Proceedings of the 2015 Industrial and Systems Engineering Research Conference
S. Cetinkaya and J. K. Ryan, eds.

Information Diffusion: A Study of Twitter During Large Scale Events

Christa Rogers-Pettie
University of Maryland-College Park
College Park, Md

Dr. Jeffrey Herrmann
University of Maryland-College Park
College Park, Md

Abstract

The diffusion of information through a population about a natural disaster or other emergency affects how and when the public reacts to the situation, including evacuation and the demand for assistance. Thus, it is important to understand how and at what speed important information spreads. Social media are an important part of this diffusion and provide a convenient and effective way to measure it. This study used data about a social network of 15,000 Twitter users and their tweets. Information such as the time of a tweet, the user name, the tweet content, and the tweet ID was analyzed to measure the diffusion of information and track the trajectory of retweets. The spread of information was visualized and analyzed to determine how far and how fast the information spread. The results showed how different types of information spread indicated the importance of different topics to these users. The network of popular users also demonstrated how effectively a message can pass through several users. Understanding how information spreads will benefit policy makers and emergency managers who want to get the right information to the right people, so that they can respond optimally to an emergency.

Keywords

Information Diffusion, Emergency Management, Twitter

1. Introduction

1.1 Motivation

This research consists of a study about information diffusion of large scale events. There are many ways people can receive and pass information such as, television, phone, email, text, tweets, blogs, articles, and posts. With all of these communication channels available, it is important to discover which is the most effective and why. This research attempts to understand characteristics of information diffusion specifically within social media. The purpose of this research is to find patterns of information diffusion on Twitter and learn about user behavior to help emergency managers craft appropriate messages and send them through the most efficient channels. By looking at user activity, this research will increase our understanding of user connectivity; furthermore, determining what information is important to users and for how long will provide insights into information diffusion of specific content.

There are several questions this research aims to answer:

- What types of messages are most popular?
- What users are most popular?
- How long does it take a tweet to travel through a data set?
- How connected are the users?

1.2 Literature Review

There have been various studies about information diffusion as the popularity of Twitter has increased. Previous research has characterized popular tweets by studying a single large scale event [1]. Trending topics (the most popular hashtags at any given time) have been analyzed over time to understand the distribution and decay rate of these popular topics [2]. Additionally, there is a unique nature of conversations on Twitter because of the inclusion of these hashtags [3]. By studying content, researchers can specifically find out what makes certain topics spread at different rates [4]. Research by Hong et al. found several key conclusions about predicting tweet popularity based on user activity and user popularity [5]. Other research has focused on various properties of diffusion like speed, scale, and range to predict diffusion patterns [6]. An economic approach was taken by studying consumer engagement on Twitter by comparing consumer engagement to revenue [7]. There are a variety of research methods and analysis processes relating to information diffusion and social media. More research needs to be completed on (1) how information diffuses during different types of large scale events and (2) the dynamics of user behavior.

1.3 Nomenclature

There are several new or unfamiliar terms used throughout this paper that are listed in this section. Some of the definitions have been taken from sources with business or research focused on social media. Some terms are specific to this academic research and defined as such to lessen confusion.

Tweet: A message posted by a user of 140 characters or less. It can include links which will be shortened to 30 characters or less [8]
Retweet: A tweet that is reposted and unchanged by a user. The only addition to the tweet is the mention of the Originator and the letters RT signifying the tweet is reposted
Originator: The first user to post a unique tweet
Retweeter: A user that reposts a unique tweet
Chain: A set of tweets in which the unique tweet and its retweets can be identified and listed together
Period of Relevance: A time period in which the information provided by a tweet is valuable to users
Verified Tweets: A tweet that comes from a current professional or unquestionable source [3]
Opinion Leader: An influential user that shares his or her opinion on social, political, worldwide, emergency, or important events to his or her network [3]
Reciprocal Relationship: Two users involved in at least two chains where the Originator of chain A is the Retweeter of chain B and the Retweeter of chain A is the Originator of Chain B
Loyalty: A user or users that retweet an Originator more than once

2. Methods

Information diffusion occurs when people pass information and share their opinions and experiences. Collecting data about leading topics potentially produces more topic-based conversation to study. Using Twitter to collect data presents a detailed view of what people are discussing and for how long. It also provides multiple options to analyze information diffusion.

2.1 Data Collection

Twitter is an international social network that allows users to share messages or tweets instantly, publically or privately with other users. Twitter has 241 million monthly active users, an average of 500 million Tweets are sent every day, 76% of Twitter’s active users tweet from their mobile device, and Twitter supports over 35 languages [8]. Social media provides convenient access for researchers to investigate the habits of its users. Twitter provides an easy way to collect data that tracks the activity of users as well as their interactions. Twitter data was collected and analyzed from four large scale events.

Dr. William Rand of the University of Maryland provided the data for this research. There were four topic-based datasets collected. Two were about weather emergencies: Hurricane Sandy and Hurricane Irene. The others were political events that affected the United States: the 2012 Presidential Election and the death of Osama Bin Laden. The application TwEater [9] was developed within the Center of Complexity and Business at the University of Maryland. This program collected tweets from Twitter’s Application Programming Interface that contained hashtags and keywords as assigned by the event that was occurring. The Osama Bin Laden data set was collected based on a time period, then further separated to find tweets with keywords related to the aforementioned event. The users that sent these tweets came from a network of 15,000 Twitter users (the “15K network”). The 15K network was selected based on their activity and connections. When a user tweeted about a topic, their tweet was collected by TwEater [9] and placed in a dataset. For example, a tweet with keywords “Obama” or “election” would be gathered by TwEater because those words match preselected key words. The same pool of 15K users is not used in each data set. Only those users that tweeted about a topic had their tweets included. So, not all users are represented in every data set, and the datasets do not have identical sets of users. More details about the data collection and 15K network is available in two referenced papers [10,11].

2.2 Data Analysis

The following statistics were used to characterize the data: Time Period of Collection, Total Number of Tweets, Active Users, Average Tweets per User, and Average Time between Tweets. The distribution of each data set shows how frequently users tweet. Furthermore, we identified very active tweeters and focused on their activity. For retweet chains, we investigated the difference between the number of times a tweet was found to originate within the dataset.

2.3 Retweet Chains

A retweet chain is a set of tweets in which the unique tweet and its retweets can be identified and listed together. Within a dataset, all of the tweets with the same Retweet ID form a retweet chain. If there is no Tweet ID that matched the Retweet ID, then the chain is an “outside” chain, meaning it was originated outside of the data set.

Analyzing these chains provides information such as the time required to retweet and how popular content is. To learn about groups of users, tweets of the Originators and Retweeters were explored by finding the tweets and retweets of each user. Once all statistics and calculations are complete, a search for patterns between datasets was pursued. Any patterns are analyzed by tweet content, originator, time, and volume. An example of a chain is in Table 1. A “1” in the *status is retweet* column means that a tweet originated from another user. A “-1” in the *Status is retweet of* column means tweet is original; however, if there is a long number in that column it is identifying the tweet id of the originally posted the tweet.

Table 1: A Retweet Chain from the Hurricane Irene Data Set

time	*tweet_id*	*user_id*	*tweet_text*	*Status is retweet*	*Status retweet of*	*Status retweet count*
10/27 16:19	262242329132421000	616173	Google's big Monday Android event is off on account of the hurricane http://t.co/a7YQRKzW	0	-1	0
10/27 16:19	262242320345350000	203871835	RT @TheNextWeb: Google's big Monday Android event is off on account of the hurricane http://t.co/FjMSOzGY by @alex	1	262242329132421000	25
10/27 16:20	262242387542294000	46958309	RT @TheNextWeb: Google's big Monday Android event is off on account of the hurricane http://t.co/FjMSOzGY by @alex	1	262242329132421000	62

To understand the connectivity each dataset was visualized based on its retweet chains. These chains are based on tweets that originate from within the network as well as those that start outside the dataset. Following this basic analysis, a network graph was created from the users involved in retweet chains. Each node in the graph represented a user who participated in a retweet chain as an Originator or a Retweeter. For each retweet, the graph includes an edge connecting the Originator and Retweeter which represents a retweet. NodeXL was used to visualize the graph. Colored and numbered edges illustrate users involved in the same retweet chain. For example, Figure 1 is the Retweet Chain Network from the Hurricane Sandy data set.

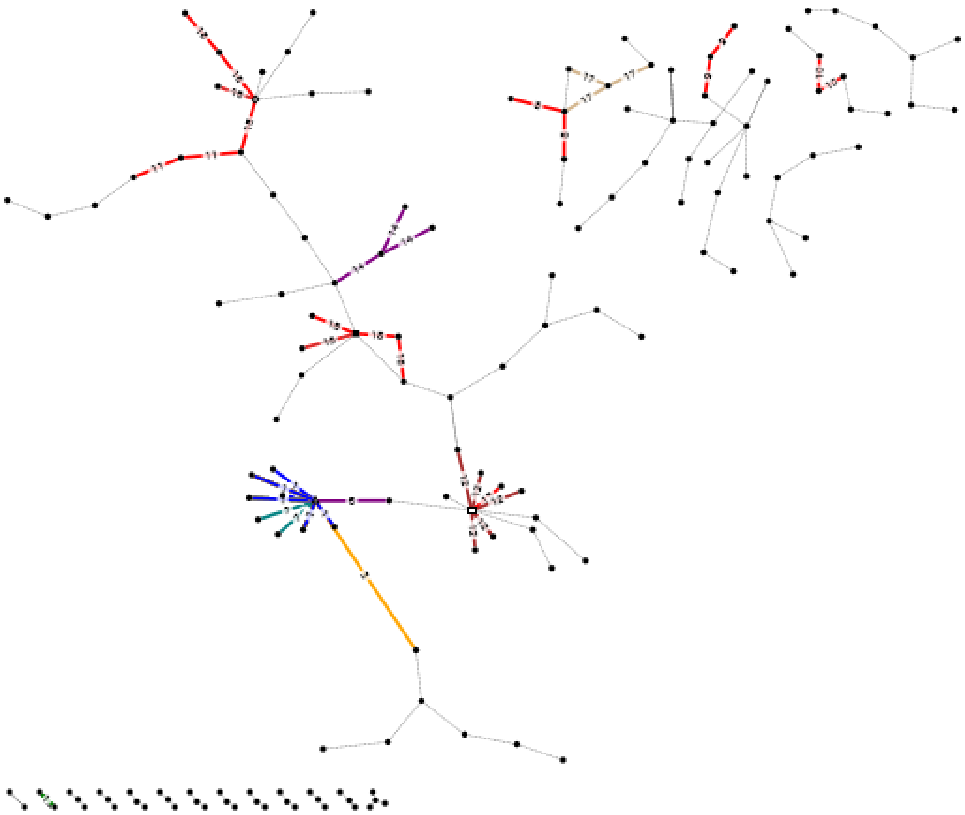
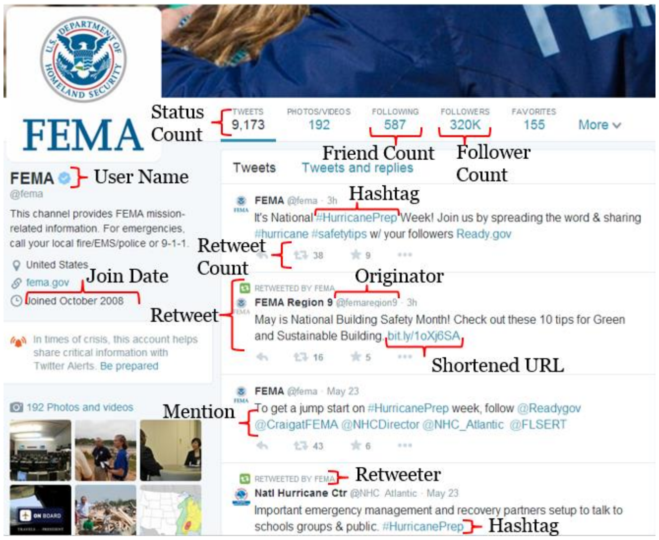


Figure 1: The Inside Retweet Chain Network for Hurricane Sandy. Colored and numbered edges illustrate users involved in the same retweet chain.

2.4 Content Analysis

To analyze the content of the tweets, the popularity of hashtag use was analyzed. Hashtags are the number symbol “#” followed by a word or phrase with no spaces (Ex: #election). The hashtag analysis process was completed by taking daily samples for the hurricane datasets and hourly sample for the election and Bin Laden data. Within the data set, the most frequent hashtags were recorded. Once the entire data set was finished, the tweets by the users involved in retweet chains were analyzed in the same fashion. When finished with both, the hashtag frequency data was charted for the entire dataset, the retweet chains, along with the distribution of tweets count over time (daily or hourly). These charts can show popularity of conversation topic in Retweet Chains compared to real time events and total tweet count.

Figure 2 shows an actual Twitter profile labeled with several key terms. The terms most important to this research are the User Name, Originator, Retweeter, Retweet Count, and Hashtags. Mentions, URLs, Friend Count, Follower Count, Status Count, and Join Date are collected but not a focus of the analysis. They could eventually be used to further this research.



FEMA @fema

This channel provides FEMA mission-related information. For emergencies, call your local fire/EMS/police or 9-1-1.

United States

fema.gov

Joined October 2008

In times of crisis, this account helps share critical information with Twitter Alerts. Be prepared

FEMA @fema · 3h
It's National #HurricanePrep Week! Join us by spreading the word & sharing #hurricane #safetytips w/ your followers Ready.gov
38 9

RETWEETED BY FEMA
FEMA Region 9 @femaregion9 · 3h
May is National Building Safety Month! Check out these 10 tips for Green and Sustainable Building bit.ly/1oXj6SA
16 5

FEMA @fema · May 23
To get a jump start on #HurricanePrep week, follow @Readygov @CraigatFEMA @NHCDirector @NHC_Atlantic @FLSERT
43 6

RETWEETED BY FEMA
Natl Hurricane Ctr @NHC_Atlantic · May 23
Important emergency management and recovery partners setup to talk to schools groups & public. #HurricanePrep

Figure 2: This Twitter profile belongs to the Federal Emergency Management Agency captured May 23, 2014. This shows actual tweet activity of the FEMA [8]

3. Results

3.1 Comparing the Data Sets

The Hurricane Irene and Hurricane Sandy data sets have the longest time of data collection; however, hurricane Irene has the lowest number of tweets and users. Table 2 shows a breakdown of these statistics. Hurricane Sandy has the highest average number of tweets per user for all users and high frequency users (users that tweet more than 20 times). These data sets have the longest time period of collection because they are prolonged weather events that people have time to track and prepare for. The Bin Laden data set had the most users and most tweets. The averages for tweets per user (normal and high frequency) were similar to the Sandy data set.

Table 2: The general statistics calculated for each data set

	Irene	Election	Sandy	Bin Laden
Time Period of Data Collection	17 days 9 hours 10 minutes	1 day 57 minutes	17 days 7 hours 1 minute	1 day 11 hours 57 minutes
Total Number of Tweets in Data Set	5948	9167	19085	27924
Number of Active Users	2210	2455	3325	4948
Average Number of Tweets Per User	2.69	3.94	5.74	5.64
Average Number of Tweets Per High Frequency User	28.68	40.13	47.20	45.51

It is important to understand how often users tweet. Some may tweet just once about a topic, while others create original tweets and pass on tweets about a topic frequently. As seen in Table 3, most users tweet only one time; however, that does not imply that those tweets make up the majority of the data set. In all data sets over 75% of the population tweets less than five times. In the Irene data set, 58% of the tweets are from users who tweeted less than 5 times total. In the Sandy and Bin Laden datasets, only 27% of the tweets were sent by users who tweeted less than 5 times. In these two data sets, at least 40% of the tweets were sent by users who tweeted 20 times or more. This statistic matches the large average of tweets from the high frequency users. The following chart and graph show statistics about how much of a data set is comprised of **tweets (PT)** from users that tweet a certain amount of times (1, <5, or >20) expressed in percentages. The chart also shows statistics of how much of the data set is comprised by the percentage of **users (PP)** tweeting a certain number of times and the **total number of tweets (N)** from these users.

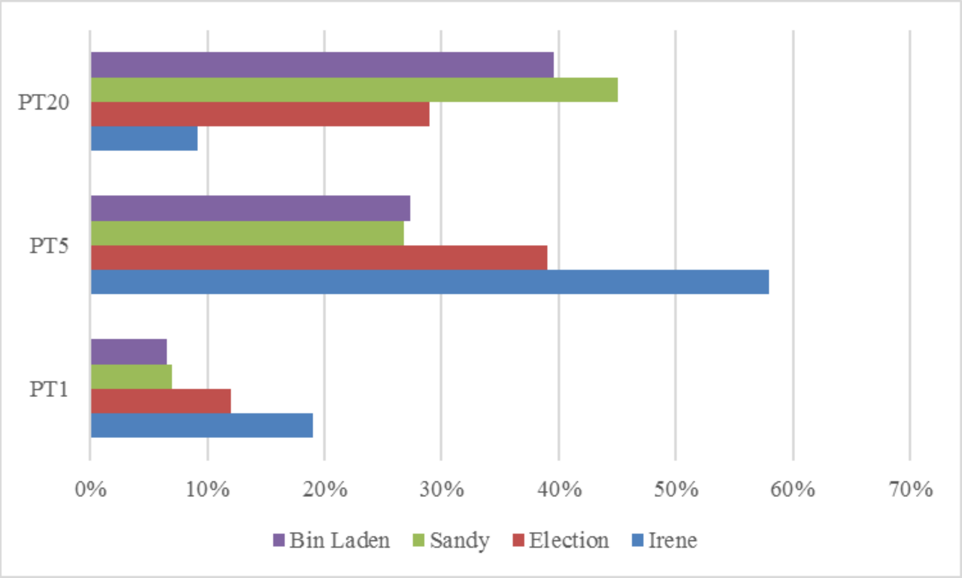


Figure 3: This graph show much (percentage) of the data set is comprised of tweets from users that tweet once (PT_1), five or fewer times (PT_5), and 20 or more times (PT_{20}). This is a comparison of *Tweet Frequency* in all data sets.

Table 3: Frequency of Tweets per user, Percentage Tweet of specific ranges (1, <5, and <20), and Percentage of Users to tweet within those ranges

	Irene	Election	Sandy	Bin Laden
N_1	1141	1142	1334	1826
PP_1	51%	47%	40%	37%
PT_1	19%	12%	7%	7%
N_1	1974	3808	2606	3765
PP_5	89%	84%	78%	76%
PT_5	58%	39%	27%	27%
N_{20}	19	70	182	243
PP_{20}	9%	3%	5%	5%
PT_{20}	9%	29%	45%	40%
Average Time to Tweet	0:04:12	0:00:09	0:01:21	0:00:35

The analysis of the chains gives a new perspective to the data sets. Sandy has the highest number of users involved in chains. Users can be involved in more than one chain by originating or retweeting, so some users may be represented more than once in these statistics. As shown in Table 4, Sandy also had the longest time to retweet. This means there was a lag between the original tweet and those following. There were 42 chains over 3 hours long that skewed this calculation. The chains with the lowest time to retweet were from Hurricane Irene. More chains would help to give the calculations for the Election and Irene stronger validity. Removing the outliers of chains over one hour lowers the time to tweet for chains of 1 tweet, making each data set more comparable.

Table 4: Chain Characteristics

	Irene	Election	Sandy	Bin Laden
Number of Chains Inside the Data Set	23	38	206	93
Number of Originators	23	38	206	93
Number of Retweeters	27	45	259	120
Average Time to Retweet	0:07:45	0:27:01	3:30:48	0:17:44
Chains over 1 Tweet: Average Time to Retweet	0:06:10	0:05:58	2:05:02	0:02:46
Chains of 1 Tweet: Average Time to Retweet	0:38:53	0:33:50	4:11:29	0:25:36
Chains of 1 Tweet: Average Time to Retweet (Chains Less than One hour)	0:08	0:12:20	0:11:17	0:02:25

By separating the analysis of users involved in chains by action (Originator or Retweeter), a pattern was easily distinguished about behavior. Retweeters will retweet more than Originators. Table 5 shows the average number of tweets and retweets per user and in every event, the average number of retweets is high for Retweeters. Table 5 also shows the percent of tweets sent by all users that are retweets. In all cases, Retweeters tweet at least 1.5 times as much as originators. Thus, they are more likely to pass on information than originators.

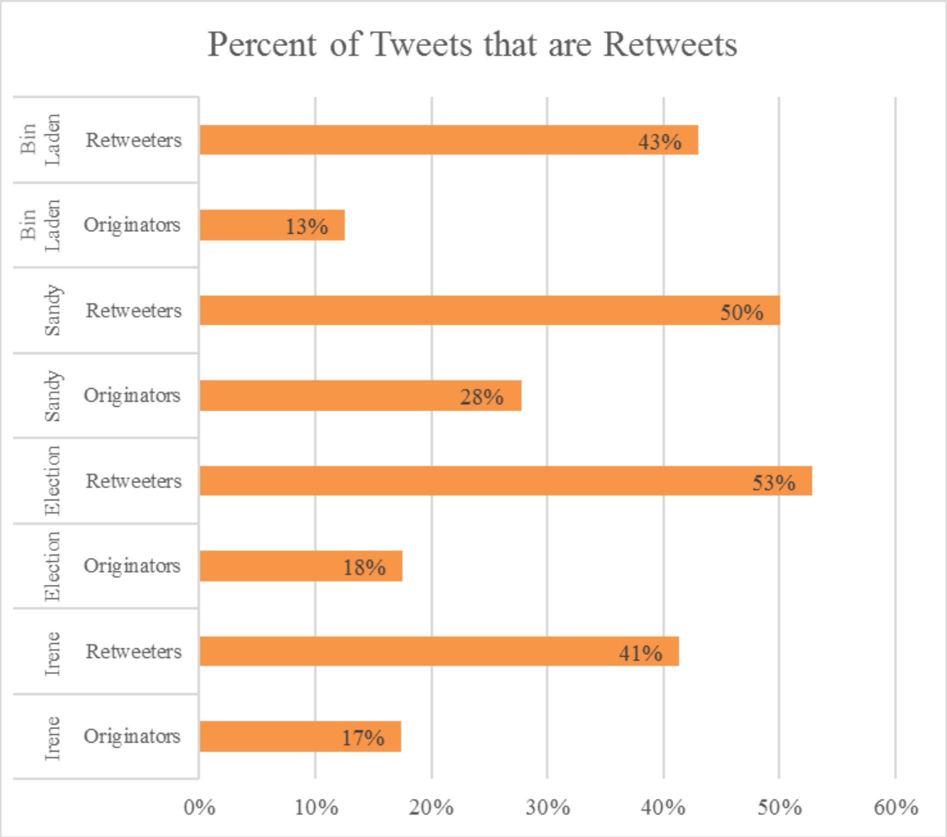


Figure 4: The characteristics of users involved in chains in all data sets were analyzed to see how much they tweeted and retweeted for their entire period of data collection. The results are separated by data set and how they were first identified within chains (Originator or Retweeter).

Table 5: Characteristics of Users involved in Chains

	Irene		Election		Sandy		Bin Laden	
	Originator	Retweeter	Originator	Retweeter	Originator	Retweeter	Originator	Retweeter
Number of Users	23	27	38	45	206	259	94	120
Total Tweets	184	198	701	528	5446	6460	2337	2694
Average Tweets Per User	8.00	7.33	18.45	11.73	26.44	24.94	24.86	22.45
Total Number of Retweets	32	82	123	279	1511	3235	336	2038
Average Retweets Per User	1.39	3.04	3.24	6.20	7.33	12.49	3.57	16.98
Percent of Tweets that are Retweets	17%	41%	18%	53%	28%	50%	13%	43%

4. Summary

This section summarizes the results of this study of information diffusion during large scale events. Section 4.1 is an overview of observations from the results section and commentary about tweets from the data sets. Section 4.2 highlights the importance and value of hashtag usage. Section 4.3 discusses the content of different types of tweets how that effects tweet popularity. Section 4.4 compares the connectivity of users within chain networks. Section 4.5 explains the bottom line and limitations of the data, results, and research method. Section 4.6 gives various ideas for additional research.

4.1 Observations

The purpose of this research was to find patterns and learn about user behavior to help emergency managers craft appropriate tweets and send messages through the correct channels. This analysis assessed connectivity and user behavior throughout a data set. Retweet chains were used to track how identical messages were passed from user to user on Twitter. From the retweet chains, it is evident that Retweeters within the chains are more likely to retweet throughout the dataset compared to Originators. Retweet chains were used to display user tendencies and simply how Twitter works. Most retweets originated from users outside of the data set.

If any tweet comes from a user with a large active and engaged network, that can lead to tweet popularity. Twitter activity is high during weather emergencies; thus, people are very willing to share weather information, and that information can come from all over the country to reach users in an affected area. The election was a different type of event, political, that can show different tendencies of users involved in more than one data set.

4.2 Importance of Hashtag Use

Hashtags are a great way to initially analyze content. It shows promise for influence metrics. The use of hashtags makes Twitter different from regular conversation. The hashtag can be interpreted as a way for a user to make a statement then add or promote a thought. By analyzing hashtag frequency for the entire data set and comparing that to tweets by users involved in chains only, shows how much influence that chain users have on how frequently the hashtags appeared.

4.3 Content

Throughout the analysis of these data sets, popular tweets were further classified and analyzed for their content. These tweets were studied to see why they were retweeted based on the type of content. This specific effort yielded many tweets from verified users and celebrities. There was further value added to the content analysis by researching the verified users and their activity. Retweets within the data set are retweeted more if they are informational, and the spread of information can be helpful if retweeted. Humorous tweets are effective in spreading information on Twitter as well. There is evidence that retweets with informational content can be spread quickly by many users. Having a celebrity or opinion leader as an originator causes the information to spread very quickly because these users often have more followers than the typical Twitter user.

4.4 Connectivity

The connectivity of the inside chain networks showed that Irene and Election data sets were not very connected by chains. Very few users were involved in more than one chain as either Originator or Retweeter. The Bin Laden data set showed more connectivity than both Irene and Election data sets; however, Hurricane Sandy's inside chain network had the most chains and users. There was a lot of connectivity with many users involved in more than one chain. Celebrity tweets always found their way into the data set by 15k users retweeting them.

4.5 Conclusions and Limitations

We developed an analysis procedure to look at the behaviors of users in various types of large scale events. The most popular tweets were informational tweets from verified sources with loyal Retweeters.

- What types of messages are most popular: Informational
- What users are most popular: Verified sources with loyal Retweeters
- How long does it take a tweet to travel through a data set: It varies depending on tweet type and event
- How connected are the users: Users sampled from 15K Network were not as connected as previously assumed

The data sets were very valuable, but several limitations were encountered during the analysis. The Election and Sandy Data sets had no limit on the retweet count. Unfortunately, the maximum retweet count of the Hurricane Irene and Bin Laden data was a limit of 101 retweets. Therefore, the number of retweets that occur not just in the data set, but throughout all of Twitter is not recorded after 101. So, there is no comparison that can be done with retweet count in and outside of the data set. If this data were available, it would allow a comparison between various types of events and how the network users perceive the importance of a tweet compared to the Twitter population..

The inside retweet chain networks show lack of connectivity, but that could mean that users are spreading the tweet information via word of mouth. A user could see a tweet and share it via text, verbal conversation, email, or another form of social media. This study could not measure if information from tweets reach users without recording their activity on Twitter.

Tweet chains travel in and out of the data sets, which is a limitation of the dataset's use in tracking trajectory. Analysis can be completed only on those tweets collected within our dataset, and the characteristics of the other tweets are unknown.

4.6 Future research

Several hypotheses could be tested in the next steps of this research. First, the use of informational messages from the same users are more effective than another type of message. Next, users with user loyalty are the most effective users to pass information. Finally, including hashtags with messages are more effective in spreading information compared messages without hashtags. These three hypotheses could provide a greater and more detailed understanding of the effects of content and user popularity.

Acknowledgments

I would like to thank Dr. William Rand, University of Maryland – College Park, for the use of Twitter data that facilitated my data analysis. I would also like to recognize the Louis Stokes Alliances for Minority Participation Bridge to Doctorate (LSAMP BD) Fellowship and the National GEM Consortium for financial support during my graduate education. I gratefully acknowledge my advisor, Dr. Jeffrey Herrmann, for his guidance throughout this research experience.

References

1. Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.
2. Asur, Sitaram, et al. "Trends in social media: Persistence and decay." Available at SSRN 1755748 (2011).
3. Murthy, Dhiraj. Twitter: Social Communication in the Twitter Age. Cambridge, UK: Polity, 201. Print.
4. Romero, Daniel M., Brendan Meeder, and Jon Kleinberg. "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter." Proceedings of the 20th international conference on World wide web. ACM, 2011.
5. Hong, Liangjie, Ovidiu Dan, and Brian D. Davison. "Predicting popular messages in twitter." Proceedings of the 20th international conference companion on World wide web. ACM, 2011.
6. Yang, Jiang, and Scott Counts. "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter." ICWSM 10 (2010): 355-358.
7. Joshi, Yogesh, Liye Ma, William Rand, and Louiqa Raschid. "Building the B[r]and." Marketing Science Institute. Marketing Science Institute, 2013. Web. Jan. 2014.
8. "About Twitter." Twitter. Twitter Inc, 2014. Web. 10 Apr. 2014.
9. Monner, Derek. "TwEater." GitHub. GitHub, 4 July 2013. Web. 01 Oct. 2013.
10. Sylvester, Jared, et al. "Space, Time, and Hurricanes: Investigating the Spatiotemporal Relationship among Social Media Use, Donations, and Disasters." Robert H. Smith School Research Paper No. RHS 2441314 (2014).
11. Herrmann, Jeffrey, et al. "An agent-based model of urgent diffusion in social media." Robert H. Smith School Research Paper (2013).