, 1633-1636.

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

[38] Zhao, W. X. and Jiang, J. 2011. An Empirical Comparison of Topics in Twitter and Traditional Media. Singapore Management University School of Information Systems Technical Papers Series.

Table 1: Conceptualization of syntactical features

Syntactical feature	**Syntax as applied to this study**	**Purpose**
At-Reply	@[username] at first position of tweet text	To directly address another individual in a public manner
At-Mention	@[username] at any point in tweet text	To highlight a tweet to another individual or to talk about someone. Mentioning them will inform them of the tweet
Retweet	RT @[username] "tweet text"	To further disseminate another individual's tweet.
URLs	http://t.co/[6-10 characters]	To include external information in a tweet. Note: Twitter uses a URL shortener
Hashtags	#[alphanumeric text]	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

Table 2: Syntactical Feature Percentage Presence Summary by Day

Syntactical Feature	Overall	Minimum Percentage	Maximum Percentage	Mean	Median	Range
URL	32.59%	16.74%	67.28%	40.92%	42.54%	50.54%

Table 3: URLs by Dataset and Date Decoded

Dataset	Number of Decoded URLs	Time Period Decoded
Complete	17,356,265	2/7/2013-2/21/2013
Retweet	10,448,022	2/1/2013-2/6/2013
At-reply	712,402	1/25/2013-1/29/2013

Table 4: Most Frequently Occurring URLs by Tweet Type

Complete	**Retweet**	**At-reply**
twitter.com	twitter.com	twitter.com
youtube.com	youtube.com	youtube.com
instagram.com	t.co	huffingtonpost.com
huffingtonpost.com	barackobama.com	t.co
barackobama.com	huffingtonpost.com	barackobama.com
t.co	instagram.com	instagram.com
mi.tt	mi.tt	twitpic.com
trib.al	trib.al	twitlonger.com
twitpic.com	twitpic.com	mi.tt
washingtonpost.com	politifact.com	facebook.com

Table 5: Top 50 Base URL Codes by Dataset

	Complete	Retweet	At-reply
Campaign	4	8	5
Mass Media	36	33	30
User Generated	7	7	10
NA	3	2	5

Table 6: Percentage of Overall Tweets with URLs by Code

	Complete	Retweet	At-reply
Campaign	7.01%	11.09%	5.53%
Mass Media	36.32%	31.48%	22.73%
User Generated	54.35%	56.25%	65.07%
NA	2.33%	1.19%	6.67%

Table 7: Syntactical Feature Contribution to Overall URL Presence

URL	Retweet	At-reply	Total Percentage
twitter.com	80.74%	4.54%	85.28%
youtube.com	64.60%	7.86%	72.46%
instagram.com	43.46%	1.44%	44.90%
huffingtonpost.com	81.35%	3.12%	84.47%
barackobama.com	75.34%	2.14%	77.48%
twitpic.com	77.71%	4.44%	82.15%
washingtonpost.com	57.27%	3.59%	60.86%
nytimes.com	47.95%	3.18%	51.12%
breitbart.com	65.76%	4.14%	69.91%
thinkprogress.org	75.09%	3.93%	79.01%
facebook.com	26.20%	4.86%	31.05%
politifact.com	82.68%	4.02%	86.70%
tumblr.com	13.53%	0.91%	14.44%
news.yahoo.com	27.53%	2.29%	29.82%
bbc.co.uk	52.16%	1.42%	53.58%
foxnews.com	65.64%	3.89%	69.52%
politico.com	70.73%	3.20%	73.93%
dailykos.com	72.83%	5.54%	78.37%
motherjones.com	79.42%	2.98%	82.40%
abcnews.go.com	67.38%	3.99%	71.36%

Table 8: Percentage of URLs unable to be Decoded By URL Shortener

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