ABSTRACT

Title of Document: INFORMATION DIFFUSION: A STUDY OF A TWITTER DURING LARGE SCALE EVENTS

Christa D. Rogers, MS, 2014

Directed By: Associate Professor, Dr. Jeffrey Herrmann,
Mechanical Engineering

The diffusion of information through population affects how and when the public reacts in various situations. Thus, it is important to understand how and at what speed important information spreads. Social media platforms are important to track and understand such diffusion. Twitter provides a convenient and effective way to measure it. This study used data obtained from 15,000 Twitter users. Data was collected on the following events: Hurricane Irene, Hurricane Sandy, Osama Bin Laden's capture, and the United States’ 2012 Presidential Election. Information such as the time of a tweet, the user name, content, and the 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 show how information spreads and the content analysis of data sets indicate the importance of different topics to users.

INFORMATION DIFFUSION: A STUDY OF A TWITTER DURING LARGE SCALE EVENTS

By

Christa D. Rogers

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2014

Advisory Committee:
Associate Professor Dr. Jeffrey Herrmann, Chair
Dr. William Rand
Dr. Michael Fu

Dedication

This dissertation is dedicated to my loving husband, Clifton Pettie IV. I appreciate your support and patience throughout my graduate education.

Acknowledgements

I would like to give a special thanks to the GEM Consortium and LSAMP Bridge to Doctorate Program. They have greatly assisted in funding and support during my graduate education. I would also like to thank my committee members Dr. William Rand, Dr. Michael Fu, and my research advisor Dr. Jeffrey Herrmann.

Table of Contents

List of Tables

List of Figures

Chapter 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. The purpose of this research is to find patterns and learn about user behavior to help emergency managers craft appropriate tweets and send messages through the correct 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 enlighten researchers of information diffusion. User connections and content analysis shows who (other users or people mentioned in tweets) are important to users. One focus of this research is to use retweet chains to uncover patterns of information diffusion within various data sets. This research attempts to understand characteristics of information diffusion specifically within social media. The tools used to collect all of this data are Twitter (Twitter, 2014) and TwEater (Monner, 2013). The applications used to analyze the data were NodeXL, MATLAB, and Microsoft Excel.

1.2 Twitter Data

Social media provides easy access 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 is a social network that allows users to share messages or tweets instantly publically or privately with other users. Twitter is not unique to the

United States. Twitter has 241 million monthly active users and an average of 500 million Tweets are sent every day. 76% of Twitter active users tweet from their mobile device and this application supports over 35 languages. (Twitter, 2014)

Twitter data was collected and analyzed from the Hurricane Irene, Hurricane Sandy, Osama Bin Laden's capture, and the United States' 2012 Presidential Election. The data is collected with the permission of Twitter's Application Programming Interface (API). The Irene, Sandy, and Election data sets consists of active users tweeting about a specific topic from a within a larger network of fifteen thousand users. The Osama Bin Laden data set was collected based on the active users in the network but over a period time rather than by topic.

1.3 Data Dictionary

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 several publically available resources 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 (Twitter, 2014)

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

User: The owner of a Twitter account with the ability to use the account

Active User: A user that sent at least one tweet recorded in the data set

High Frequency: A user that posts more than 20 tweets and/or retweets within the data set

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

Single Chain: A set of two tweets in which one in the unique tweet and one is the only retweet

Data Set: Data provided by Twitter about the large scale events.

Stranger: A user from outside the data set

Neighbor: A user from within the data set

Neighborhood: All data and users within the dataset

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 (Murthy, 2013)

Opinion Leader: An influential user that shares his or her social, political, worldwide, emergency, or important events view to his or her network (Murthy, 2013)

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

1.4 Organization

The following section will show the initial and continuing investigation of information diffusion and its relationship with social media. The literary review will show how this research coincides with previous research efforts as documented in books and published papers.. Then, the methodology with be explained. The tools and process from the data collection and analysis will be revealed in detail. Subsequently, the results section will show statistics and charts from each data set. The characteristics discovered will be further discussed. Finally, a summary of this research will provide the limitations, future plans, and new questions posed by this research.

Chapter 2: Literature Review

2.1 Overview

Capturing statistics about information diffusion was once limited to tracking information spread in-person or phone conversation, or mail. There is great difficulty in tracking this type of diffusion, and there are many associated population assumptions. The concept of an online social network is more recent. Social media platforms provide an easier way to gather data about the diffusion of messages travelling through social networks. Now that social media messages are readily available, a proper model must be used to capture the characteristics of the population. There is a history of various rumor spread models that will be explored to enhance our understanding of the diffusion process. As information is diffused, how is it verified to be truthful? There are several studies that investigate rumors spread throughout social networks. There are various ways to evaluate the characteristics of data mined from Twitter. Investigating a social network provides insight about user connectivity. This connectivity can be graphed or visualized to represent the social network. The influence produced by either content or users is demonstrated by a rising popularity of selected content, showing what users value. Content and sentiment analysis can capture the connotation of tweets passed along a network. In addition, some researchers have used a combination of these techniques to create algorithms to predict the popularity of various tweets, topics, and users. Social media is used in many ways by everyday people, government, businesses, politicians, musicians, and even activists. This chapter will review previous efforts to analyze information diffusion and discuss how the research described in this thesis extends our knowledge of this phenomenon.

2.2 Models

A model that is often referred to when discussing diffusion is the Bass Model. This model was originally meant to match the consumer adoption process. This shows the growth pattern and produces the probability of a purchase based on a linear function of prior purchases. The equation for this model is: $\frac{F'(t)}{1-F(t)} = p + qF(t)\text{and } F(0) = 0$. Where *p* is the innovation or advertising coefficient and *q* is the imitation coefficient, *F'(t)* is the rate of change of the installed base fraction and *F(t)* is the installed base fraction. There are five possible stages of the adoption process that a person can be in: innovator, early adopter, early majority, late majority, or laggard. This model helps to predict the probability that purchase will be made at time *t* based on the purchases made by the population before time *t* (Bass, 1969). This model had limitations as it was unable to account for marketing mix variables (Radas, 2005). The marketing mix variables are product, price, place, and promotion (BusinessDictionary, 2014). These variables are modified to find the right mixture keeping customers satisfied and bringing in optimal profit.

Early information diffusion models were based of models on disease epidemics. These models were modified for information spread throughout a population. Daley and Kendal (1964) introduced a model that represented a constant closed population through which a rumor spread. In the D-K model, there are three classes of people:

- the Ignorants do not know about the rumor,

- the Spreaders have heard the rumor and choose to actively spread it, and
- the Stiflers have heard the rumor but do not propagate the information.

The transitions or interactions between these classes are defined by pair-wise contacts. There is an assumption that a Spreader that comes in contact with another member of the population tells this person the rumor. If the receiver is an Ignorant, they become a Spreader; however, if the receiver is a Spreader or a Stifler, both classes are now “discouraged” from further passing along the rumor. The initial assumptions of this population are there is one initial Spreader and the entire population outside of that individual is an Ignorant. At any time *t*, the total population remains the same regardless of the transitions occurring. Let *X(t)*, *Y(t)*, and *Z(t)* bet the number of Ignorants, Spreaders, and Stiflers in the population at time t. The transitions and assumptions are presented in Tables 2.1 and 2.2 (Deitz, 1967).

Table 2.1 The initial conditions in the Daley Kendall Model (Dietz, 1967)

Class	Initial Conditions
Ignorant	X(0) = N
Spreader	Y (0) = 1
Stifler	Z(0) = 0
Total Population	X(t)+Y (t)+Z(t) = N +1

Table 2.2 The dynamics of the Daley Kendal Model (Dietz, 1967)

	Transitions	Mathematical Representation	At a rate proportional to:
1	Infections (X,Y,Z)	(X-1, Y+1, Z)	XY
2	Removals (X,Y,Z)	(X, Y-1, Z+1)	YZ
3	Spreader removal (X,Y,Z)	(X, Y-2, Z+1)	$\frac{1}{2}$Y(Y-1)
4	Spreader removal (X,Y,Z)	(X,Y,Z)	-Y(X+$\frac{1}{2}$(Y-1)+Z)

In line 1, an Ignorant becomes infected when they come in contact with a Spreader. In line 2, when a Stifler comes in contact with a Spreader, the Spreader becomes a Stifler. Line 3 and 4 are two ways that the DK model is different from previous models. In line 3, an active Spreader is removed when he or she interacts with some who has heard the rumor before. If two Spreaders meet one becomes a Stifler. In line 4, if a Spreader meets a Stifler, he too becomes a Stifler.

Maki and Thompson (1973) further examined the D-K rumor spread model and kept many similar features. The M-K model assumes constant population and similar types of interactions; however, there is one important change in the outcome of a meeting between a Spreader and Stifler. Also, the interactions are a result of directed contacts of Spreaders to the rest of the population rather than pair-wise. When a Spreader meets a Stifler and attempts to pass on a rumor, the Spreader will lose interest in spreading the rumor after realizing that the Stifler already knows about the rumor. Thus, the Spreader becomes a Stifler.

Hayes (2005) attempted to use both the D-K and M-K models to predict the number of people who would never hear a rumor, or the undisturbed population of Ignorants. His simulations aimed to match the theorized value of 0.203188 or about 20% unknowing members of the population, as shown in Figure 2.1. After much trial and error, he was able to get both models to converge to the same correct value.

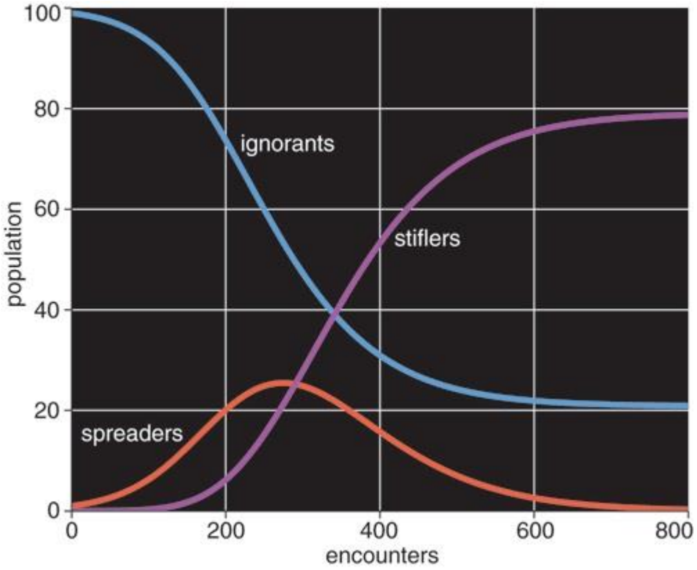


Figure 2.1 The dynamics of the population as encounter increase (Hayes, 2005)

Zhang and Zhang (2009) created a model to demonstrate the interaction of rumor spreading and emergency development, shown in Figure 2.2. This led to the conclusion that a fast spreading rumor can cause panic while a slow rumor may not reach enough people in time. Their results showed that, if the spread rate is appropriate, the public can react and take proper measures for safety.

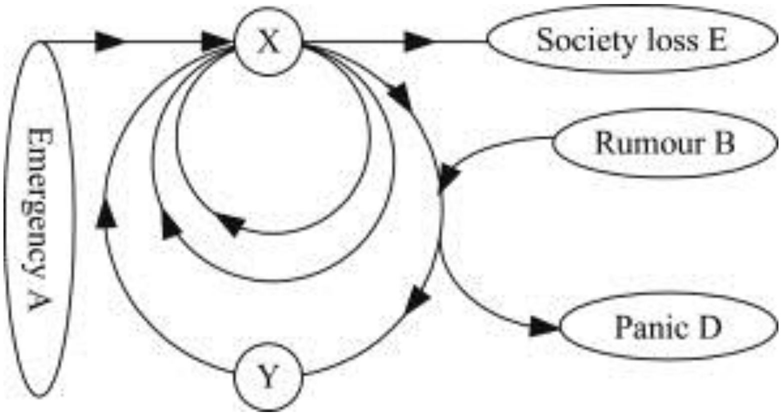


Figure 2.2 The interaction of rumor spreading and emergency development (Zhang and Zhang 2009)

Huo and Guo (2012) included many of these models and studied the dynamics of a system of differential equations. They started with the D-K model and included belief rate, spread rate, reproduction number of the global dynamics, and a varying population. By first understanding the various models of information diffusion, these models can be adjusted for the social networking population and possible behaviors.

Another way to look at modelling information diffusion is using an agent based approach. Herrmann et al. (2013) compared both the Bass and Independent Linear Cascade models to see which best represents the information diffusion of large scale events as conveyed on Twitter. They found that neither had great advantage over the other and both cater to long term events rather than urgent diffusion. They furthered their research by incorporating these as sub-models in their agent based model. By comparing simulated data vs. real data, they found that events with longer timelines fit the model better. They aimed to discover the probability that new information would be adopted and use this information to assist those that manage crisis using social media.

Wang et al. (2013) proposed a linear diffusion model that uses a simple linear partial differential equation to account for influence, population density, and decay of interest in a topic. Social interactions and human dynamics determine information diffusion dynamics. There are new parameters considered in this model: spatial spreading power and temporal news decay. Creating a model to demonstrate the behavior of diffusion is not an easy task. As shown in this section, there are many variables to consider to depict an accurate model.

2.3 Rumor Transmission

The models previously mentioned assume that the information shared is the truth. An important part of studying information diffusion is verifying that the information is accurate. There has been much research about information diffusion and the spread of rumors. Various models have been presented to best represent rumor transmission through a population. Zhao *et al.* (2011) defined a rumor as an unconfirmed elaboration of events, or issues that spread through various channels, in itself neither nor false. In many cases, tweets are rumors until confirmed by verified sources. Doerr et al. (2012) simulated rumor spreading on a large scale social network and used a basic push-pull model to understand the network communication; however, this model did not include forwarding or retweeting.

As a rumor spreads throughout a network, there are many behaviors to consider. Once people receive the information they pass it on, ignore it, hibernate (Zhao et al., 2012) or possibly forget about the information all together (Zhao et al., 2011). Hibernators can extend the life of a rumor and change the spread process. A rumor spread through social networks can fuel a social epidemic. Berger (2013) defined a social epidemic as instances where products, ideas, and behaviors diffuse through a population. Once information is spread, what metric can verify belief of that information? Sanmitra et al. (2012) defined degree of belief as a fraction calculated by dividing the number of relevant supporting tweets by the total number of relevant tweets. Belief is an important concept when tackling emergency management. Social networks allow information to spread quickly in the event of a weather emergency, giving warnings to those capable of receiving that information. To properly assist a community this question by Raghusvanshi (2013) needs to be answered: When is a weather forecast confirmed as the truth and when do people react?

2.4 Understanding Social Networks

A social network involves many connections within a large population. Proper visualization helps to illustrate these connections. In my research, nodes represent each user involved in a connection and the arcs (lines between each node) are retweets that connect users. The use of layered visualization shows graph decomposition for networks of high complexity. Abello and Queyroi (2013) studied the mathematical properties of complex networks and gave values to the nodes and connections of the networks.

Easley and Kleinberg (2010) suggested that there are also social, economic, and natural processes that need to be examined for an accurate and realistic understand of how networks work. They described two theories in this book: graph theory and game theory. Graph theory explores the strength of ties and the structural balance of social networks. Game theory focuses on behavior by looking at the actions and decisions made by a group of people.

Network structure can play a huge role in how information is passed from one node to another. In emergency situations, the structure can effect emergency procedure. Hui *et al.* (2010) studied how 5 different types of network structures (grid, regular, random, scale-free, and group) diffused information. These researchers also examined different levels of trust between nodes. The more trust there is, the greater the weight of the connection is thus facilitating information diffusion.

The right connections within a network can have many positive residual effects, especially for businesses. Gupta et al. (2013) worked in conjunction with Twitter to create a targeted marketing algorithm that suggests "Who to Follow." In summary, each user is assigned a value (all users combined equal 1). A similarity score is found for each consumer and relevance score is obtained for each producer from the consumer-producer graph connections. The highest valued users based on connections are recommended to consumers. This is important for businesses that want to connect with consumers, but not at random. This type of algorithm can be used to recommend followers to one another based on their activity and network connections.

2.5 Influence

Once a network is analyzed, determining influence among these connections can better characterize user behavior. For this research influence is the ability of a user, topic, or actions of a user to cause action in other users. It is important to understand who and what motivates the diffusion of information and why it is influential Influencers are defined as "someone who exhibits some combination of desirable attributes whether personal attributes like credibility, expertise, or enthusiasm, or network attributes such as connectivity or centrality-tat allows them to influence a disproportionately large number of others" (Gladwell, 2000). Stickiness is another important term; it is a "unique quality that compels the phenomenon to stick in the minds of the public and influence their future behavior" (Gladwell, 2000).

Many ways to measure influence have been developed. Cha *et al.* (2010) found that popularity alone does not define the influence of Twitter users. A user can earn influence by focusing on one topic rather than engaging in conversation.

Anger and Kittl (2011) used the term "alpha users" to describe influencers. They introduced the Follower/Following Ratio (r_f), Retweet and Mention Ratio (r_{RT}), and the Interactor Ratio (r_i). According to Anger and Kittl, when there is an attempt to influence a user with a tweet, users react three possible ways: Compliance, Identification, or Internalization. Compliance is public agreement, Identification is an attempt to interact because of user status, and Internalization is accepting a belief or behavior, publically and privately (e.g., retweets). When a user is not influenced, the possible reactions are Neglect (ignore) or Disagreement (comment and possible unfollow). This study was limited to Austria's top 10 users as ranked by influence using an application called "Klout".

Ye and Wu (2010) classified three different influence metrics: Follower Influence, Reply influence, and Retweet Influence. Each influence is characterized by the action associated with the metric such as receiving the message or being a follower, replying, and retweeting, respectively. Bakshy et al. (2011) tracked and computed influence based on URL clicks or visits. They have three different ways to assign credit of influence to multiple sources.

The Linear Influence Model considered global influence of individuals on the diffusion rate throughout the network. This research was different than others, in assuming there was no prior knowledge about the network. By combing through tweets and blogs for a one year period, researchers focused on 10,000 Twitter users. Memetracker methodology and hashtag focused data collection was completed to gather various forms of data (Yang & Leskovec, 2010). They found that users with about 1000 followers are the most effective in diffusion and adoption of hashtags.

Wang *et al*. (2013) separated the influences coming from both distance and time in search of an inherent news decay pattern. Their linear diffusive model was validated using “Digg” social network. They found that model captures essential factors shaping information diffusion and plan to apply their work to learn about influence on Twitter.

2.6 Content and Sentiment Analysis

The content being diffused it just as important as the user passing along the message. This research focuses on large scale events. Bakshy *et al.* (2011) selected tweets with content about Michael Jackson and started their collection two days after his death. This research found that the most popular tweets were short phrases, slang, and automatically generated messages such as spam. Additionally, Murthy (2013) stated that Twitter does not reflect normal conversation. Hashtags give users an opportunity to include what they are thinking as well as what they are saying in their main message. Self-promotion is an opportunity for a user’s tweet to be seen by a larger audience by including a trending hashtag.

Asur et al. (2011) conducted a study on trending topics (the most popular hashtags at any given time) to study the distribution and decay rate of these trending topics. Prolonged trends stemmed from traditional media sources such as news agencies and were propagated by chains of users. Romero et al. (2011) focused a study on content, specifically finding out what makes certain topics spread at different rates. They analyzed about 500 hashtags from over 3 billion tweets and defined "persistence" as the "relative extent to which repeated exposures to a piece of information continue to have significant marginal effects on its adoption" (Romero et al., 2011).

Content analysis opened the door for a deeper look at textual analysis. Sentiment Analysis seeks to understand the meaning behind text. It is hard to recognize emotions such as sarcasm or disdain in text. Wu et al. (2013) created a tool called a Recursive Neural Tensor Network to assist in sentiment analysis. This tool takes phrases input by users and gives sentiment to each word. Sentiment is based on their word bank and the past sentiment assignments, ranging from very negative, negative, neutral, positive, and very positive. The word bank tool takes feedback from the users who know the actual sentiment of their phrase to improve the accuracy of this tool. A more mature version of this tool could prove helpful in the future of this research.

2.7 Prediction

The content, sentiment, and influence can be visualized as inputs to the prediction of information diffusion. By creating conversations on Twitter, marketers can estimate the popularity of an item with consumers. Asur and Huberman (2010) defined several predictive measurements: Tweet Rate, Subjectivity, and Polarity.

Tweet Rate is the specific tweet count divided by the hours of the tweet collection. Subjectivity is the absolute value of the positive and negative tweets divided by the neutral tweets. The polarity is defined by the PNratio which is the tweets with positive sentiment divided by the tweets with negative sentiment.

Research by Liangjie et al. (2011) found several key conclusions about predicting tweet popularity. Twitter users are more likely to pass on retweets from their "first level" friends; naturally, if a user has scores of followers, they will see more retweets. Users with a restricted amount of followers will not have as many retweets. This method worked well for 2 types of posts: tweets and retweets with a volume of more than 10,000.

Other research has focused on various properties of diffusion like speed, scale, and range, to predict diffusion patterns (Yang and Counts, 2010). Focusing on the active interaction network of users, rather than the follower network, is theorized to be a stronger network representation (Yang and Counts, 2010). We complete our network graphs in this way as well.

The lifetime of a tweet can be viewed as a period of relevance. Links shared in social media are considered to have half-lives. Half-life is a term used in chemistry, but for Bitly (2011) it means when a link has received half the clicks or visit, it will ever receive in it's lifetime. Bitly (2011) concluded that the content of a link determines its half-life. Over time, if the source of the link constantly retweets the same tweet, it will extend the half-life of the link (Sullivan, 2011).

2.8 Uses of Social Media and Social Networks

Information diffusion is studied in various fields including marketing. From a marketing perspective, getting clicks or visits is the beginning of the customer engagement process (Smith 2011). Businesses intend to build customer relationships through social media and utilize metrics such as number of advocates, number of comments, and sentiment (Murdough 2009). Companies want to understand how to best use social media to their advantage and connect with their consumers. Rand et al. (2013) studied consumer engagement on Twitter for various bands. This research focused on the content shared by the bands as well as their reputation and compared that to sales. They observed consumer behavior at various levels of engagement. This information was used to measure brand establishment for various bands, linking revenue to engagement.

2.9 Summary

Many models have been established and further modified as diffusion research has matured. There are several methods to investigate the characteristics of data mined from Twitter. The study of connectivity and influence conveys the popularity of users and content. Furthermore, content and sentiment analysis techniques are used to create algorithms to predict the regard of various tweets, topics, and users. More research needs to be completed about comparing how information spreads in different types of large scale events. Also, there are many different conclusions on measuring influence associated with many different styles of analysis.

Two independent established networks of users need to be compared to find patterns of diffusion and to utilize analysis techniques. Analyzing two controlled population and completing identical analysis will validate assumptions about different types of events.

This research holds many similarities to studies previously conducted. This research builds on previous work by focusing on a topic-based data set of a known network of users. Although this dataset is smaller, there are many connections within the dataset, which alleviates the need for data collection. The data was collected before, during, and after these major events. The networks are represented by active connections such as retweets, but do not consider time in these visual representations. The overall retweet count is utilized and compared to the number of retweets found within the data set. Similar statistics such as tweet rate are found, but there is not a focus on sentiment. There is focus on the users who tweet very often and those that a retweeted most often in an attempt to understand the very active and popular users. This data has the unique perspective of focusing on emergency weather situations, as well as a nationwide event where every adult American had the opportunity to share their opinion.

This research focused on retweet chains rather than overall population behavior. The data collected is for various types of events rather than a singular focus. Varying behaviors can be analyzed based on event that lead to the data collection. The collections were already topic based, but finding patterns with chains will show a new perspective on diffusion research.

Chapter 3: Methods

3.1 Organization

The purpose of the data collection is to gather real instances of information diffusion about large scale events. Data is collected about two types of events: weather emergencies and political affairs. To facilitate information diffusion, a lot of people need to be passing information or feel compelled to share. Collecting data about leading topics potentially produces more topic-based conversation to study. Using Twitter to collect data gives a detailed view of what people are discussing as well as multiple options to analyze information diffusion.

This chapter describes the data collection process (Section 3.1), the data set collected (Section 3.2), the analysis procedures (Section 3.3), and a glossary of key terms (Section 3.4). The goals of the data analysis are to track the trajectory of an initial tweet through the data set and further understand information diffusion occurring on social media.

3.2 Data Collection

Dr. William Rand 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 (Monner, 2013) 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 was selected based on their activity and connections. When a user tweeted about a topic, their tweet was collected by TwEater (Monner, 2013) and placed in a dataset. For example, a tweet with keywords "Obama" and "election" would be gathered by TwEater because those words match predetermined 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 is available in Rand et al. (2013) and Herrmann et al. (2013).

3.3 Data Sets

As Twitter evolves, so does the information that it provided to research organizations. Not all datasets have the same fields of data available. The following are examples from each data set demonstrating the similarities and differences. The data is organized into Excel files and each line contains important characteristics such as the Date/time, Tweet ID, User ID, Tweet Text, Retweet ID, Tweet Status, and possibly location data. A given data set contains from 5,000 to 30,000 lines to analyze. Specifically, the data sets have the total number of tweets: Irene, 5948; Election, 9167; Sandy, 19085; Bin Laden, 27924. The following are the definitions of the headers of the data collected:

Row Added At/Time – The date and time that the tweet was sent.

Status ID/Tweet ID - Each Tweet has a unique ID assigned to it, so even if it is retweeted a new ID will be assigned.

Retweet ID – If the Tweet is original the value in this column is -1, if a tweet is a retweet then the retweet ID will match a tweet id that was previously output.

Status Text/Tweet Text- The actual message posted by a user.

User ID – Each user is assigned a unique user number. The user name can change, but the user id will not change.

User name – The name that a user chooses for their account.

User Status Count – How many status the users has created

Join Date – The date that a user created an account in UNIX time.

Status Retweet Count/Retweet Count - Total number of times the tweet has been retweeted throughout Twitter (not network specific)

RT Status/Status is Retweet – The value is either 0 or 1: if the tweet is a retweet (0) or if the tweet is not a retweet (1).

Status Retweet of- The values is either -1 or a Tweet ID: if the tweet is not a retweet (-1) or if the value is a retweet (the original tweet ID).

User Followers– The number of users following a user

User Friends- The number of people the user follows.

User Listed- The number of times the user has been listed by another user

User Verified - Whether or not the user has been verified by Twitter

User Location – Self Reported by users via text.

User Lang – The Self-Reported language used by each user.

Latitude & Longitude – Coordinates of location.

User ETC Offset - The time difference from Universal Time.

User Info from status – Unique Id also assigned to each tweet.

Status Date- The date and time that the tweet was sent in UNIX time.

The following tables will assist in understanding how data looked in Excel form and what type of information was collected in each set. Table 3.1 shows the headers and an example for each data set. Table 3.2 lists some tweets from the Hurricane Sandy dataset. This sample will be referred to throughout this chapter.

Table 3.1 Examples of differences in data provided in each data set

	Bin Laden	**Irene**	**Sandy**	**Election**
row_added_at	5/1/2011 22:27	8/26/2011 10:00	10/25/2012 23:57	11/6/2012 0:02
status_text	WOW RT @keithurbahn: So I'm told by a reputable person they have killed Osama Bin Laden. Hot damn.	@MySears enjoy the weekend but be careful on the hurricane. Be safe! Ps loved the Sears hurricane email great idea	Apparently weatherbug hasn't gotten word that #Frankenstorm is on it's way #weather #hurricane #perfectstorm http://t.co/nWNfuSpA	@metumbleweed\ Anyone\ with\ LOVE\ OF\ AMERICA\ will\ vote\ AGAINST\ OBAMA.
status_id	64878655444234200	107090067004211000	261677875990781000	265680573538893000
user_id	14944471	18231339	14706004	66951419
status_is_retweet	0	0	0	0
status_retweet_of	-1	-1	-1	-1
status_retweet_count	0	0	0	0
status_latitude	0	0		0
status_longitude	0	0		0
status_date	1304303251000		1351223842000	1352178159000
user_name	GLB62			
user_status_count	15241			
user_followers	1133			
user_friends	1044			
user_listed	42			
user_join_date	1212071733000			
user_verified	0			
user_lang	en			
user_location	Bristol, CT			
user_utc_offset	-18000			

Table 3.2 Tweets from the Hurricane Sandy dataset

Line	*time*	*tweet _id*	*user _id*	*tweet_text*	*Status is retweet*	*Status retweet of*	*Status retweet count*
1	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
2	10/27 16:19	262242320345350000	20387183 5	RT @TheNextWeb: Google's big Monday Android event is off on account of the hurricane http://t.co/FjMSOzGY by @alex	1	262242329132421000	25
3	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
4	11/2 11:58	264338926603476000	22256515	RT @ATT: Our trucks hit the streets of CT to get service up and running. #Sandy http://t.co/5FGcegWG	1	264138150207692000	15
5	11/2 19:03	264466516395843000	46958309	RT @TheNextWeb: Twitter mobile usage in NYC doubled during Hurricane Sandy's peak http://t.co/P1U1XJRX by @harrisonweber	1	264466402650492000	50
6	11/2 19:05	264466989190365000	21703120	RT @DFW_SocialMedia: Twitter mobile usage in NYC doubled during Hurricane Sandyâ€™s peak http://t.co/P323edjz	0	-1	0
7	11/4 5:49	265042793691488000	116706360	RT@Yo_ItsSpongebob: Patrick=Laziness, Mr. Krabs=Greed, Squidward=Anger, Sandy=Pride, Gary=Gluttony, Spongebob=Lust. #7SinsInSpongebob	1	263613379950825000	223
8	11/4 11:08	265083188655816000	15233366	Please Help http://t.co/K1Mr8X7B #sandy	0	-1	0
9	11/4 11:10	265083333288005000	53941287	â€œRT @MarinkaNYC: Please Help http://t.co/xqaoyZox #sandyâ€ •	0	-1	0
10	11/4 11:28	265086985306271000	23866339	RT @MarinkaNYC: Please Help http://t.co/K1Mr8X7B #sandy	1	265083188655816000	3
11	11/11 11:45	267669325098393000	14218186	Think about giving to the @RedCross to help #Hurricane & #Storm victims on the east coast http://t.co/7Ljz1s0G	0	-1	0
12	11/12 11:52	268033428073639000	21460496	RT @USATODAYlife: Has comedy helped you deal with #Sandy? Send us a joke! (We'll RT our faves) http://t.co/mYZcbGdF	1	268025382282874000	5

3.4 Data Analysis

3.4.1 Summary Statistics

Basic statistics were found to initially 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 within the dataset as well as the overall tweet count. The following definitions explain the basic statistics calculated for every data set.

- Time Period of Collection: the difference between the day and time of the first tweet in the dataset and the day and time of the last tweet in the dataset. This is measured in seconds.
- Total Number of Tweets: Count of all Tweet within each dataset.
- Active Users: the number of unique User IDs in the dataset.
- Average Tweets per User: the total number of tweets divided by the number of Active Users.
- Average Number of Tweets per high Frequency User: the total number of tweets by Active Users who tweeted over 20 times divided by the number of such users.
- Number of Users to Tweet Once: Number of Users who tweeted only once in a given data set.
- Percentage of Population that only tweeted once: This quantity is the ratio of Users to Tweet Once divided by the number of Active Users.

- Percentage of Single Tweets: This quantity is the ratio of Users to Tweet Once divided by the Total Number of Tweets.
- Number of Users to Five or Fewer Times: Number of Users in who tweeted five or fewer times in a given data set.
- Percentage of Population that only Tweeted Five or Fewer Times: This quantity is the ratio of Users to Tweet Five or fewer divided by the Total Active Users.
- Percentage of Five or Fewer Tweets: This quantity is the ratio of Users to Tweet Five or Fewer times divided by the Total Number of Tweets.
- Number of Users to Tweet Twenty or More: Number of Users who tweeted more than twenty times in a given data set.
- Percentage of Population that Tweeted twenty or More Times: This quantity is the ratio of Users to Tweet Twenty or more times divided by the Total Active Users.
- Percentage of Twenty or More Tweets: This quantity is the ratio of Users to Tweet Twenty or more divided by the Total Number of Tweets.
- Total Tweets (Subset): Number of tweets by a subset of users.
- Total Retweets (Subset): Number of retweets by a subset of users.
- Average Tweets Per User (Subset):This quantity equals the number of tweets divided by the number of users

- Percentage of Tweets that are Retweets: This quantity equals the number of retweets by a user divided by the number of all tweets the user sent.
- Average Time between tweets: the time period of collection divided by the total number of tweets.
- Distribution: the distribution graph including each user and how many tweets they sent within a given data set.

3.4.2 Retweet Chains

A retweet chain is a set of tweets in which the unique tweet and its retweets can be identified and listed together. In the example presented in Table 3.2, lines 1, 2, and 3 form a retweet chain because these tweets are all retweets of the same original tweet. 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 chains 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. Irene and Election datasets were analyzed using MATLAB. However, after the analysis of these two data sets, we found that Microsoft Excel functions were much easier to organize and output the data, and the analysis of the Sandy and Bin Laden datasets was completed with Microsoft Excel.

To learn about groups of users, tweets of the Originators and Retweeters were explored by finding the tweets and retweets of each user. The difference between a tweet and a retweet is evident if "RT" exists before a tweet and Tweet Status is 1 (retweet) or -1 (tweet). In the Table 3.2, lines 2 and 3 are retweets of tweet 1. As previously mentioned, chains can originate from outside the dataset, meaning the Originator is a stranger, but the Retweeter is a neighbor. In Table 3.2, line 12 is part of a chain that was started outside of the dataset. The originator USATODAYlife is not a user within the dataset. Just like chains within the dataset, these chains are analyzed in the same manner. This type of analysis is completed for every dataset. Once all statistics and calculations are complete, a search for patterns between datasets was pursued. These statistics were previously defined in Section 3.3.1. Any patterns are analyzed by tweet content, originator, time, and volume.

To understand the neighbor connectivity each dataset was visualized based on retweet chains. These chains are based on tweets that originate from within the network as well as those that start outside the dataset. Using NodeXL, the User IDs of those involved in retweet chains were copied from the data set chains and pasted as vertices and edges. The nodes represented by shapes are active users in the neighborhood. The active users are also identified on the network by the listed User ID numbers or written text. An edge is a retweet represented by a line connecting the Originator and Retweeter. Colored and numbered edges illustrate users involved in the same retweet chain.

For the retweet chains, the following statistics were determined.

- Number of Chains Inside the Data Set: The total number of Chains that originated inside the data set
- Number of Originators: The total number of unique users that start chains inside the data set

- Average Number of Tweets per originator: This quantity equals the ratio of sum of all tweets by originators divided by the number of total originators
- Number of Retweeters: The total number of users that retweeted a tweet from an originator
- Average Number of Tweets per Retweeters: This quantity equals the rather of sum of all tweets by Retweeters (involved in chains) divided by the number of total Retweeters
- Average Time to Retweet: This quantity equals the sum of time between the original tweet and each retweet divided by the total number of retweets in each chain.

3.4.3 Content Analysis

To analyze the content of the tweets, the popularity of hashtag (#) use was analyzed. Hashtags are (#) followed by a word or phrase with no spaces. In Table 3.2, hashtags are used in tweets on lines 4, 8, 9, 10, 11, and 12. Tweets with the same hashtag are searchable on Twitter. Thus, people talking about the same thing and using the same hashtag can see who else used the same hashtag or any hashtag. Once a hashtag is the most used on Twitter, it is considered a trending topic. This process was completed by taking daily samples for the hurricane datasets and hourly 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 data was charted for the dataset, the retweet chains, and total tweets for the varying time periods. These charts can show popularity of conversation topic in Retweet Chains compared to real time events and total tweet count.

3.5 Glossary

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 several available resources with business or research focused on social media. Some terms are specific to this academic research and defined as such to lessen confusion.

Active User: A user that sent at least one tweet recorded in the data set.

Chain: A set of tweets in which the unique tweet and its retweets can be identified and listed together.

Dataset: Data collected from Twitter about the large scale events.

High Frequency User: A user that posts more than 20 tweets and/or retweets within the data set.

Loyal User: A user who retweets an originator more than once.

Mention: A tweet that names another user with a "@" preceding the user.

Neighbor: A user from within the data set.

Neighborhood: All data and users within the dataset.

Opinion Leader: An influential user who shares his or her social, political, worldwide, emergency, or important events view to his or her network. (Murthy, 2013)

Originator: The first user to post a unique tweet.

Period of Relevance: A time period in which the information provided by a tweet is valuable to users.

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.

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.

Retweeter: A user that reposts a unique tweet.

Single Chain: A set of two tweets in which one in the unique tweet and one is the only retweet.

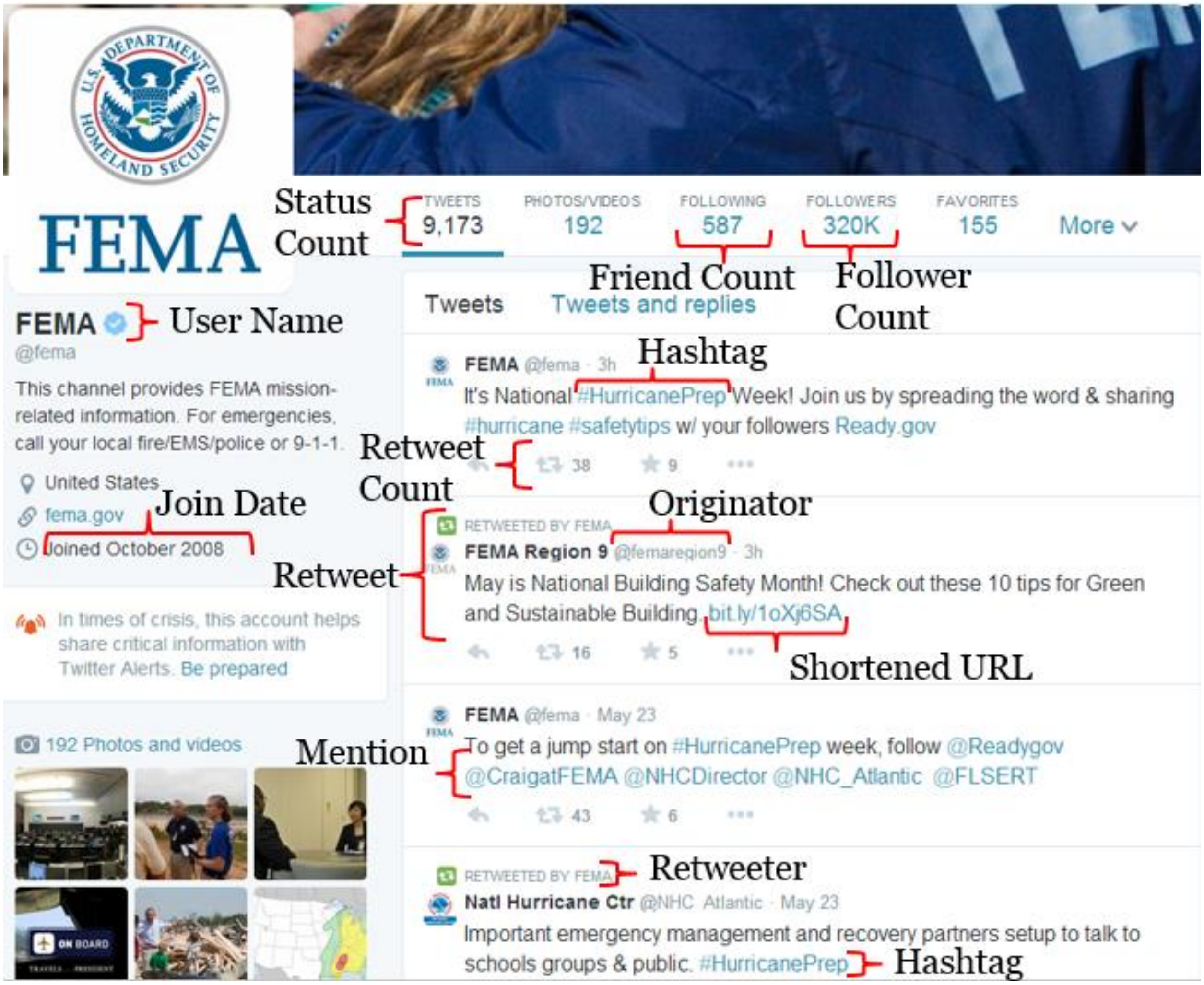
Stranger: A user from outside the data set.

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 (Twitter, 2014).

User: The owner of a Twitter account with the ability to use the account.

Verified Tweets: A tweet that comes from a current professional or unquestionable source. (Murthy, 2013)

Figure 3.1 shows an actual Twitter profile labeled with several definitions mentioned in the glossary. 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

TWEETS 9,173 · PHOTOS/VIDEOS 192 · FOLLOWING 587 · FOLLOWERS 320K · FAVORITES 155 · More

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

192 Photos and videos

Tweets · Tweets and replies

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 3.1 This Twitter profile belongs to the Federal Emergency Management Agency captured May 32rd, 2014. This shows actual tweet activity of the FEMA and also a lot of data defined in sections 3.2 and 3.4 (Twitter, 2014)

Chapter 4: Results

4.1 Organization

This section highlights the results generated by analyzing the datasets using the techniques described in Chapter 3. The results for each data set are presented in its own section. The following results are presented: a summary of basic statistics, retweet chain characteristics, chain networks, and hashtag use. These results give some insights into the activity of the users in each data set. The results will be compared to search for patterns among the data sets. Finally a discussion will be presented analyzing the results of the data sets.

4.2 Irene

4.2.1 General Information

The Irene data was collected over a period of 17 days. There are 5948 tweets and 2210 active users in the Irene neighborhood. Approximately 51% of neighbors tweet only once and of the users that only tweet once, those tweets account for 19% of the data set. As shown in Figure 4.1 only a small number of users tweet more than 5 times. Only 19 of the 2210 users tweeted more than 20 times in the 17 day time period. The average number of tweets per user is 2.69. The high frequency users tweeted an average of 28.63 times. Figure 4.1 shows the frequency that users tweet.

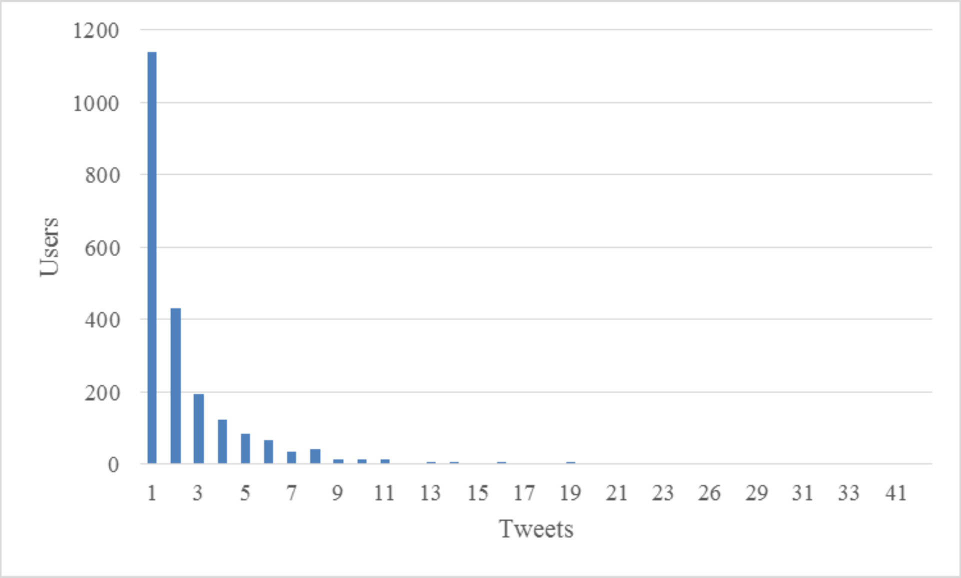


Figure 4. 1 Frequency at which users tweet within the Irene data set

4.2.2 Irene Retweet Chain Characteristics

Within this neighborhood, 23 chains were found with Originators. Thus, there were 23 Originators and 27 retweets and Retweeters. The average number of tweets per Originator and Retweeter are 8 and 7.333 respectively. The average number of retweets per Originator and Retweeter are 1.391 and 3.037 respectively. Finally the percent of tweets that are retweets are 17.39 and 41.414 respectively. Clearly, the Originators and Retweeters on average tweet about the same. The difference is obvious with the Retweeters retweeting over 2.1 times as much as Originators. The ratio is a little more for Retweeters; moreover, their overall tweets contain 2.3 times as many retweets as Originators.

Figure 4.2 shows the time is takes retweets to travel in chains of up to 3 retweets long. This is a visualization to understand the trajectory of these chains. The average time from the original tweet to a retweet is 7 minutes; however, if the chain ends with only one retweet the average time is 38 minutes. Taking out extreme values of retweet times lasting over one hour, the average retweet time is recalculated for chains of single retweets to be 8 minutes. So the average time to retweet for all chains is very similar to that of single chains.

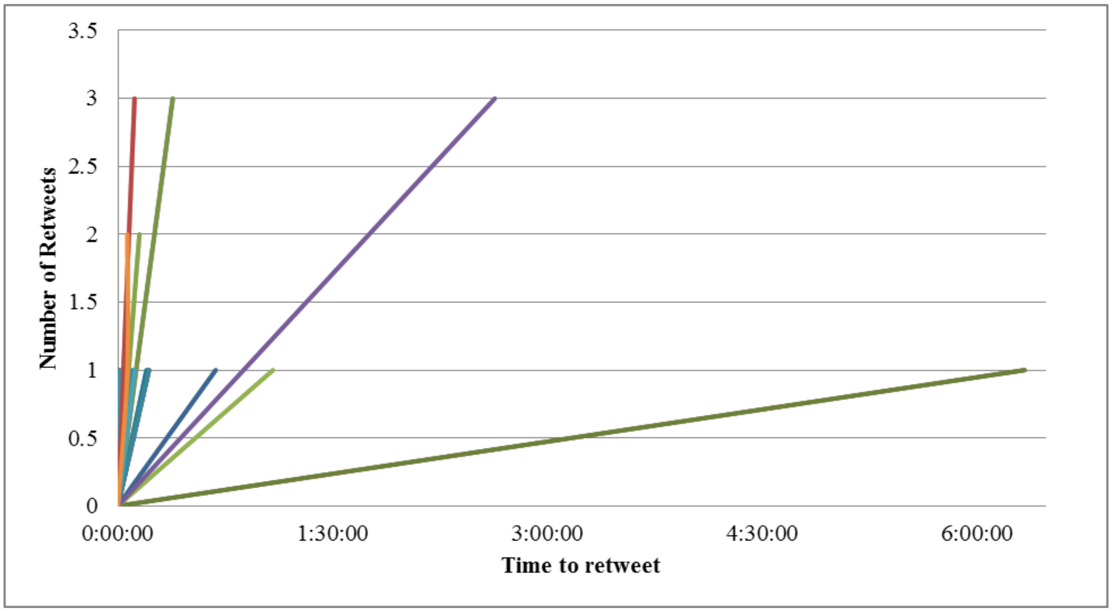


Figure 4.2 Retweet Chains in the Irene data set up to three tweets long

4.2.3 Chain Network

The network visualization Figure 4.3 shows a lack of connectivity between users involved in chains within the Irene network. A lot of the chains involved are just two tweets long. The nodes represented by shapes are active users involved in more than one chain. This network contains 27 connections and contains no duplicate connections or reciprocal relationships.

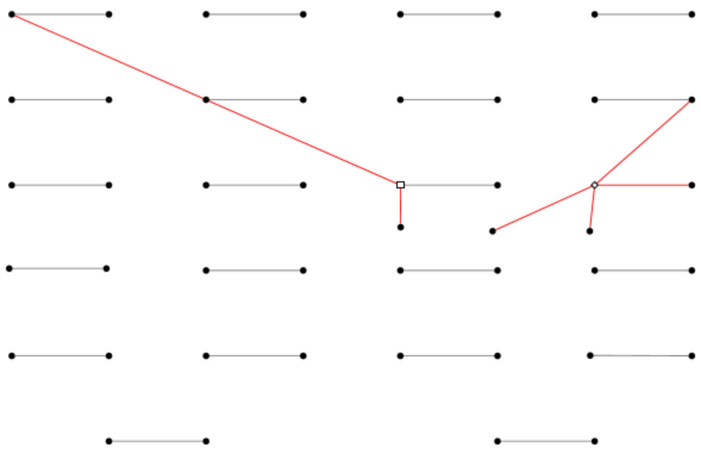


Figure 4.3 The inside chain network for Hurricane Irene

Figure 4.4 shows the network of the chains that were started outside of the data set. There are more connections in this network. Hurricane Irene had a Twitter account created by someone. Many people retweeted Hurricane Irene. These outside chains show that many celebrity tweets made their way into the data set.

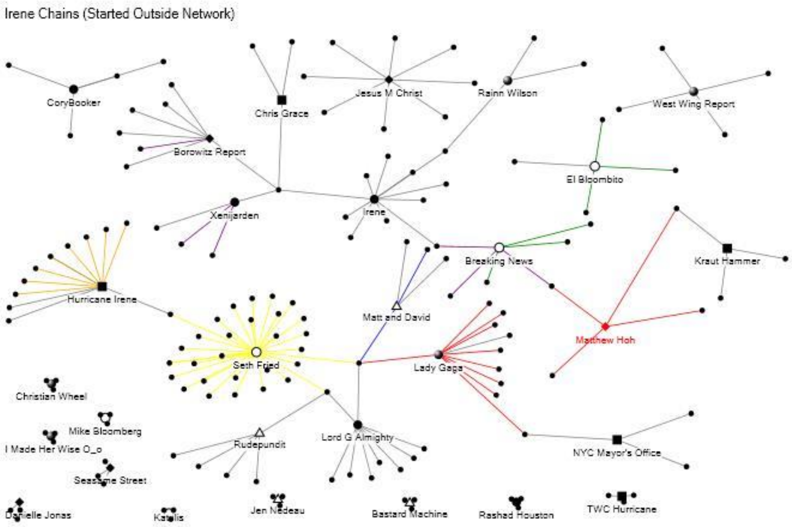


Figure 4.4 The outside chain network for Hurricane Irene

4.2.4 Hashtags

The most popular hashtags used were #Irene and #hurricane. Hashtag use is not very high throughout the data set. Most were used early in the data collection. Figures 4.5 and 4.6 show the frequency of use of these two hashtags by all users as well as just chain users.

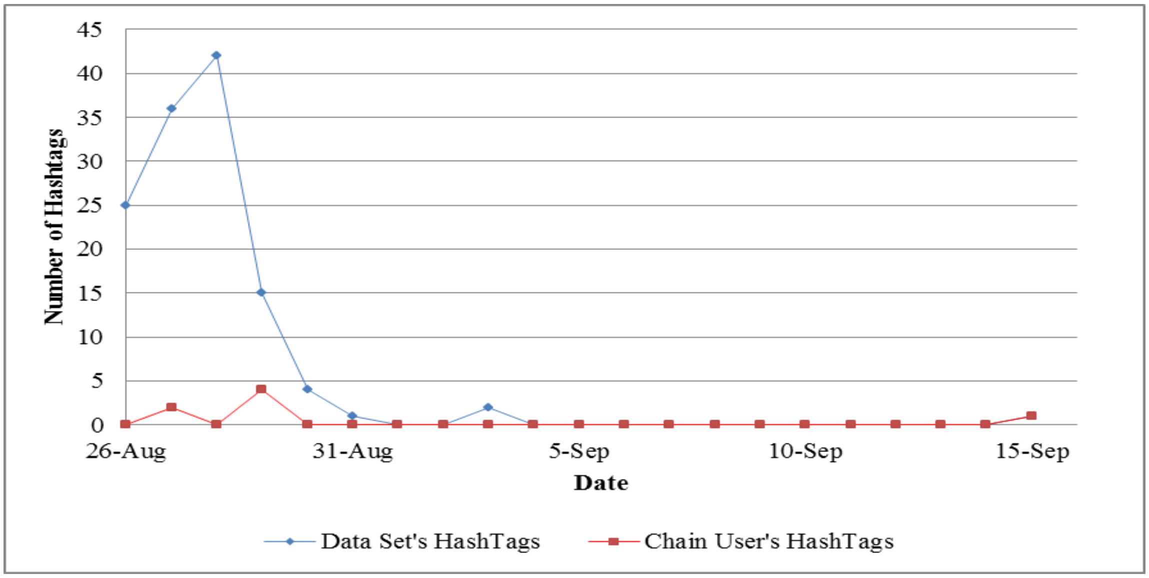


Figure 4.5 The frequency of the use of #Irene by all users and chain users

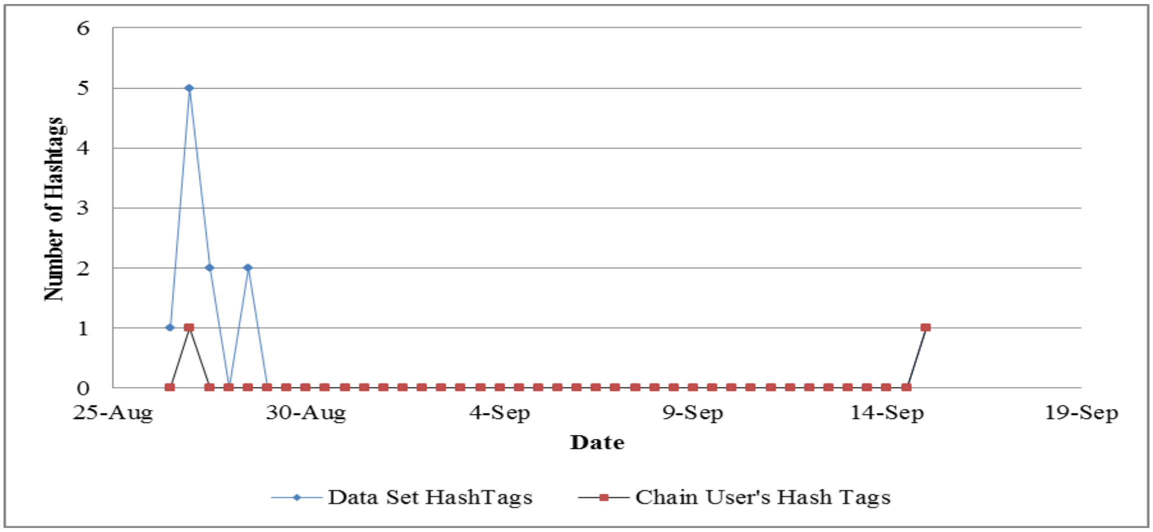


Figure 4.6 The frequency of the use of #Hurricane by all users and chain users

4.3 Election Results

4.3.1 General Election Neighborhood Information

The Election data was collected over a period of 25 hours. There are 9671 Tweets and 2455 active users in the Election neighborhood. Approximately 47% of neighbors tweet only once and 84% of neighbors tweet 5 times or less. Of the users that only tweet once, those tweets account for 12% of the data set. 39% of tweets come from users that tweet 5 times or less. Only 70 of the 2455 users tweeted more than 20 times in the 25 hour time period. The average number of tweets per user is 3.94. The high frequency users tweeted an average of 40.13 times. Figure 4.7 shows the frequency at which users tweet. As previously stated, most users tweet only once.

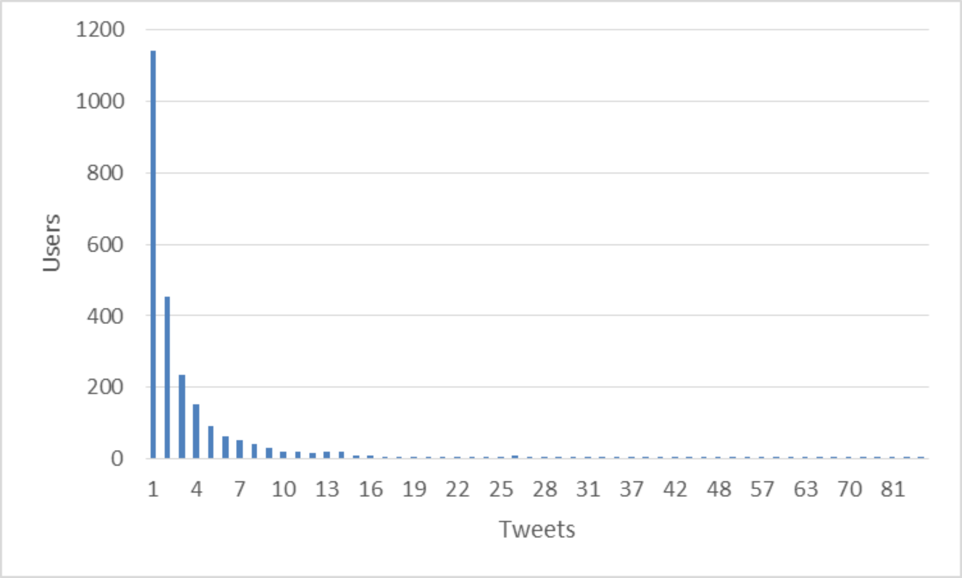


Figure 4.7 Frequency at which users tweet within the Election data set.

4.3.2. General Election Retweet Chain Characteristics

There were 38 Chains that started within this neighborhood. Those chains had 38 Originators and 45 Retweeters. There were a few examples of tweet chains in which a user retweeted an Originator and later those two users were in another chain in which the Originator was the Retweeter and the Retweeter was the Originator. Users involved in retweet Chains were not limited to reciprocal relationships.

The average number of tweets per Originator and Retweeter are 15.21 and 5.55 respectively. The average number of retweets per Originator and Retweeter are 3.23 and 6.2 respectively. Finally the percent of tweets that are retweets are 17.55 and 52.84 respectively. Unlike the Irene data, the Originators and Retweeters on average tweets are very different with the average tweets being much higher. This is because there was a single Originator that skewed the data by retweeting over 200 times.

The average time from the original tweet to a retweet is 27 minutes and 1 seconds; however, if the chain ends with only one retweet the average time is 33 minutes and 50 seconds. Taking out extreme values of retweet times lasting over one hour, the average retweet time is recalculated for chains of single retweets to be 12 minutes and 20 seconds.

4.3.3 Chain Network Visualization

Figure 4.8 shows a lack of connectivity between users with the Election network. A lot of the chains involved are just two tweets long. The nodes represented by shapes are active users involved in more than one chain. This network shows many chains and contains no duplicate connections or reciprocal relationships. There are no users involved in more than one long chain. There is more connectivity in the chains that start outside of the 15k network. Many network users tweeted similar Originators. Also, popular Originators are involved in more than one chain, as shown by the different colors.

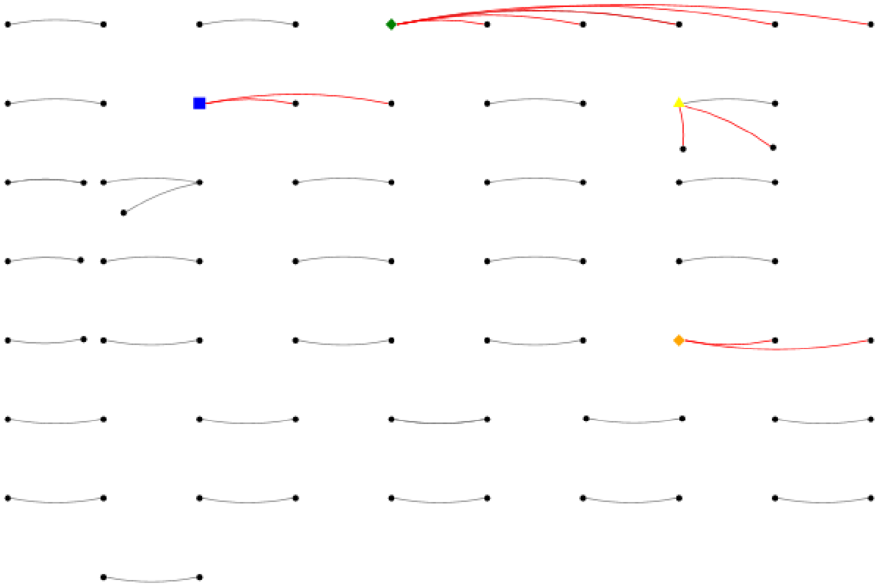


Figure 4.9 The inside chain network for the 2012 Presidential Election

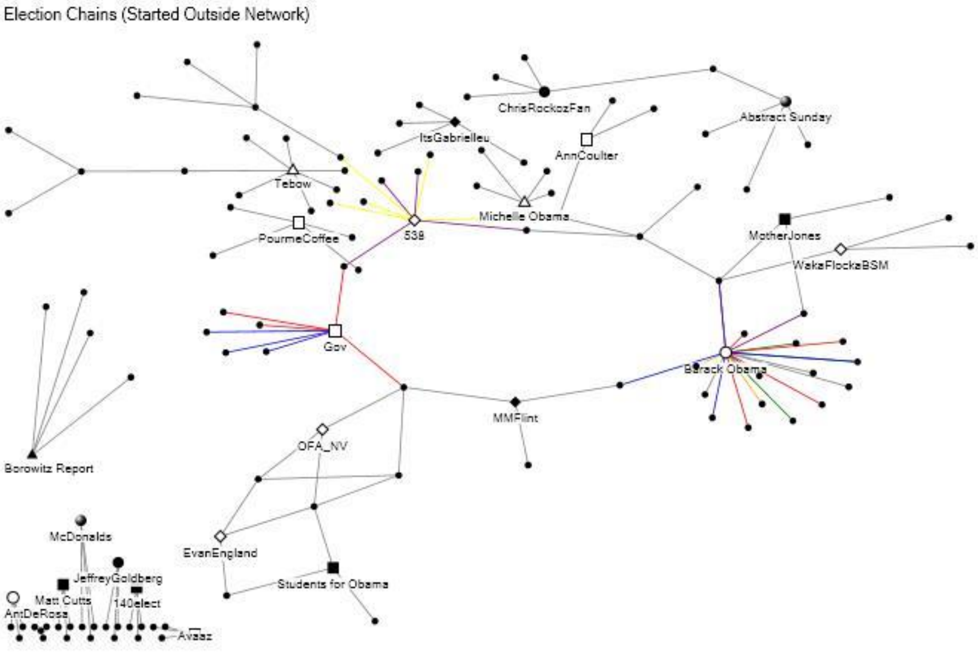


Figure 4.8 The outside chain network for the 2012 Presidential Election

4.3.4 Content

Figures 4.10 and 4.11 illustrate the use of various popular hashtags in the data sets over time. Figure 4.10 shows the overall use of popular hashtags such as #election, #vote, #Obama, #Romney, #iVote, #govote. Figure 4.11 shows the hashtag use referring to various states. #Election and #Vote and #Oh and #Fl are the two most popular hashtags for each graph respectively. If the election was decided by hashtag frequency President Obama would win with 406 hashtags and Romney would once again concede with 320. Of course #election (1862) and #vote (523) would be most popular on Election Day. Ohio and Florida were considered battleground or swing states in the 2012 election. Obama won both of these states by a margin of less than 5% (CNN, 2012). Many users were urging Americans to stay in line and make sure that their vote was counted in those states. #Oh was seen 131 times, and #Fl was seen 79 times.

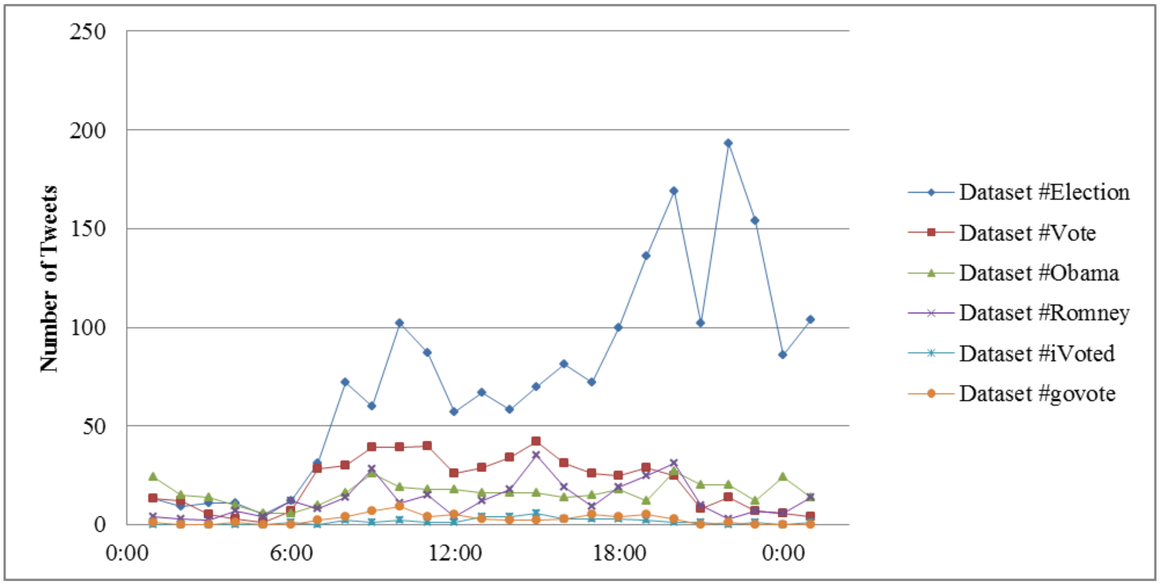


Figure 4.10 The frequency of the use of the most popular hashtags by all users

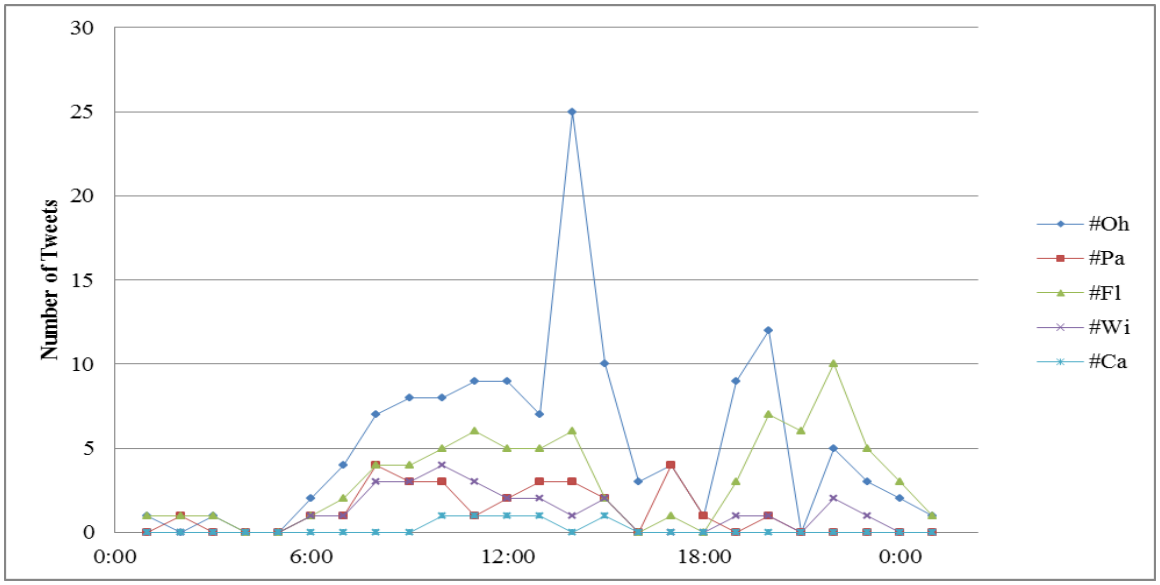


Figure 4.11 The frequency of the use of the most popular state hashtags by all users

4.4 Sandy Results

4.4.1 General Sandy Neighborhood Information

The Sandy data was collected over a period of 17 days. There are 19085 Tweets and 3325 active users in the Irene neighborhood. Approximately 40% of neighbors tweet only and of the users that only tweet once, those tweets account for 7% of the data set. So, 93% of the tweets in the data set come from users that tweet more than once. Figure 4.12 shows how frequently each user tweets. The average number of tweets per user is 5.74. 182 of the 3325 users tweeted more than 20 times in the 17 day time period. These are considered high frequency users. The high frequency users tweeted an average of 47.2 times. These values are about twice as high as the Irene statistics but the time period is about the same.

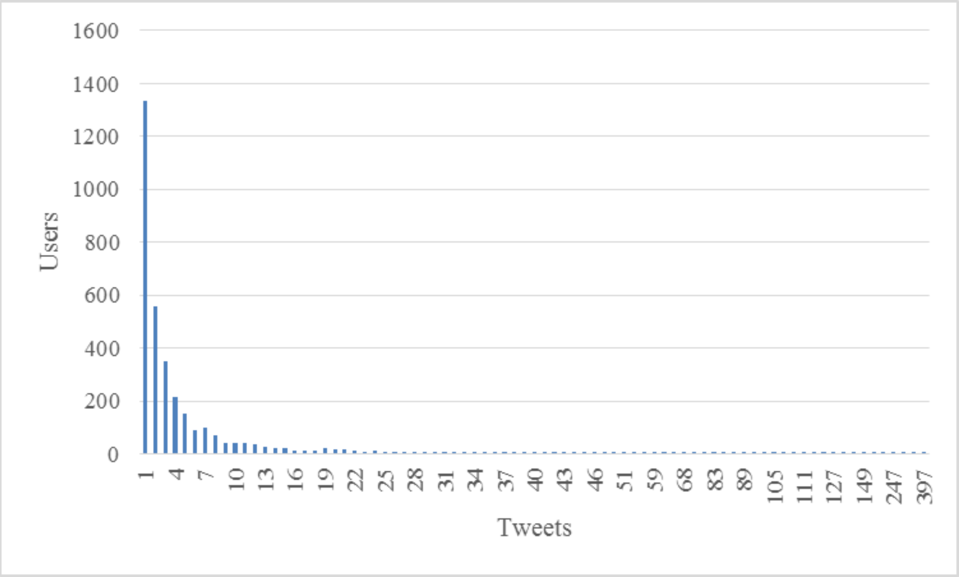


Figure 4.12 Frequency at which users tweet within the Hurricane Sandy data set

4.4.2 Chain Characteristics

There were 204 Chains with and 253 Retweeters. The Average number of tweets per Originator and Retweeter were 26.69 and 25.53, respectively. The average number of Retweets per Originator and Retweeter were 7.41 and 12.78, respectively. Additionally, the percent of tweets that are retweets are 21.72% for Originators and 33.37% for Retweeters. So, the average tweets per user are about the same for each group, but the average and percent of retweet for Retweeters is over 1.5 times higher. The average time from the original tweet to a retweet is 3 Hours 30 minutes; however, if the chain ends with only one retweet the average time is 41 minutes longer. There are many chains in this neighborhood and the data collection time period is long; therefore, the chains are longer, but they also have more time to spread.

4.4.3 Chain Network Visualization

There was a lot of connectivity between the chains initiated within the Sandy neighborhood. Figure 4.13 shows this connectivity. The nodes represented by shapes are active users with more than 5 connections. The highly active nodes could be either the Originator or Retweeter in a relationship. This network contains 167 connections and does not exclude duplicate connections or reciprocal relationships. There are 72 users and 128 chains in this network. Figure 4.14 focuses on two popular nodes within this chain network. These nodes are in the first network but they have a lot of follower loyalty so all connections are not showing because they are layered. The two nodes now show all connections regardless if they include the same user connections. Thus, the chains are visible via the different colored edges connected to vertices.

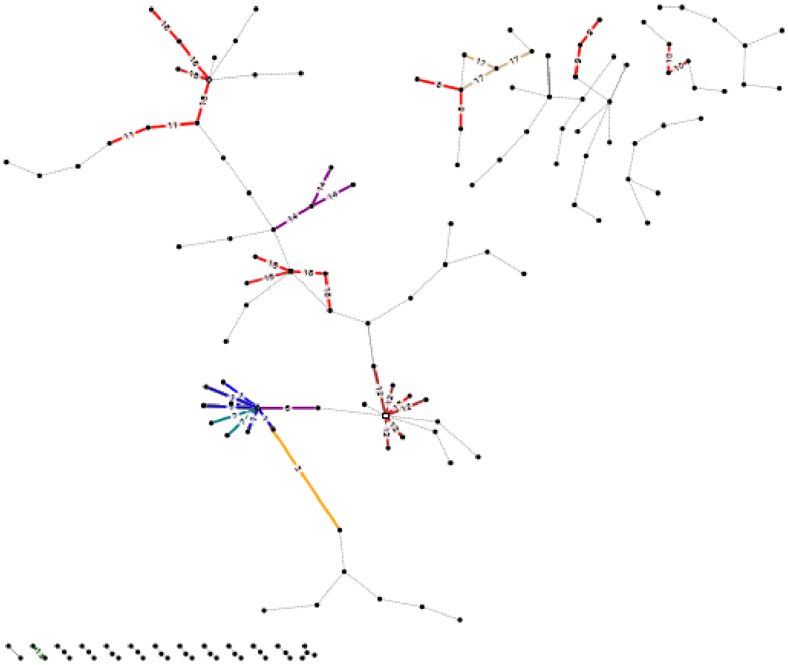


Figure 4.13 The inside chain network for Hurricane Sandy

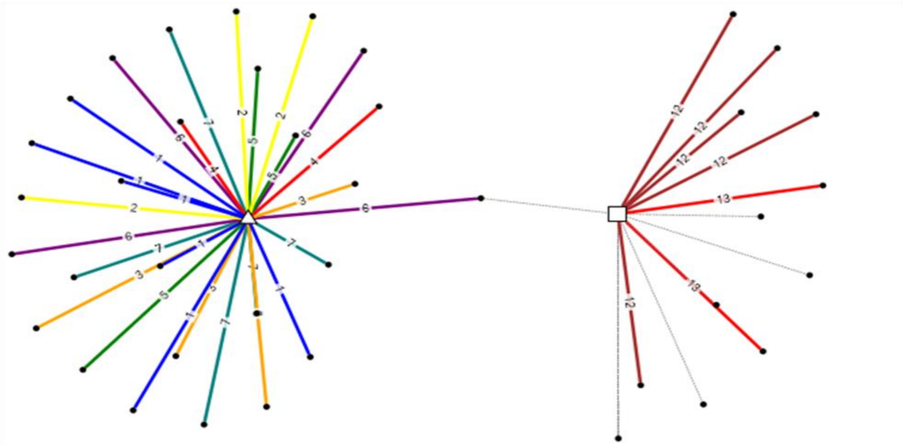


Figure 4.14 A more detailed view of two popular nodes within the inside chain network for Hurricane Sandy

Sandy has many more outside Chains then inside chains. The Chains pictured in Figure 4.15 are only those above 2 tweets long. Once again you can see that users like NY Times, FDNY, GovChristie, Jeff Weiner, NYPL Labs show where the event is occurring. All of these users are connecting to New York and New Jersey. Many users have different retweets that are tweeted several times by the users within the data set.

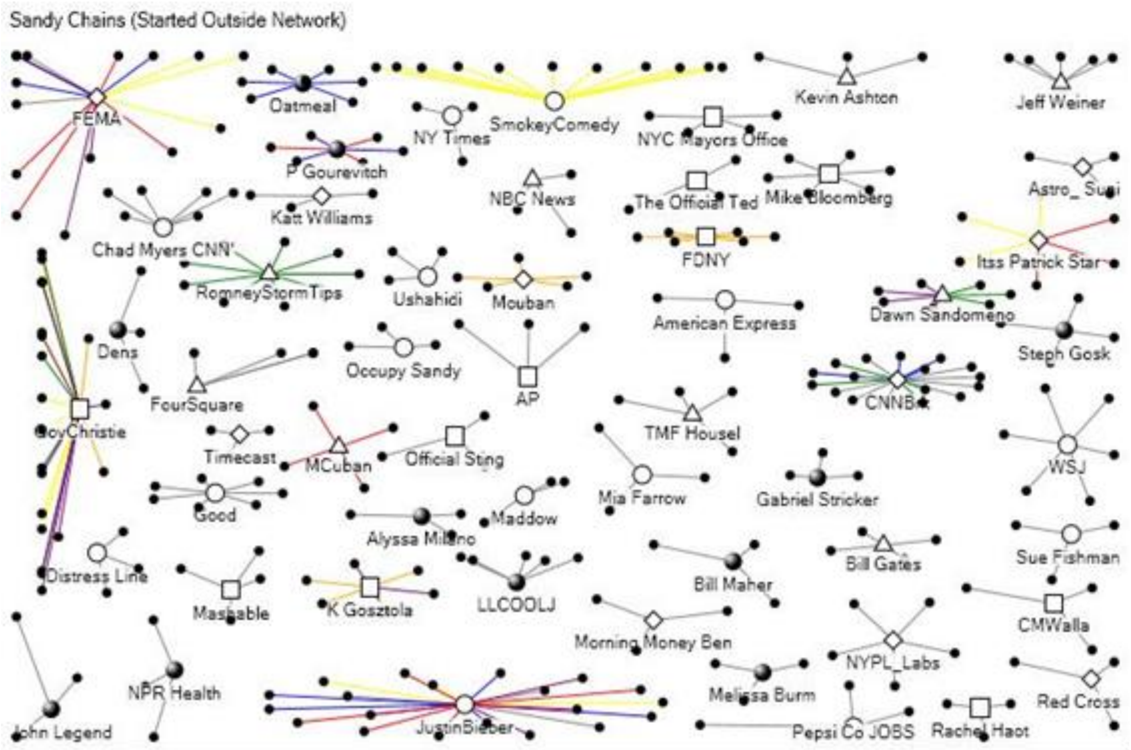


Figure 4.15 The outside chain network for Hurricane Sandy

4.4.4 Content Statistics

The use of #sandy is in 42% of the total tweet count (Figure 4.16). In Figure 4.17, use of #NJ is in 1% of the total dataset but the Chain users account 45% for of the #NJ use. Figure 4.18 shows the use of the #How2Help is visibly a very small amount of tweets within the total dataset; however, it is important to note that it peaks when the data set peaks and 97% of the hashtags come from chain users. Also, the timing of this hashtag is after the storm has passed, destruction is being assessed, and the relief effort has begun. All three of these instances show influence of users using three different hashtags.

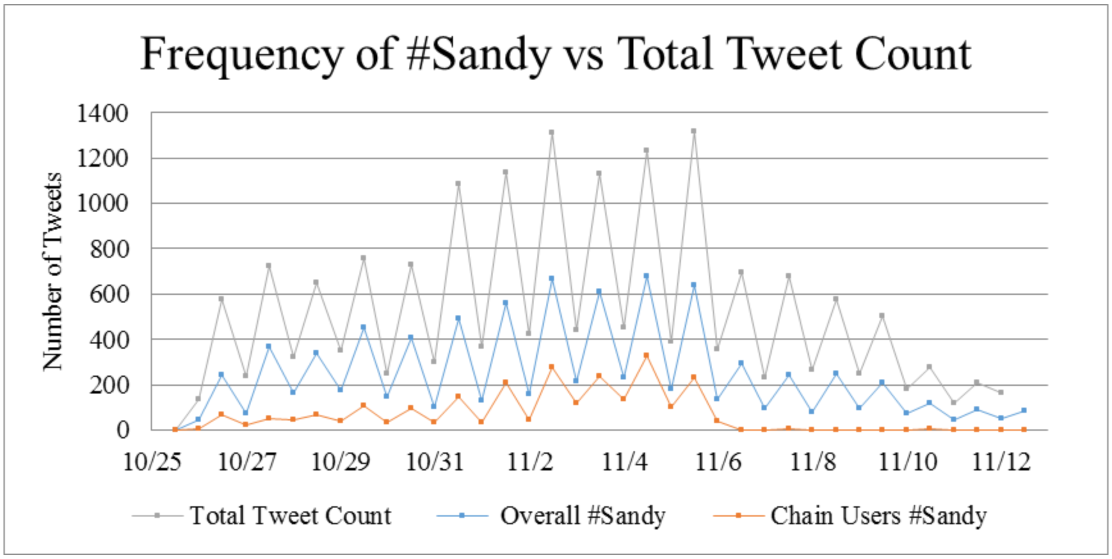


Figure 4.16 The frequency of the use of #Sandy by all users and chain users compared to the total tweet count of the data set.

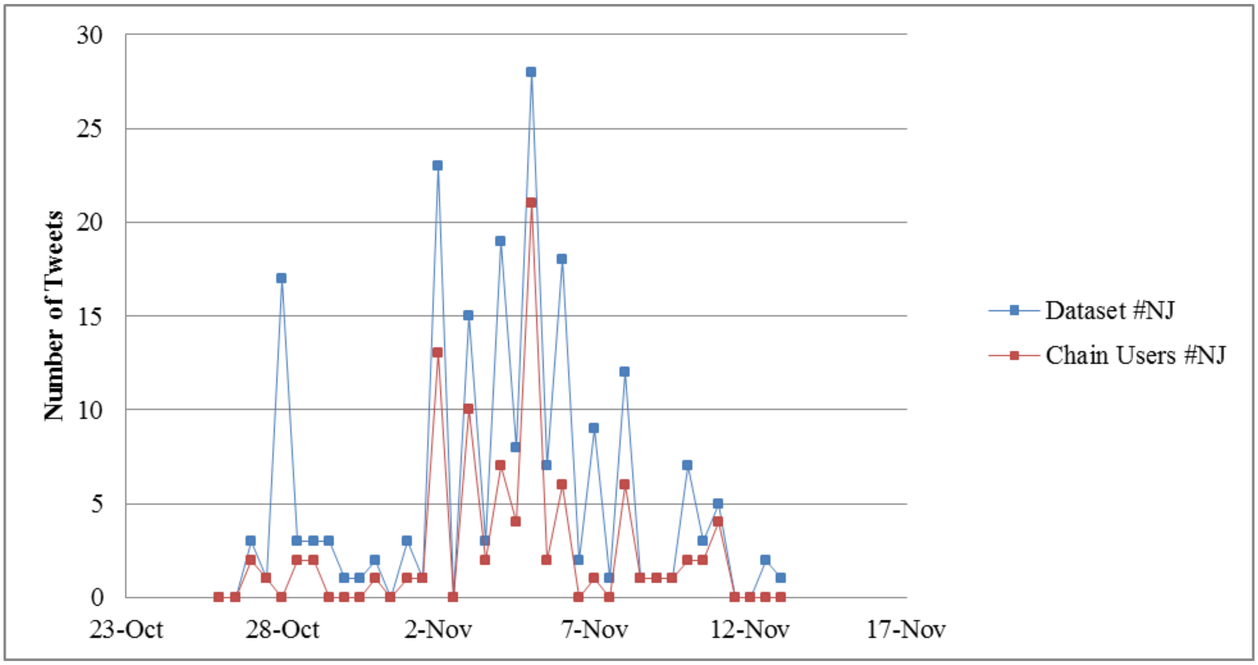


Figure 4.17 The frequency of the use of #NJ within the data set compared to chain

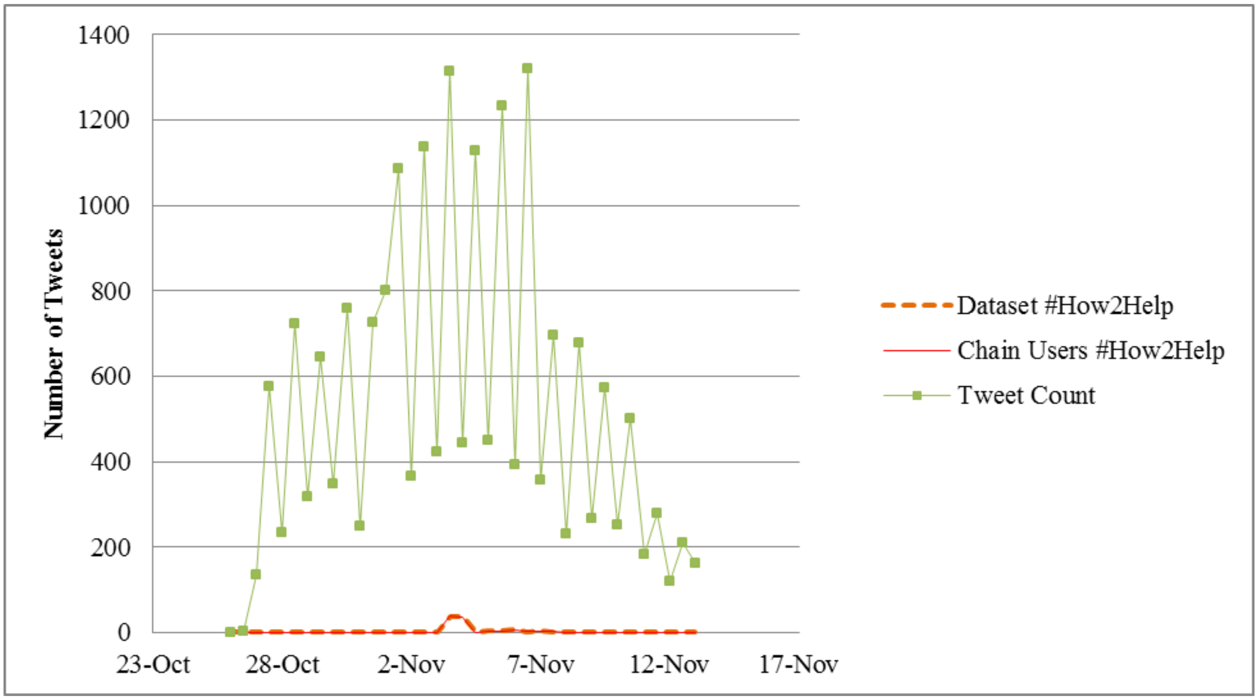


Figure 4.18 The frequency of the use of #How2Help within the data set compared to chain users and overall tweet count.

4.5 Bin Laden Results

4.5.1 General Bin Laden Neighborhood Information

The Bin Laden data was collected over a period of 33 hours. There are 27924 Tweets and 4948 active users in the Irene neighborhood, making it the largest data set. Approximately 37% of neighbors tweet only once and 76% of neighbors tweet 5 times or less. Of the users that only tweet once, those tweets account for 7% of the data set. 27% of tweets come from users that tweet 5 times or less. 5% of the population users tweeted more than 20 times in the 17 day time period. The average number of tweets per user is 5.64. The high frequency users tweeted an average of 45.51 times.

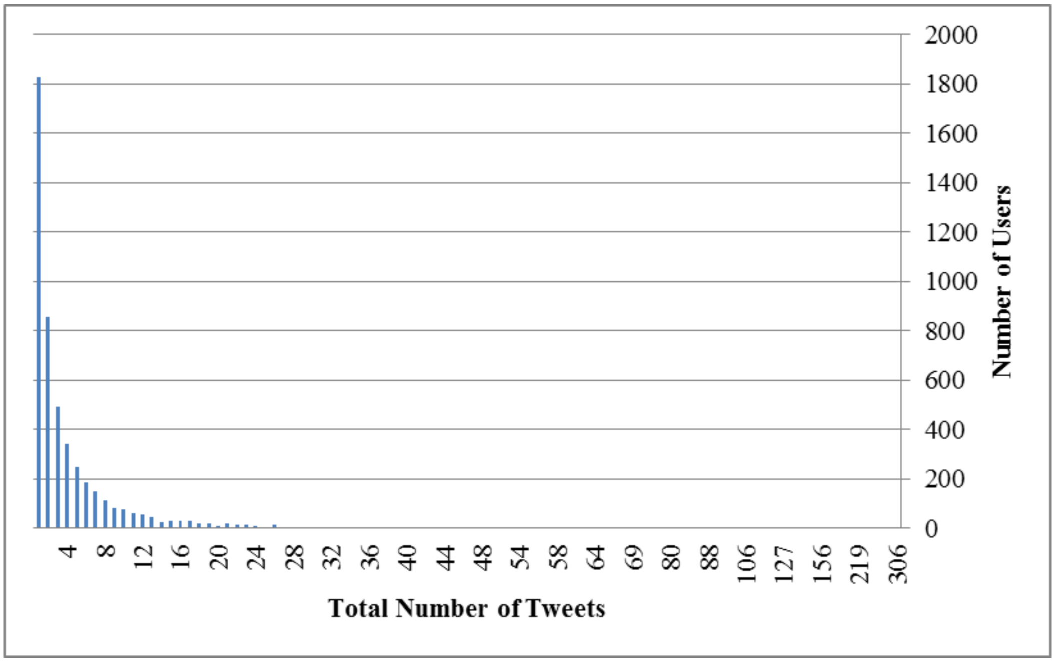


Figure 4.19 Frequency at which users tweet within the Bin Laden data set.

4.5.2 Chain Characteristics

There were 94 Chains with 94 Originators and 120 Retweeters. The average number of tweets per Originator and Retweeter were 24.86 and 22.45, respectively. The average number of Retweets per Originator and Retweeter were 3.57 and 16.98, respectively. Additionally, the percent of tweets that are retweets are 13% for Originators and 43% for Retweeters. So, the tweets are about the same for each group, but the average and percent of retweet for Retweeters is over 2.5 times higher. The average time from the original tweet to a retweet is 17 minutes; however, if the chain ends with only one retweet the average time is 25 minutes.

4.5.3 Chain Network Visualization

There was little connectivity between the chains initiated within the Bin Laden neighborhood, as shown in Figure 4.20. The nodes represented by shapes are active users with more than 5 connections. The more active nodes could be either the

Originator or Retweeter in a relationship. This network contains 120 connections and does not exclude duplicate connections or reciprocal relationships. There are 214 users and 94 chains in this network.

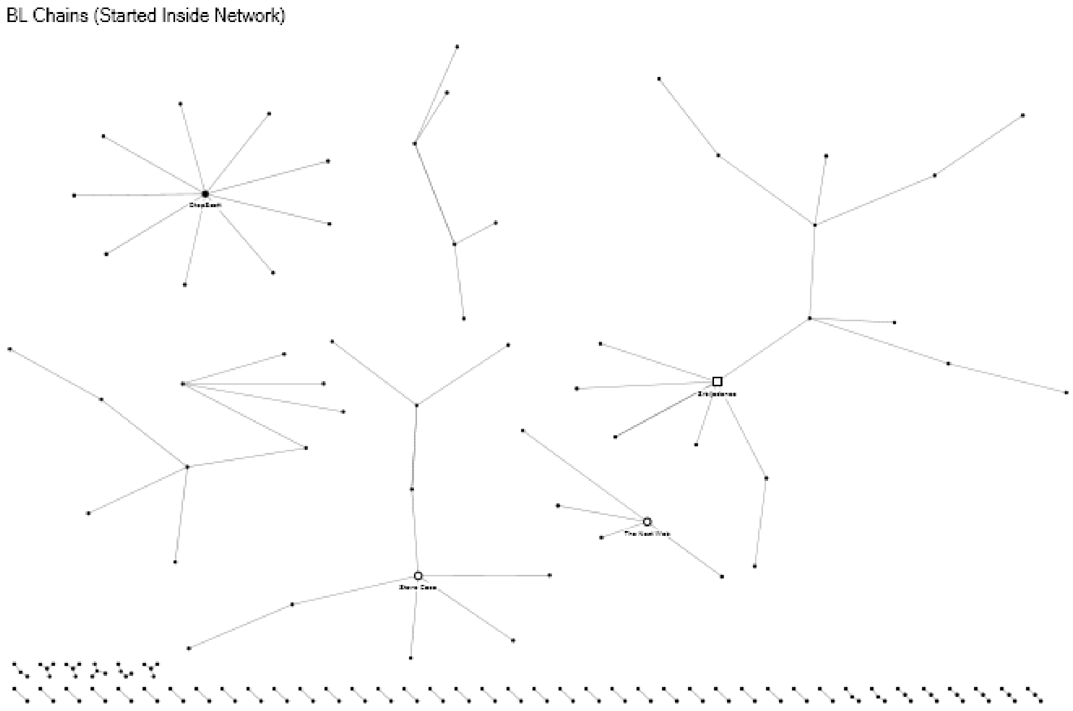


Figure 4.20 The inside chain network for Bin Laden

This data set has many more outside chains then inside chains. The chains picture are only those above 2 tweets long. If a user is involved in more than one chain it is indicted by color. There is connectivity within this network but it seems that there is are smaller networks within.

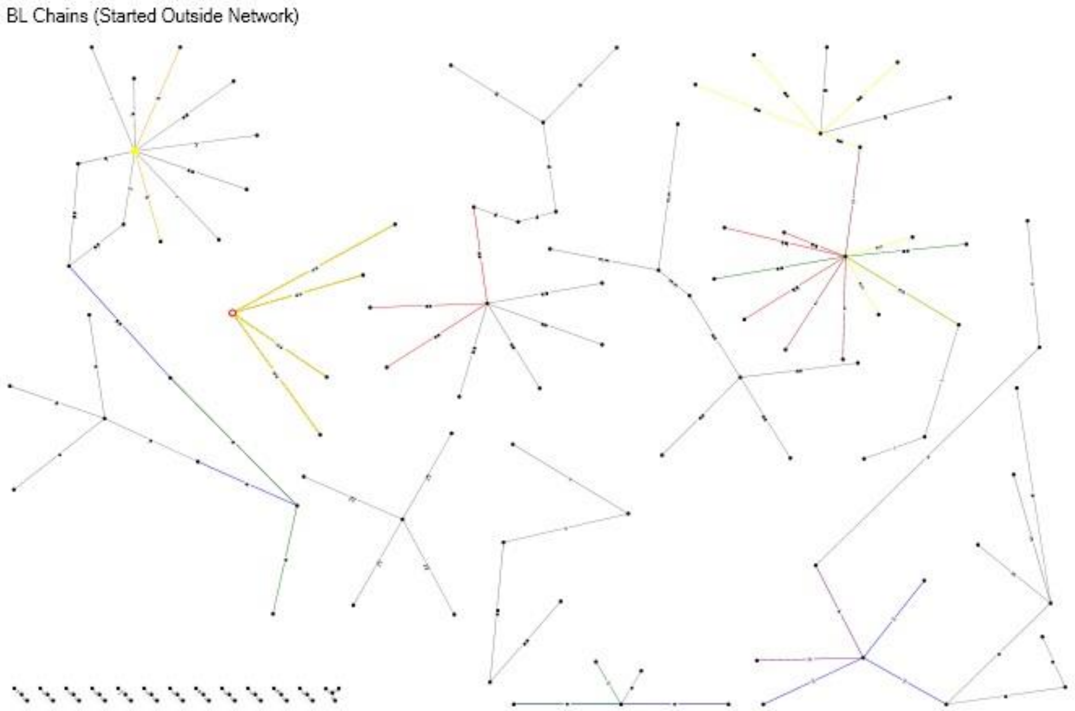


Figure 4.21 The outside chain network for Bin Laden

4.5.4 Content Statistics

Figures 4.22 and 4.23 show the use of hashtags throughout the data set and specifically by chain users. In Figure 4.22 the use of #Osama is in 4% and #Binladen is 1% of the total tweet count. Thus chain users account for 50% for of the #Osama use. #News use goes up and down throughout the dataset and Figure 4.23 shows much chain users also use #news. Also, this news announcement is a single event that does not linger or change. So, the life of this topic on Twitter will fade fast.

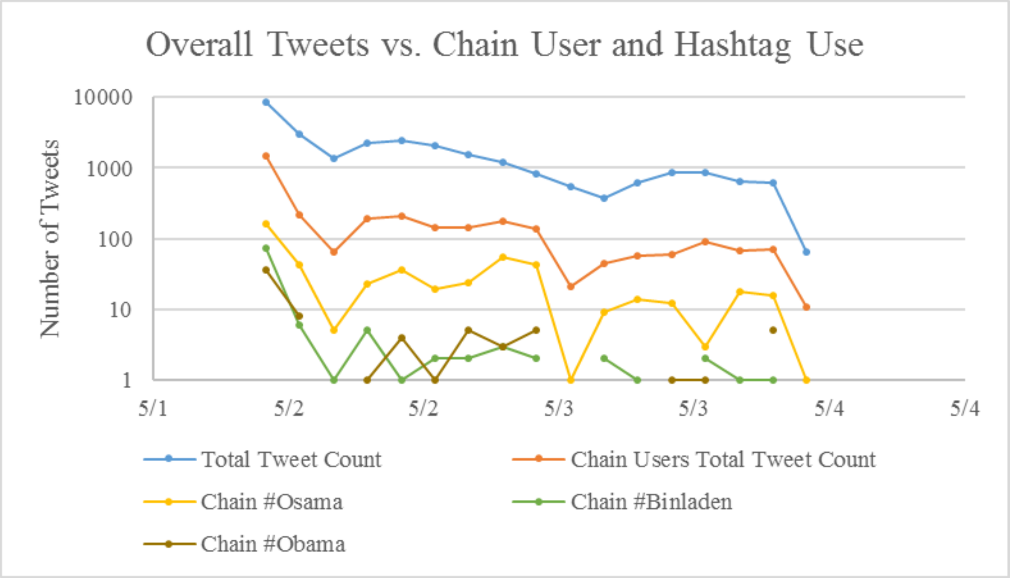


Figure 4.23 The frequency of the use of the most popular hashtags within the Bin Laden data set compared to chain users and overall tweet count. (Logarithmic Scale)

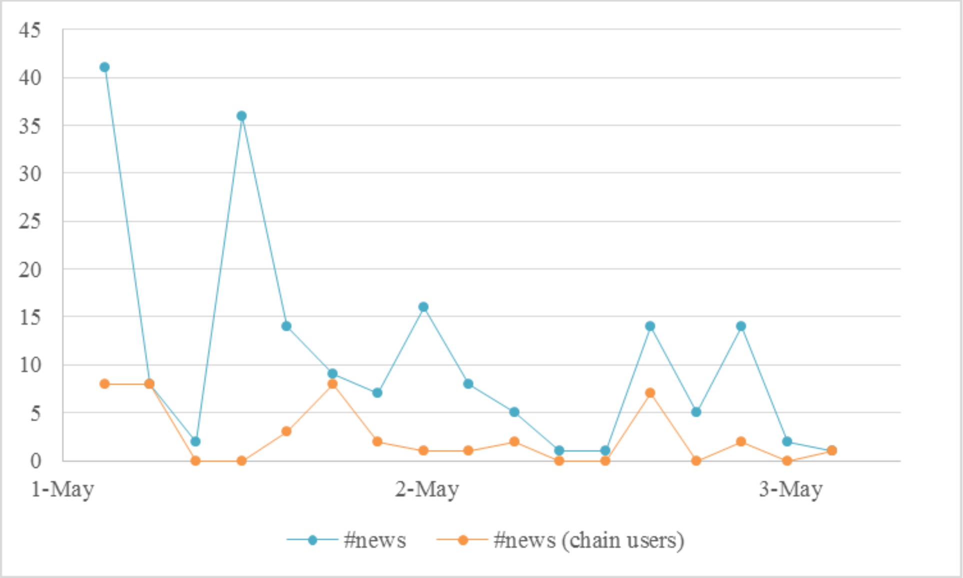


Figure 4.22 The frequency of the use of #News within the data set compared to chain users.

4.6 Comparing the Data Sets

The hurricane Irene and hurricane Sandy data sets have the longest time of data collected; however, hurricane Irene has the lowest number of tweets and active users. Table 4.1 shows a breakdown of these statistics. Hurricane Sandy has the highest average number of tweets per user for all users and high frequency users. These data sets have 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 active users and most tweets. The averages for tweets per user and high frequency user were similar to the Sandy data set.

Table 4.1 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 hours	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 and pass on tweets about this topic frequently. As seen in Table 4.2, 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. Figure 4.24 shows 58% of that Irene data set is made up of tweets from users who tweeted less than 5 times total. In the Sandy and Bin Laden datasets, however, 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.

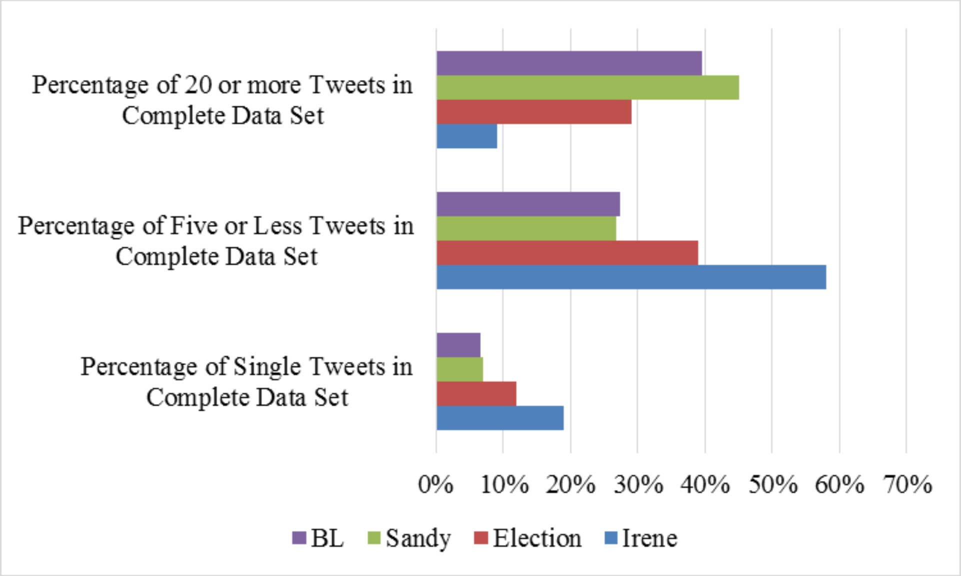


Figure 4.24 Comparison of how much of the data set is made up of single, five or less, and 20 or more tweets

Table 4.2 Frequency of Tweets per user

	Irene	Election	Sandy	Bin Laden
Number of Users to Tweet Once	1141	1142	1334	1826
Percentage of Population that Tweets Only Once	51%	47%	40%	37%
Percentage of Single Tweets in Complete Data Set	19%	12%	7%	7%
Number of Users to Tweet Five or Less Times	1974	3808	2606	3765
Percentage of Population that Tweets Five or Less Times	89%	84%	78%	76%
Percentage of Five or Less Tweets in Complete Data Set	58%	39%	27%	27%
Number of Users to Tweet 20 or More Times	19	70	182	243
Percentage of Population That Tweet 20 or more Times	9%	3%	5%	5%
Percentage of 20 or more Tweets in Complete Data Set	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. As shown in Table 4.3 Sandy also had the longest time to retweet. There were a 42 chains over 3 hours long that skewed this calculation. The lowest time to retweet were the Irene chains. 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 4.3 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 (Of 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 originators and Retweeters, a pattern was easily distinguished about behavior. Retweeters will retweet more than originators. Figure 4.25 shows the average number of tweets and retweets per user and in every case, the average number of retweets is high for retweets. Table 4.4 also shows the percent of tweets sent by all of these users that are retweets. In all cases, Retweeters tweet at least 1.5 times as much as originators.

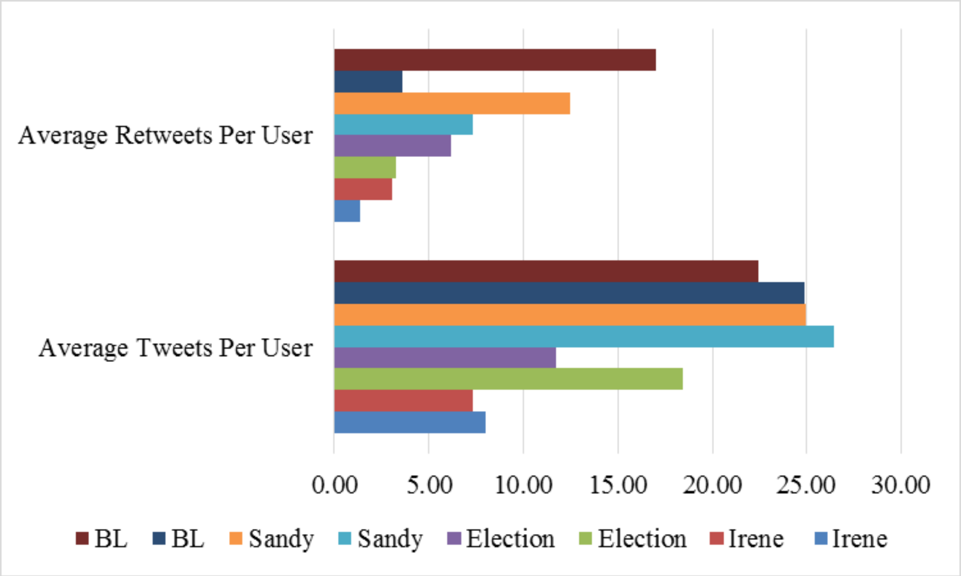


Figure 4.25 Average number of tweets and retweets for chain users in all data sets

Table 4.4 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%

Chapter 5: Summary

This chapter aims to summarize the results of this study of information diffusion during large scale events. The contributions of this work and final thoughts will conclude this portion of the research. Section 5.1 is an overview of observations from the results section and commentary about tweets from the data sets. Section 5.2 highlights the importance and value of hashtag usage. Section 5.3 discusses the content of different types of tweets how that effects tweet popularity. Section 5.4 compares the connectivity of users within chain networks. Section 5.4 explains the limitations of the data, results, and research method. Section 5.5 gives various ideas for additional research.

5.1 Observations

The frequency distribution of how often users tweeted shows that most users tweeted fewer than five times in each data set. Of course, there are several users who tweet more than five times, even more than twenty; however, this study was not focused on their activity specifically. 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 was view 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 found in the neighborhood start from outside of the neighborhood.

In the data sets, the popular tweets are of various types. The most popular tweets were informational, verified, and humorous. Informational tweets that can benefit many people if spread (geographically and socially) are very popular, as they can affect the health and safety of a population. Verified tweets will be spread quickly because of the validity of the source. A verified tweet could come from a celebrity, a social influencer, or an opinion leader. Humorous tweets can spread quickly with the attempts to share humor with a large group of people. On the other hand, selfish tweets, such as rants or emotional information that do not directly affect anyone else in the neighborhood are ignored. Also, 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. People are very willing to share weather information, and that information can come from all over the country and reach users affected by that area. The election was a different type of event, political, that can show different tendencies of users within the same neighborhood.

5.2 Importance of Hashtag Use

The 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. For example, in face to face communication a user may say "That was an insane speech. Utterly brilliant. Let's keep this man," in reference to a campaign speech by President Obama. In tweet text that becomes: "RT @steveweinstein: That was an insane speech. Utterly brilliant. Let's keep this man. #Obama2012 #Vote". The user is thinking about the desire for Obama to stay in office and win the election and also wants other to vote with similar sentiment. The hashtag organizes tweets by topic and puts those tweets in one bin, so Twitter users can search by hashtag and find tweets that include that hashtag. If a hashtag is used a lot in recent tweets it becomes a trending topic on twitter. This work also analyzed data by hashtag use over time.

Graphs were presented for each data set to illustrate the use of various popular hashtags over time. These charts were completed for the total data sets and the subset of tweets from users involved in chains. By completing an overall analysis as well as an identical analysis for chain user, shows how much influence that chain users have on how frequently the hashtags appeared.

5.3 Content

Throughout the analysis of these data sets, popular tweets were further analyzed for their content and classified. These tweets were studied to see why they were retweeted based on the type of content. This point in the research 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.

Example 1

Retweets within the data set are retweeted more if they are informational and the spread of information can be helpful if retweeted. For example, the following tweet shares the contact information to follow Craig Fugate the Federal Emergency Management Association Director. This tweet received 29 retweets in only 1 hour and seven minutes: “FEMA director Craig Fugate is using Twitter to provide hurricane updates. To follow him: @CraigatFEMA”. The next tweet shares the time period of Irene’s most expected destruction. This chain included 5 tweets within the neighborhood. Overall, there were 52 tweets in 10 minutes: “My emergency management team continues 2 tell me the worst of Irene will hit from 6am - 12. It'll get progressively worse through the night”. This tweet was very important to get out quickly considering the originator sent the initial tweet at 10pm the evening before.

Example 2

The following show popular retweets from strangers. The New York City Mayor’s Office sent one tweet that appeared as a retweet three different times in the dataset within 2 hours and 12 minutes. The retweet counts for these tweets were 36, 38, and then 80. The tweet text was “RT @NYCMayorsOffice: Because #Irene's winds could bring down trees, all NYers should stay out of City parks Sunday, and their backyards ...” During this time NYC Mayor’s office was sending many tweets about the preparation for and current status of the hurricane. Please see the following tweets for all the tweets sent by the mayor’s office between 1:12 P.M. and 1:29 P.M. (WXII12, 2011). This is the time period between the first and second tweet appearing in the neighborhood:

- As #Irene arrives, safety will be increasingly important. From 9pm Saturday until 9pm Sunday, NYers should stay indoors.
- High-rise residents: there's a risk of flying debris shattering windows, and that risk increases on the 10th floor or higher. #Irene
- For your safety, stay in rooms with no/few windows. If you live above the 10th floor, consider staying in an apt on a lower floor. #Irene
- Many bldgs have basement/rooftop mechanical equipment that may get flooded, so a good precaution is filling a bathtub/sink with potable H2O

- The Buildings Dept is issuing a stop work order to suspend all construction in the five boroughs from 2pm Sat to 7am Mon. #Irene
- Yellow & livery cabs move to “zone-fare” plan on Saturday w/reduced fares, group rides, & liveries allowed to make street pick-ups. #Irene
- Staten Island Ferry service will be suspended if winds reach 46 mph and seas become too rough. #Irene

Example 3

Humorous Tweets are effective in spreading information on Twitter. For example, the following tweet was retweeted over 101 times, “RT @irene: Btw, tweeting @irene doesn't deliver any messages to the hurricane. Sorry.” Considering hurricane Irene was a negative event, there is very little room to find a positive message. This tweet makes light of the situation and adds sarcasm and humor to the dataset. The tweet was started by a stranger, but viewed within the dataset 12 tweets in 44 hours. The overall increase in tweets could not be measured because the retweet count maxed out at 101 tweets. Within the neighborhood, this tweet was retweeted every 3.66 hours. Based on the amount of times it was retweeted, it seems that neighborhood was saturated with this information; so, even though the information was important and it spread throughout the network, it was less relevant as time progressed. This tweet was spread 11 times in 11 hours and 37 minutes. The last tweet came 32 hours 28 minutes after the 11th tweet. After the 12th tweet, the period of relevance ends for this neighborhood.

Example 4

Kirstie Alley is an actress who does not live on the East Coast; however, her tweet was found once within the dataset, but its total retweet count was over 101. She was attempting to motivate others to evacuate with humor. It is interesting to see how popular this tweet was even though the originator lives on the West Coast. The tweet was found as follows “RT @kirstiealley: If you find the need to evacuate for the impending Irene..PLEASE take your pets.......and of course your kids....LEAVE your bad lovers.” One can assume that her celebrity status and humor facilitate the popularity of this tweet.

Example 5

As previously mentioned, retweets with informational content can be spread quickly. Having a celebrity or opinion leader as an originator will pass the information quicker, than an everyday user because they most often have more followers than an everyday user. Sesame Street is a children’s show that has been on television for over 40 years and teaches children basics like counting and spelling (Hello Design, 2014). This television show uses puppets to teach this information and they also have a large amount of business from retail products. The tweet found in the Irene dataset gives parents a simple opportunity to educate their children about the upcoming and disastrous weather. The tweet was previously sent out, then sent again by Sesame Street, “RT @sesamestreet: In case you missed it: Looking for a way to talk to your children about hurricanes? Here is our hurricane toolkit: htt ...”. The first retweet occurred within the neighborhood. Over the next 8 hours and 39 minutes it was seen twice and the final occurrence showed the max number of tweets

(101+) were sent. This message was very popular even though the message was a copy of one previously sent.

5.4 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.

The connectivity of the outside chain network showed how many more chains came from outside Originators as opposed to those within the data set. The Bin Laden outside network had a lot of chains and moderate connectivity with sub-networks. The Irene and Election data show the most connectivity by far. Sandy's outside network had a lot of chains but not a lot of connectivity; however, there were many Originator's from outside the network that had several tweets appear within the data sets. Celebrity tweets always find their way into the 15k network by 15k users retweeting them.

5.5 Limitations

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 chains networks show lack of connectivity, but that does not mean that users do not see other tweets and simply spread that information via word of mouth. A user could see at tweet and share it via text, verbal conversation, email, or another form of social media. This study could not measure if tweets reach users without recording their activity on Twitter.

The Bin Laden Data set was collected by time period and not by topic; however, the data was separated by keywords to filter out tweets about Bin Laden. This data set was actually the largest and had the most hashtags to analyze. The Irene and Election networks are not as connected as assumed. For Irene, of the 5948 tweets within the neighborhood, only 23 chains were found. The longest chain had 52 tweets throughout all of Twitter but only four of those retweets occurred within the neighborhood. Tweet chains travel in and out of the neighborhood, 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.

5.6 Future research

Further research should attempt to find patterns in similar events such as weather emergencies. In addition, focusing on the same active users in the Irene and Sandy data sets could potentially show how users change their behavior overtime.

The research focus would be about tweets stemming from similar events to test if response is similar and because events are similar.

The social family of a given user and their activity level is important to study. This research would benefit from an analysis of popular individual users and analyze their network's activity (Retweeters) as well as their tendencies. Patterns could be discovered in popular tweets and the users that retweet those tweets. Rumor propagation should be further researched in such an event. The knowledge of rumor confirmation or information verification as truth can illustrate how the chain evolved. Moving forward, an algorithm to measure influence and predict popularity would be helpful with user and chain analysis. Also, more focus on content and sentiment analysis can help us better understand the users, content, and motivation.

More data collection is needed to see how Twitter has changed since the last data collection. Table 3.1 is an example of this evolution showing every data set analyzed collected different types of data. Also, Twitter increases the amount of users every year and past users may have different behavior. A new data collection and analysis would show network growth as well as any additional records provided by the API. Furthermore, analyzing two separate networks tweeting about the same topics can give a real-time comparison of activity. This type of analysis could improve upon the conclusions made by this 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 have provide a greater and more detailed understanding of the effects of content and user popularity.

Bibliography

"About Twitter." Twitter. Twitter Inc, 2014. Web. 10 Apr. 2014.

Asur, Sitaram, & Huberman, Bernardo A. (n.d.). Predicting the Future with Social Media. 2010

Asur, Sitaram, Huberman, Bernardo A., Szabo, Gabor, & Wang, Chunyan. (2011).Trends in Social Media : Persistence and Decay.

Bass, Frank M. "A New Product Growth for Model Consumer Durables." Management Science. 50 (2004): 1825-1832. Print.

Bauckhage, Christian, and Kristian Kersting. "Strong Regularities in Growth and Decline of Popularity of Social Media Services." ArXiv (2014): n. page. 25 June 2014. Web. 12 Aug. 2014.

Beaumont, Peter. "The Truth about Twitter, Facebook and the Uprisings in the Arab World." The Guardian. Guardian News and Media, 25 Feb. 2011. Web. Mar. 2014.

Benjamin Doerr , Mahmoud Fouz , Tobias Friedrich, Why rumors spread so quickly in social networks, Communications of the ACM, v.55 n.6, June 2012 [doi>10.1145/2184319.2184338]

Berger, Jonah. Contagious: Why Things Catch on. New York: Simon & Schuster, 2013. Print.

Bitly. "You Just Shared a Link. How Long Will People Pay Attention?" Bitly Blog. N.p., 6 Sept. 2011. Web. 04 Feb. 2014.

Business Dictionary. "What Is a Marketing Mix? Definition and Meaning." BusinessDictionary.com. WebFinance, Inc, 2014. Web. 21 Aug. 2014.

Cha, Meeyoung; et al. "Measuring User Influence in Twitter: The Million Follower Fallacy", Association for the Advancement of Artificial Intelligence, 2010, PDF.

Chew, C, and G Eysenbach. "Pandemics in the Age of Twitter: Content Analysis of Tweets During the 2009 H1n1 Outbreak." Plos One. 5.11 (2010). Print

Daley, D J, and D G. Kendall. "Epidemics and Rumours." Nature. 204.4963 (1964): 1118-1118. Print.

Daniel, Romero, Meeder Brendan, and Kleinberg Jon. Differences in the Mechanics of Information Diffusion Across Topics. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2011. Print.

Dietz, Klaus. "Epidemics and Rumours: A Survey." Journal of the Royal Statistical Society A 130.4 (1967): 502-28. JSTOR. Web. 14 Oct. 2013.

Easley, David, and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. New York: Cambridge UP, 2010. Print.

Eytan, Bakshy, Hofman Jake, Mason Winter, and Watts Duncan. Everyone's an Influencer. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2011. Print.

Gupta, Pankaj, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. "WTF: The Who to Follow Service at Twitter." WWW 13. Proc. of International World Wide Web Conference, Windsor Barra Hotel, Rio De Janeiro, Brazil. Geneva: International World Wide Web Conferences Steering Committee, 2013. 505-14. ACM Digital Library. Web. 20 Mar. 2014.

Hayes, B. "Rumours and Errours." American Scientist. 93.3 (2005): 207-211. Print.

Hello Design. "Sesame Street." Sesame Street. Sesame Street Workshop, 2014. Web. 20 May 2014.

Hui, Cindy, Mark Goldberg, Malik Magdon-Ismail, and William A. Wallace. "Simulating the Diffusion of Information: an Agent-Based Modeling Approach." International Journal of Agent Technologies and Systems. 2.3 (2010): 31-46. Print.

Humphreys, Lee, Phillipa Gill, Balachander Krishnamurthy, and Elizabeth Newbury. "Historicizing New Media: a Content Analysis of Twitter." Journal of Communication. 63.3 (2013): 413-431. Print.

Huo, L, P Huang, and C.-X Guo. "Analyzing the Dynamics of a Rumor Transmission Model with Incubation." Discrete Dynamics in Nature and Society. 2012 (2012). Print.

Isabel, Anger, and Kittl Christian. Measuring Influence on Twitter. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2011. Print.

James, Abello, and Queyroi François. Fixed Points of Graph Peeling. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2013. Print.

Joshi, Yogesh, Liye Ma, William Rand, and Louiqa Raschid. "Building the B[r]and." Marketing Science Institute. Marketing Science Institute, 2013. Web. Jan. 2014.

Liangjie, Hong, Dan Ovidiu, and Davison Brian. Predicting Popular Messages in Twitter. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2011. Print.

"Live Wire: Hurricane Irene Impacts East Coast." WXII12. Hearst Properties Inc, 26 Aug. 2011. Web. 01 Apr. 2014. <http://livewire.wxii12.com/Event/Hurricane_Irene_Impacts_East_Coast?Page=0>.

Maki, Daniel P, and Maynard Thompson. Mathematical Models and Applications: With Emphasis on the Social, Life, and Management Sciences. Englewood Cliffs, N.J: Prentice-Hall, 1973. Print.

Matsubara, Y, Y Sakurai, B.A Prakash, C Faloutsos, and L Li. "Rise and Fall Patterns of Information Diffusion: Model and Implications." Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. (2012): 6-14. Print.

Monner, Derek. "TwEater." GitHub. GitHub, 4 July 2013. Web. 01 Oct. 2013.

Murdough, Chris. "Social Media Measurement: It's Not Impossible." Journal of Interactive Advertising 10.1 (2009): 94-99. EBSCO. Web. Feb. 2014.

Murthy, Dhiraj. Twitter: Social Communication in the Twitter Age. Cambridge, UK: Polity, 201. Print.

Nekovee, M., Y. Moreno, G. Bianconi, and M. Marsili. "Theory of Rumour Spreading in Complex Social Networks." Physica A: Statistical Mechanics and Its Applications 374.1 (2007): 457-70. Web.

Penny, Laurie. " Revolts don't have to be tweeted." New Statesman. New Statesman, 15 Feb. 2011. Web. Feb. 2014.

Radas, Sonja. "Diffusion Models in Marketing: How to Incorporate the Effect of External Influence?" Economic Trends and Economic Policy 15.105 (2006): 30-51. Portal of Scientific Journals of Croatia. Web. 20 Aug. 2014.

Raghuvanshi, Gaurav. "Tweeting to Keep Disaster at Bay - India Real Time - WSJ." India Real Time RSS. Dow Jones & Company, 18 Oct. 2013. Web. Nov. 2013.

Sanmitra, Bhattacharya, Tran Hung, Srinivasan Padmini, and Suls Jerry. Belief Surveillance with Twitter. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2012. Print.

Shaomei, Wu, Hofman Jake, Mason Winter, and Watts Duncan. Who Says What to Whom on Twitter. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, 2011. Print.

Smith, Jason. "The Engagement Trajectory: How Consumers Socially Engage with Brands." Social Media Today. Social Media Today LLC, 8 July 2011. Web. 04 Nov. 2013.

Sullivan, Danny. "Why “Second Chance” Tweets Matter: After 3 Hours, Few Care About Socially Shared Links." Search Engine Land. Third Door Media, 7 Sept. 2011. Web. Oct. 2013.

Sylvester, Jared and Healey, John and Wang, Chen and Rand, William M., Space, Time, and Hurricanes: Investigating the Spatiotemporal Relationship Among Social Media Use, Donations, and Disasters (May 23, 2014). Robert H. Smith School Research Paper No. RHS 2441314. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2441314

Wang, F, H Wang, K Xu, J Wu, and X Jia. "Characterizing Information Diffusion in Online Social Networks with Linear Diffusive Model." Proceedings - International Conference on Distributed Computing Systems. (2013): 307-316. Print.

Wu, Jean, Richard Socher, Rukmani Ravisundaram, Tayyab Tariq, and Jason Chuang. "Sentiment Analysis: Live Demo." Recursive Neural Tensor Network. Stanford University, Aug. 2013. Web. 20 Jan. 2014.

Yang, J, and J Leskovec. "Modeling Information Diffusion in Implicit Networks."Proceedings - Ieee International Conference on Data Mining, Icdm. (2010): 599-608. Print.

Ye, S, and S.F Wu. "Measuring Message Propagation and Social Influence on Twitter.com." Lecture Notes in Computer Science. (2010): 216-231. Print.

Zhang, Z.l, and Z.q Zhang. "An Interplay Model for Rumour Spreading and Emergency Development." Physica A: Statistical Mechanics and Its Applications. 388.19 (2009): 4159-4166. Print.

Zhao, Laijun, Jiajia Wang, Yucheng Chen, Qin Wang, Jingjing Cheng, and Hongxin Cui. "Sihr Rumor Spreading Model in Social Networks." Physica A: Statistical Mechanics and Its Applications. 391.7 (2012): 2444-2453. Print.

Zhao, Laijun, Qin Wang, Jingjing Cheng, Yucheng Chen, Jiajia Wang, and Wei Huang. "Rumor Spreading Model with Consideration of Forgetting Mechanism: A Case of Online Blogging LiveJournal." Physica A: Statistical Mechanics and Its Applications 390.13 (2011): 2619-625. Web.