Online Content Popularity in the Twitterverse: A Case Study of Online News

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Md Nazmul Hasan

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

With the advancement of internet technology, online news content has become very popular. People can now get live updates of the world's news through online news sites. Social networking sites are also very popular among Internet users, for sharing pictures, videos, news links and other online content. Twitter is one of the most popular social networking and microblogging sites. With Twitter's URL shortening service, a news link can be included in a tweet with only a small number of characters, allowing the rest of the tweet to be used for expressing views on the news story. Social links can be unidirectional in Twitter, allowing people to follow any person or organization and get their tweet updates, and share those updates with their own followers if desired. Through Twitter thousands of news links are tweeted every day.

Whenever there is a popular new story, different news sites will publish identical or nearly identical versions ("clones") of that story. Though these clones have the same or very similar content, the level of popularity they achieve may be quite different due to content agnostic factors such as influential tweeters, time of publication and the popularities of the news sites. It is very important for the content provider site to know about which factor plays a important role to make their news link popular. In this thesis research, a data set is collected containing the tweets made for the 218 members of 25 distinct sets of news story clones. The collected data is analyzed with respect to basic popularity characteristics concerning number of tweets of various types, relative publication times of clone set members, tweet timing and number of tweeter followers. Then, several other factors are investigated to see their impact in making some news story clones more popular than others. It is found that multiple content-agnostic factors i.e. maximum number of followers, self promotional tweets plays an impact on news site's stories overall popularity, and a first step is taken at quantifying their relative importance.

Acknowledgements

I would like to express my sincere appreciation to Dr. Derek Eager for all his invaluable guidance, support and encouragement as my supervisor. His endless energy and enthusiasm in research had motivated me and provided me the rhythm of doing research. His encouragement, support and guidance shape me a lot to be a good researcher. In addition, he was always accessible, motivated and willing to help me in my research even in his tight schedule. As a result, my research life became smooth and rewarding for me. It has been my great pleasure and honor to have worked with him. I would also like to thank Dr. Niklas Carlsson for their advice, guideline and feedback on my research output.

I would also like to thank the members on my supervisory committee, Dr. Dwight Makaroff and Dr. Christopher Dutchyn, as well as my external examiner Dr. Ha Nguyen, for their helpful comments and constructive suggestions. Besides, I would like to express my gratefulness to all the faculty, staff and graduate students in the Department of Computer Science, for their caring and support during my study at the University of Saskatchewan.

Finally, I would like to thank my family and friends for their continuous encouragement and help, especially my parents who sacrificing the companionship of their only son for this research work. My parents are the idol of my life; they did lots of sacrifice for their family. Some cases it was difficult for them to achieve their own dream because of the surroundings. I am very happy as I am fulfilling their dream. I am grateful for their continuing understanding and countless support throughout my entire life. I am also thankful to my lovely wife who accompanying me here and make my life comfortable, though sometimes I feel without her presence, I would have completed this thesis earlier.

This thesis is dedicated to my father Md Abdus Sattar, who is my Super Hero as well as my Inspiration; and my caring mother Nasima Akter without whose contribution in my life I am nothing. Thanks a lot for all of your contributions, support and impact in my life....

-Nazmul

Contents

List of Tables

List of Figures

List of Abbreviations

API	Application Programming Interface
CDF	Cumulative Distribution Function
CCDF	Complementary Cumulative Distribution Function
CSV	Comma Separated Value
HTTP	Hypertext Transfer Protocol
URL	Uniform Resource Locator

Chapter 1

Introduction

Social networking sites have become very popular for sharing views, ideas, user profiles, and other information. Different social networking sites are being used for different purposes, such as for professional use, for communication with friends and family, and for online content sharing. Due to the advancement of the Internet, different kinds of online content have become popular and easily accessible to the users. Social networking sites can play a large role in promoting online content items and increasing their popularity.

Twitter is currently one of the most popular social networking and micro blogging sites. It was launched in 2006[1] and every day thousands of people become new users of this site.[2] Unlike many social network sites, Twitter allows unidirectional connections, enabling people to get updates concerning any other Twitter user without requiring mutual interest from that other user. An important application of Twitter is the propagation of links to online content, alerting users to potentially interesting content without requiring those users to surf the original content provider site. This thesis concerns the characterization of aspects of content popularity in Twitter, focusing on online news in particular, and the factors impacting its popularity.

Sections 1.1 and 1.2 present basic concepts regarding social networking sites, and online content, respectively. Basic information about Twitter is provided in Section 1.3. The popularity characteristics of online content and the factors impacting popularity are discussed in Section 1.4. The motivation of the thesis, the thesis contributions, and the thesis organization are described in Sections 1.5, 1.6, and 1.7, respectively.

1.1 Social Networking Services

A social networking site is an online service that builds social networks or social relations among people who can share their interests, activities, backgrounds, or real-life connections among themselves. Facebook is the most popular social networking site, with 750 million unique monthly visitors as of December 2012.[3] Other popular sites include Twitter, with 250 million unique monthly visitors, LinkedIn with 110 million unique monthly visitors, and Google Plus with 65 million unique monthly visitors, as of December 2012.[4] Additional

[1] http://www.crunchbase.com/company/twitter; accessed on 30 August - 2013
[2] http://mashable.com/guidebook/twitter/; accessed on 30 August - 2013
[3] http://www.bizjournals.com/dayton/news/2012/12/17/top-15-most-popular-social-media-sites.html; accessed on 15 March - 2013
[4] http://www.bizjournals.com/dayton/news/2012/12/17/top-15-most-popular-social-media-sites.html; accessed on 15 March - 2013

social networking sites are popping up, and accumulating users and gaining popularity.

Social networking sites such as Facebook provide diverse services to their users including chatting, online content sharing, photo uploading, and video conferencing. Although social networking sites primarily focus on building and supporting relationships between people with similar interests and activities, they are also used for business purposes. Different types of advertisements are published on social network sites which are one of the sources of income for those sites. Many businesses use at least one social networking site for advertising. LinkedIn is a site focused on professional connections where people can advertise their professional activity and skills, and get connected with other people in their fields.

Twitter provides a microblogging service that makes it one of the fastest sites for propagating messages to an individual's followers. So, for example, a link to interesting online content can be very quickly propagated. One person can follow others without the relationship being reciprocated. The immense popularity of Twitter has prompted other social networking sites such as Facebook to introduce similar features.

1.2 Online Content

Many types of online content are available today in the Internet through a variety of popular sites. For example, sites such as YouTube and Dailymotion support publishing and watching user generated videos, while Flickr is dedicated to photo sharing. Other sites are dedicated for providing news, which can be updated on the site as events unfold. Such sites make it unnecessary to wait for a morning newspaper or a TV/radio broadcast to learn the latest local or world news. Many news sites are available today, including popular English-language sites such as BBC News Online, Yahoo News, CBC News Online, FOX News, CNN, Guardian Unlimited, The New York Times, and Reuters.[5] These news sites provide a wide variety of news stories and update their news frequently. Micro blogging is a relatively new phenomenon in the Web 2.0 world of user generated content [14]. Micro blogging site like Twitter plays a vital role in order to propagating online contents like news link along with the comments of Twitter users about the links.

Sometimes the same, or at least highly similar, content is present on multiple sites or in different files on the same site. For example, there could be multiple YouTube videos showing the same event, uploaded by different uploaders. In the case of online news content, the same news event could be written about by different reporters for different news sites, or a news story produced by a news agency such as Reuters or Associated Press could be shared by different news sites. Following previous practice [6], in this thesis identical or nearly identical online content items are termed "clones", and a set of such content items is termed a "clone set".

[5] http://news.nettop20.com/; accessed on 18 March - 2013

1.3 Twitter

Twitter is an online social networking website and microblogging service, which recently has gained much attention and following [41]. Twitter was launched as a social networking site in July 2006. Twitter has rapidly gained worldwide popularity, with over 500 million registered users as of 2012, generating over 340 million tweets daily and handling over 1.6 billion search queries per day.[6] It has became one of the ten most visited websites on the Internet, and has been treated as the SMS of the Internet.[7] Twitter enables its users to send and read text-based messages of up to 140 characters, known as "tweets". Tweets can contain a variety of types of content such as daily activity updates, comments, references to photos or videos, and interesting Web URLs. Each Twitter user can choose which other users to follow and the tweets from the followed users are aggregated in a single reverse chronologically ordered stream [14]. Unregistered users can read tweets, while registered users can post tweets through the website interface, SMS, or a range of applications for mobile devices.[8]

The relationship between the followers and the followed is unidirectional and need not be reciprocated or individually approved, making it very easy to follow any individuals such as a celebrity or a Twitter user representing a news agency, for example. Although Twitter allows its users tweets of at most 140 characters, a URL can be included in a tweet using only 20 characters because of Twitter's URL shortener.

Some basic Twitter terminology and concepts are summarized below.

- Tweet: A text-based message of at most 140 characters.
- Retweet (RT): Re-sharing or giving credit to someone else's tweet. There are two ways to retweet. One of these is using the Retweet button when hovering over a tweet in the stream of tweets being displayed for the user. Another way is by copy and pasting the tweet into the user's own tweet, adding the letters RT and @username where user name is the name of the individual from whom that tweet was received.
- Feed: The stream of tweets being displayed to a user.
- Handle: The Twitter username of a user.
- Mention (@): The way to reference another user using that user's Twitter username in a tweet (e.g. @mashable). Users are notified when @mentioned. Thus, this is a way to conduct discussions with other users in a public realm.
- Direct Message (DM): A private, 140-character message between two people. It is possible to send a DM to a user only if that user is a follower.[9]

[6] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city; accessed on 20 March - 2013

[7] http://www.business-standard.com/article/technology/swine-flu-s-tweet-tweet-causes-online-flutter-109042900097.html; accessed on 20 March - 2013

[8] https://support.twitter.com/groups/34-apps-sms-and-mobile/topics/153-twitter-via-/sms/articles/14014-twitter-via-sms-faq; accessed on 22 March - 2013

[9] http://mashable.com/2012/06/05/twitter-for-beginners/; accessed on 24 March - 2013

- Hashtag (#): This can be used to denote a topic of conversation or participate in a larger linked discussion (e.g. #IPL, #Obama, #jamat). A hashtag is a labelling tool that allows others to find your tweets, based on topic. By clicking on a hashtag, a user can see all the tweets that include this hashtag in real time, even those from people that are not being followed by the user.

- Lists: A user can organize the individuals that the user is following by putting them in different lists.

- Favourite: A tweet can be marked as a favourite by a user, which will bookmark the tweet for later viewing, as well as possibly notify the sender of the tweet. A user's favourites can also be publically viewed.

- URL shortener: A URL shortening service is a Web service that provides short aliases for redirection of long URLs. For example, the homepage of the TinyURL service includes a form that is used to take input of a long URL for shortening. For each URL that is entered, the server adds a new alias in its hashed database and returns a short URL. If the URL has already been requested, TinyURL will return the existing alias rather than create a duplicate entry. The short URL forwards users to the long URL whenever they enter that URL into the browser.[10] Twitter has it own URL shortener named T.co. It is embedded with Twitter, so users can just copy and paste a URL to the Twitter status update field and Twitter will automatically shorten the URL to 20 characters. In addition to TinyURL and T.co, other popular third party URL shorteners include Bit.ly, ls.gd, Goo.gl as well as many others.[11]

1.4 Online Content Popularity

A number of studies have examined the popularity characteristics of online content (e.g. [8, 18, 30]). Often of interest is the popularity distribution within a set of content items (such as Web pages or videos, for example), with most studies observing a heavy-tailed distribution (e.g. [26], [10]). Also of interest has been popularity evolution over time (e.g. [10], [7]). For some types of online content, such as user generated video, popularity can be surprisingly resilient, with content remaining popular for years [19]. The popularity of online news items, in contrast, is generally short-lived. Online news content popularity typically lasts at most around four days [3].

The most important factor impacting the popularity of an online content item is, of course, the content itself. However, content-agnostic factors such as the "first mover advantage", "rich-get-richer" behaviour, content age, and publisher popularity, can also play an important role in determining popularity. Borghol *et al.* proposed the study of such content-agnostic factors through measurement of the popularity of online content items that contain the same or nearly the same content ("clones")[6]. They applied their proposed

[10] http://en.wikipedia.org/wiki/TinyURL; accessed on 25 March - 2013
[11] http://sproutsocial.com/insights/2011/09/twitter-url-shorteners/; accessed on 26 March - 2013

methodology to YouTube videos. It does not appear that this methodology has been applied to any other types of online content in any work prior to this thesis.

1.5 Thesis Goal and Methodology

Online content is very popular in today's world and often many versions of almost the same content are available in the Internet. Different versions of almost the same content frequently achieve different levels of popularity. An important issue for online content providers therefore concerns the content-agnostic factors that impact popularity and their relative levels of importance.

The goal of this thesis is to contribute to the understanding of online content popularity, specifically the popularity of content within the Twitterverse. Of interest is both characterization (e.g., popularity distribution among sets of content items, popularity evolution over time) and the content-agnostic factors that impact the observed popularity (what are they, which are the most important). A particular type of online content, online news, is focused on as a case study in this thesis. An advantage of considering this particular type of content is that the clone methodology of Borghal *et al.* [6] could be applied here to allow study of the factors impacting popularity, but for a quite different type of online content than the YouTube video content considered by Borghal *et al.* This should yield improved understanding of the factors impacting the popularity of online content.

In order to achieve this goal, news stories were identified for which there were identical or nearly-identical versions on different news sites. Each such news story yielded a clone set, with the multiple versions being the clone set members. The Twitter Search API was used to collect the tweets that referenced each version. A different API was used for finding information concerning the Twitter users who posted these tweets, such as numbers of followers. The raw data collected in this manner was then parsed and used for various analyses. These analyses included both characterization of the collected data (with respect to numbers and types of tweets, site popularity, relative clone set member publication times, evolution of tweeting activity over time, and popularity of the tweeters as measured by number of followers), and analysis of the impacts of various factors (tweeter popularity, news site popularity, relative publication time, and number of "self-promotional" and "redundant" tweets) on clone set member popularity as measured by number of tweets.

1.6 Thesis Contributions

The contributions of this thesis are the following.

- A dataset was collected containing information for 25 clone sets each with 2 - 15 clone set members. Each clone set corresponds to a different news story, with the clone set members corresponding to identical or nearly identical versions of that news story posted on different news sites. For each news story version (clone set member), the tweets referencing that news story version were collected, as well

as information concerning the tweeters.

- The collected data was analyzed with respect to its basic popularity characteristics. It was found that both the number of tweets for a clone set member (version of a news story published on some Web site), and the elapsed time from its first tweet to the times of subsequent tweets, appear to have light-tailed distributions, reflecting the ephemeral popularity of news stories. The elapsed time distribution appears to have a similar form as the exponential distribution. With respect to the characteristics of the tweeters for a clone set member, it was found that the distribution of their maximum number of followers also appears to be light-tailed, with a similar form as the exponential distribution over most of its range.

- The possible impacts of various content-agnostic factors on clone set member popularity were investigated. Both the total number of tweets and the total number of unique tweeters were investigated as possible measures of popularity. A strong correlation (as measured by the Spearman correlation coefficient) was found between the popularity of a clone set member and the maximum number of followers of its tweeters, as well as with the number of self-promotional tweets (from a Twitter account associated with the news site). Relative publication time, and overall site popularity (at least among similarly-prominent sites), in contrast, appeared to have weaker correlations with popularity. These results suggest that in the Twitterverse, popularity of online news content is possibly strongly impacted by having influential tweeters, and less so by factors such as the "first mover advantage" and the overall news site popularity. At a nutshell, it could be summarized as "On line news popularity is influenced by content agnostic factors that emerge from social network dynamics". More detailed analysis requires a larger data set.

1.7 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 describes related work concerning Twitter data characterization, tweet propagation, and the factors impacting the popularity of online content. Chapter 3 describes the available Twitter APIs that can be used for data collection, and outlines the data collection methodology used in this thesis. This chapter also presents a summary of the collected data set. Chapter 4 presents the results from characterization of the collected data, while Chapter 5 describes the results from analysis of several content-agnostic factors that could impact the online content popularity. A summary of the thesis and of its contributions, and a discussion of future work, are presented in Chapter 6.

Chapter 2

Related Work

2.1 Twitter Data Characterization

Different types of crawling systems have been used to collect data from Twitter. The "whitelist" is the list of services which are allowed to make very higher number of requests per hour. Previously until February 2011, Twitter allowed whitelisted accounts [1] to issue up to 20,000 Twitter API queries per hour. At that time it was possible to collect a vast amount of Twitter data, which could than be analyzed to investigate research questions of interest. Later on, Twitter imposed several more restrictive limitations which make it difficult to collect large amounts of data and divert researchers to other systems or more specific and limited data collection in Twitter. Bosnjak *et al.* [5] divided Twitter data collection systems into two categories according to whether they make use of whitelisted accounts or not. They built an open source Twitter crawler for collecting Twitter data concerning human relations and communications. This crawler can continuously collect data from a particular user community while respecting Twitter query rate limits.

Using whitelisted accounts Kwak *et al.* [21] performed one of the first large crawls of the Twitter social network, to study its topology and its power for information sharing. Their crawl was carried out from July 6th to 31st 2009, and used 20 whitelisted machines, on each of which was imposed a limit of 10,000 queries per hour. Using breadth-first search, they started their crawl from Perez Hilton who had over one million followers at that time. Additional searches were conducted to collect data from users who might not be reachable from that starting point, by crawling profiles of users who tweeted concerning one of Twitter's identified "trending topics" during the measurement period. Such tweets were also collected. They crawled in total 41.7 million users, 1.47 billion social relations and 106 million tweets.

By analyzing this data they found a follower distribution that was not power law, and had a short effective social network diameter, with a low reciprocity which they considered to be deviations from the known characteristics of human social networks [29]. Specifically, they found that only 22.1% of the user pairs with any link between them had a reciprocal relationship (following), whereas for the other 77.9% there was only a link in one direction. The authors suggested that such one-way relationships make Twitter more of the information sharing medium than a social network. The average social network path length between two Twitter users was estimated to be only 4.12, which is lower than has been reported for other social

[1] http://readwrite.com/2011/02/11/twitter_kills_the_api_whitelist_what_it_means_for; accessed on 11 January - 2014

networks.

Homophily is the tendency of an individual to associate with others of the same kind. Kwak *et al.* also studied homophily with respect to reciprocated relationships. They found that users in reciprocated relationships tended to live in similar time zones, particularly users with at most 50 reciprocated relationships. Further, there was found to be some positive correlation (for users with at most 1000 followers), between the number of followers of a user i and the number of followers of users with which i had a reciprocal relationship. In another analysis, Kwak *et al.* compared different measures of user influence. They found that ranking the users according to the number of followers gave similar results to ranking using the Page Rank algorithm used by the Google Web search engine to rank web sites. On the other hand, ranking according to the number of retweets of a user's tweets gave a quite different ranking, indicating a difference between the influence inferred from number of followers and the popularity of one's tweets. With respect to the impact of retweets, it was found that the average number of additional users receiving a tweet owing to retweeting was independent of the number of followers of the original tweeter (for tweeters with up to about 1000 followers).

Perera *et al.* [32] studied the temporal behaviour of messages arriving in a social network. They developed a software architecture using Python and MySQL in conjunction with Twitter APIs to obtain tweets sent to specific users. They extracted the user id and exact time stamp for each tweet. They used this architecture to characterize the interarrival times between tweets, the number of retweets and the locations of the users, for tweets and retweets sent to U.S. president Barak Obama for a period of 6 days from the 14th to 20th of August, 2010. An exponential probability density function was fitted to model the interarrival times of tweets with 97% accuracy (measuring accuracy in terms of root-mean square error) and a geometric probability mass function was used to model the number of retweets with 93% accuracy. According to their study, the arrival process of tweets sent to a user could be modeled as a Poisson process.

Sakaki *et al.* [39] investigated how Twitter could be used to detect the occurrence of important events in the real world (such as earthquakes). They devised a method for tweet classification to identify tweets associated with a target event, and a method for estimating the event location from the locations of the tweeters. As an application of their methods, they developed a warning system for earthquakes in Japan. They found that their approach can detect an earthquake with high probability: 96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more that occurred during a two month period in 2009 were detected by their system. Their system allows much faster notification to the registered users of their system than the announcements that are broadcast by the JMA.

Social networks can also be used for finding answers to different questions from their users. Sometimes people turn to their social network to fulfill their information needs. In some social networks, groups are frequently created where particular types of information can be shared among the group users. People interested in a particular issue can join the corresponding group and get valuable suggestions from group users with expertise on that issue.

Paul *et al.* [31] conducted a study of question asking and answering (Q&A) behaviour on Twitter. They

obtained 1.2 million tweets from the public Twitter stream and processed them to identify tweets containing questions. They removed the tweets or retweets directed to specific users, tweets containing a URL link, tweets containing obscene words, non-English tweets, and finally removed tweets that did not contain a question mark since Morris *et al.* [27, 28] had found that a question mark is contained in most questions asked on social networking sites. They used Amazon Mechanical Turk [2] to identify 1152 questions from the remaining tweets. It was found that rhetorical questions were the most popular type of question followed by factual questions and polls. The most popular topic for questions was entertainment. A significant number of personal and health related questions were also identified although they had a low response rate. In general, the majority of questions didn't get even one response, and those that did get a response typically only received one or two responses. The low response rate could be due to the high volume of tweets that users may be exposed to, particularly users following many other people, which can result in questions getting buried in tweet streams. The authors suggested that this problem could be solved by using a separate question answering feature such as found in Facebook. Other findings included a positive correlation between the number of followers of the question asker and the likelihood of a response. However, no significant correlation was observed between the response likelihood and the number of tweets posted by the question asker or the asker's frequency of use of Twitter.

Use of a social network could be different in different countries. It is important to know how patterns of behaviour differ in different countries, which could be useful for purposes such as improving the design of the social network, for influence marketing, and for political campaigning. Poblete *et al.* [33] presented a summary of a large scale analysis of Twitter using a dataset they gathered containing 5.2 billion tweets made in 2010 from a collection of about 4.7 million users. These users were chosen to be from the ten most active countries with respect to the volume of tweets. The United States had the largest number of users, and greatest proportion of the total tweets. English was found to be the most popular language, used in 53% of tweets. Indonesia had the highest ratio of tweets to users. The authors carried out an analysis of the content of the tweets, including analysis of the sentiments they expressed. They measure the weighted average happiness level based on the algorithm of Dodds *et al.*[11]. For example, they found that the tweets from Brazilian users expressed the highest "happiness level", while the Netherlands users had the highest usage of hashtags and those from the U.S. had the most mentions of URLs. The latter property suggests that U.S. users more frequently view Twitter as a vehicle for news dissemination. Poblete *et al.* also collected the followers/followee relationships within their collection of users, and used this information to identify differences among the countries concerning social network structural properties such as reciprocity, diameter, average path length, modularity, degree distribution, and assortativity.

[2] http://aws.amazon.com/mturk/; accessed on 10 January - 2014

2.2 Tweet Propagation

Information can be spread over a social network by being forwarded/retweeted or copied/tweeted repeatedly from friend to friend. This phenomenon is a form of "information cascade", and has received much study. Various social activities tend to have different ways of forming a cascade on a social network [34].

Gruhl *et al.* [16] studied the dynamics of information propagation in environments of low-overhead personal publishing. They used a large collection of weblogs over time as their example domain. They characterized and modeled their collection at two levels. First of all, they presented a macroscopic characterization of discussion topics in terms of sustained discussion that the authors term "chatter", and short-term high intensity increases in postings called spikes. Secondly, they presented a microscopic characterization of propagation from individual to individual, drawing on the theory of infectious diseases to model the flow. The parameters of the model capture how a new topic spreads from blog to blog. Their algorithm learns the parameters of the model from the data. They applied their algorithm to real blog data by which they were able to identify particular individuals that were highly responsible for contributing to the spread of "infectious" topics. They proposed, validated, and employed an algorithm to discover the underlying propagation network from a sequence of posts, and reported on the results.

Characterization of URL propagation in Twitter is done by Galuba *et al.* [14]. They tracked 15 million URLs exchanged among 2.7 million users over a 300 hour period and measured several statistical properties of the data. They found that the Twitter follower graph is a "small world" where the mean shortest path length between two randomly-selected users is 3.61. It was also found that the tweeting frequency and the frequency of URL posting by the different users followed a power law distribution. Information cascades through the social graph were found to typically be shallow and wide having an exponentially distributed depth, with the cascade for each URL composed of smaller connected components. The diffusion delay between URL tweets in a cascade (i.e., the time from when one of the users that a user i is following tweets a URL, until user i tweets that URL) was found to be lognormally distributed with a median of 50 minutes. They also proposed a propagation model which can predict more than half of the individual future URL mentions in the test data set with a false positive rate of less than 15%.

Lerman and Ghosh [24] carried out a study on the spread of news in Digg and Twitter. Though Digg and Twitter have several differences in their user interface and functionality, they use similar ways to spread information. With both sites, users actively build up their network by adding friends, whose activity they want to follow. Users employ their network to track the tweeting or voting activity of their friends, and then by using their own tweeting and voting actions they make this information visible to their own fans or followers. Although in both sites, the number of fans/followers per user was found to exhibit a heavy-tailed distribution, Digg's social network was found to be more denser and more interconnected than Twitter's as measured by the proportion of reciprocated links and the network clustering coefficient. Activity measures of users on both sites were characterized by power law distributions with different exponents. Most votes

or tweets received by a story occurred within a single day. Another observation was that stories posted on Digg initially spread quickly through its network, but then the spread slowed significantly. On the other hand, information on Twitter spread slowly compared to Digg but continued spreading at that rate as the information aged and spread farther than did Digg stories.

Hui *et al.* [17] studied information cascades in social media that occur in response to a crisis. The main focus of their work was how messages spread among the users on Twitter during the occurrence of the crisis and what kinds of information cascades or patterns are observed. One of their major contributions was to build a model of the diffusion of actionable information in the social media that incorporates the concept of "trust" (likelihood that a tweet will be believed), which is designed to capture the network dynamics due to the actions of the individuals in response to warnings or other disaster related information. The findings of this work could be useful to information managers for improving the propagation of accurate information such as warnings to move to safety, or impeding the flow of inaccurate messages.

Rattanaritnont *et al.* [34, 35, 36] characterized cascade patterns according to user influence and posting behaviours for various topics. Their results concerning patterns of information diffusion and behaviours of participating users in Twitter could be useful for evaluating the effectiveness of marketing and publicity campaigns. They explored four measures: cascade ratio, tweet ratio, time of tweet, and exposure curve. The cascade ratio determines how much people can influence their friends, the tweet ratio determines how much people talk about each topic, the time of tweet determines how long a topic is still popular in the network, and lastly the exposure curve determines how easily people are influenced by their friends.

In all the papers of Rattanaritnont *et al.*, the data sets that were analyzed were obtained by crawling Twitter for a period of time (March 11, 2011 to July 11, 2011) immediately following a large Japanese earthquake. Their data set included 783 million tweets from 260 thousand users. They selected the 500 most frequently used hashtags from the data set and categorized them according to six popular topics: earthquake, media, politics, entertainment, sports, and idiom. They found distinctive patterns of hashtag cascade for the different topics. For example, the earthquake topic had a low cascade ratio, low tweet ratio, short lifespan, and high persistence, while the political topic had a high cascade ratio and high persistence. However, some hashtags even on the same topic had different cascade patterns. For instance, the earthquake hashtags could be divided into the hashtags directly related to the recent Japanese earthquake, the media-related hashtags, and the politically-related hashtags or the hashtags about the nuclear power plant damaged as a result of the earthquake. It was discovered that hidden relationships between topics could be revealed by the cascade patterns. Finally, their results showed that the cascade ratio and the time of tweet were the most effective measures to distinguish cascade patterns for different topics.

Instead of browsing and Internet search, people now often discover content through its posting on social networking sites, which works like word-of-mouth, where content spreads via conversations between people. Tiago *et al.* [37] presented a detailed analysis of word-of-mouth based content discovery on the Web. They analyzed the sharing of Web links (URLs) on Twitter. They found that Twitter propagation trees are wider

than they are deep. Their analysis on geo-location of users indicated that geographically close users are more likely to share the same URLs. They also found that a single URL could be spread to a large portion of the user population, in some cases to millions of users. Popular URLs could be spread by multiple disjoint propagation trees, where each of the propagation trees has a large number of nodes. Their analysis also reveled that even content published on relatively unpopular domains could be propagated to a large audience.

2.3 Factors Impacting Popularity of Online Content

2.3.1 Social Influence

Social influence can be an important factor in making content popular. Salganik *et al.* [40] attempted to experimentally measure the impact of social influence, in contrast to the actual quality of the content, on content popularity. The authors created an artificial music market comprising 14,341 participants, who were asked to rate some previously unfamiliar songs from one to five stars. There was also an option for downloading the song. Users were assigned to two different groups. In one group (the control group) users were just given the list of songs and their band names. In the other group, social influence was a factor, since users were also informed about how many times each song was downloaded by other users. The social influence factor appeared to have a large impact on song popularity. Although song popularity within the second group of users was correlated with that in the control group, popularity within the second group was highly variable, in the sense that two songs with similar popularity in the control group could have highly different popularities within the second group. When users were aware of the choices made by others, popularity became very unpredictable. Although the Salganik *et al.* study was limited to a small set of songs created by unknown bands, its conclusions about inequality and unpredictability of success due to social influence may have important implications for understanding online content popularity.

Leavitt *et al.* [22] gathered tweets, and responses to those tweets, from 12 Twitter users deemed to be highly influential, over a 10 day period in 2009. These 12 users had 15,866,629 followers and 899,773 followees, and in response to the 2,143 tweets generated by these users over the 10 day period 90,130 responses were published by other users. The authors analyzed the relative influence of the 12 users as measured by number of responses (replies, retweets and mentions). Both the absolute number of responses was considered, as well as the number of responses relative to the number of original tweets.

The social influence of "ordinary" users can also be substantial. Bakshy *et al.* [2] studied the ability of Twitter users to influence others in the sense of posting URLs in tweets which other users will then subsequently repost. For this purpose, the authors collected over one billion public tweets over a two month period in 2009. A subset of 74 million of these were identified that contained a distinct URL, and for which the "seed user" (original poster of the URL) was active in both the first and second month of the measurement period. The authors also crawled the Twitter social network starting from the set of all users who had tweeted at least one URL during the measurement period. They found that the largest cascades were generated by

users who had also been influential in the past, and who had a large number of followers. However, it was also found that prediction of which particular user or URL would generate large cascades was unreliable. Furthermore, in the context of a marketing campaign, there may be cost associated with targeting particular individuals; in particular, users may charges for each "sponsored tweet", with particularly influential users charging more. It was found that often the most cost-effective performance can be realized using "ordinary influencers" with average influence and connectivity. This is consistent with previous theoretical work done by Watts *et al.* [44], in which the strategy of trying to target highly influential users was questioned.

Cha *et al.* [9] presented an empirical analysis of influence patterns in Twitter. Using a large amount of data gathered from Twitter, including data on about 54 million users and 1.7 billion tweets, they compared three distinct measures of influence: number of followers, number of retweets, and number of mentions. The number of followers is a measure of the popularity of a user, while the number of retweets indicates how valuable others find the user's tweets. The number of mentions is a measure of how well the user is able to engage others in conversations. Comparing these three types of influence, Cha *et al.* found that among highly popular users, user rank with respect to number of retweets was highly correlated with that based on number of mentions. The correlation between either of these metrics and number of followers, again considering only highly popular users, however was found to be weak. Cha *et al.* then focused on three diverse topics that prompted considerable activity in Twitter at the time of their data collection in 2009: the Iranian presidential election, the H1N1 outbreak, and the death of Michael Jackson. They found that the most influential Twitter users with respect to one topic were typically highly influential with respect to the other two topics as well. Finally, Cha *et al.* also investigated the dynamics of an individual's influence by topic and over time, and attempted to characterize the important behaviours that make ordinary individuals achieve high influence over a short period of time.

2.3.2 Rich Get Richer

With the "rich get richer" behaviour, a popular online content item may become even more popular because of its already acquired popularity. Barabasi *et al.* [4] describe this behaviour as being potentially present in a variety of domains, including the Web. In the simplest model of "rich get richer" behaviour, additional popularity is acquired in direct linear proportion to the currently acquired popularity. "Rich get richer" behaviour can result in a power law popularity distribution, in which the probability mass function $P(n)$ is asymptotically proportional to $n^{-\alpha}$ for some constant $\alpha>1$.

Borghol *et al.* [6] collected sets of videos from YouTube, such that the videos in each set were nearly identical items ("clones") uploaded by different people at different times. The authors analyzed the content-agnostic factors that create differences in popularity among the videos within each clone set. It was found that "rich get richer" behaviour provides a good model of video popularity evolution except for very young videos.

2.3.3 Recommendation Systems

YouTube is currently one of the most popular user generated content sites, and is the third most accessed Internet site overall. Hosting a collection of hundreds of millions of videos, YouTube offers several features such as video search, related video recommendations, and front page highlights to help users to discover videos of interest. Zhou *et al.* [45] performed a measurement study of the impact on video views of the YouTube video recommendation system. They found that related video recommendations accounted for about 30% of overall views. Also, the fraction of videos for which related video recommendations were the most important source of views was higher than that for any other category of views. They found a strong correlation between the view count of a video and the average view count for its top referrer videos. They also discovered that the positioning of a video in a related video list had a substantial impact. Finally, they found that the YouTube recommendation system helped increase the diversity of videos viewed.

The issue of whether recommendation systems increase or decrease diversity was also examined by Fedler and Hosanagar [12] [13], in the context of product sales at online stores. There are two different views on this issue. One view is that recommenders help consumers discover new products and thus increase diversity. The other view is that recommenders decrease diversity by promoting already popular products. The authors explored this issue using both an analytical model and simulations. For most scenarios they investigated, recommenders were found to decrease diversity.

Zhou *et al.* [46] designed a new hybrid recommendation algorithm with the goal of improving simultaneously both diversity and accuracy. Here accuracy refers to the algorithm's ability to recommend items of interest to the user. They showed that their algorithm could be tuned to achieve its goal, through experiments using three datasets: a randomly-selected subset of the dataset provided for the Netflix Prize, a music ratings dataset, and a dataset obtained from a social bookmarking website.

2.3.4 Comparisons among Factors and Popularity Prediction

Popularity of online content in social media may be difficult to predict. Early or late popularity can be measured in terms of the number of views or votes of the contents which are somehow correlated [25, 42] but the actual factors behind making one piece of content item popular may be difficult to determine. It could be content's inherent quality, consumers' response about the content, social influence, or other factors that could play the important role here [25]. Salqanik *et al.* [40] worked on popularity of cultural artifacts and experimentally showed that the popularity of content is only weakly related to its inherent quality; social influence leads to an uneven distribution of popularity that makes it unpredictable.

Bandari *et al.* [3] used Twitter tweets referencing news articles in order to study the problem of how to predict online news story popularity. They used an API for news feed aggregator "Feedzilla",[3] and collected news feeds containing all news articles published online in the second week of August, 2011. The feed of the

[3] www.feedzilla.com

article contained the title, short summary of the article, URL, time stamp and publisher of the news. After filtering they got around 42,000 suitable items. They then used a Twitter search engine called Topsy,[4] to find the number of times each URL was tweeted or retweeted on Twitter. They focused on four factors: publisher, category (e.g. sports vs. politics), subjectivity versus objectivity in the language, and the mentioned named entities (e.g. Obama, Bieber).

Their study found that these four features were sufficient to predict whether a news item would fall into a "low-tweet" (1 - 20), "medium-tweet" (20 - 100), or "high-tweet" (more than 100) class, with 84% accuracy. It was also found that in case of the number of retweets, some blog sites such as Mashable and the Google blog had higher popularity than the conventional popular news sites.

Bandari *et al.* also discovered that the most important predictor of popularity in their study, among the four factors they considered, was the source of the article. This finding suggests that people are substantially influenced by the news source that published the article. The category factor did not perform well as a predictor, which the authors suggested may be because Feedzilla's assigned categories were overlapping. They also found that the subjectivity of the language, and using mention in the tweets, did not provide much improvement in prediction of news popularity in Twitter.

The inherent quality of a content item also plays a prominent role in determining that item's popularity in the Internet. There are also some "content-agnostic" factors, which could have great impact in making one online content item more popular than others. Borghol *et al.* [6] collected sets of videos from YouTube, such that the videos in each set were almost identical in content but were uploaded by different people at different times. They manually identified 48 clone sets; each of them contained between 17 and 94 clone members, with a median number of clone members per clone set of 29.5 and a total over all clone sets of 1761 videos.

They developed a rigorous clone-based analysis framework to control bias introduced while studying video popularity. Using their clone-based methodology, the authors made several findings. First of all, inaccurate conclusions may be drawn when attempting to study factors impacting video popularity without accounting for differences in content. Second, controlling for video content, scale-free rich-get-richer behaviour was observed, in which additional views were acquired in linear proportion to existing total view counts, except for very young videos. Third, they found that other content-agnostic factors can help explain various other aspects of the popularity dynamics. In the case of a newly uploaded video, the uploader's social network and the number of keywords can have a strong impact on the video's future popularity. Finally, they showed the existence of a first-mover advantage for the online videos, where an early uploaded video often gets more popularity than later uploaded videos with the same content.

Lerman *et al.* [25] developed a model of social voting on Digg that describes the popularity evolution over time for news stories submitted to Digg. Their model is used to predict whether a newly posted story is promoted to the Digg front page based on the early reaction of Digg's users; they also used that model to predict the posted story's ultimate popularity. The authors claim that their model is able to separate the

[4] http://topsy.com

impact of news story quality on popularity, from the impact of social influence.

Szabo *et al.* [42] presented a method for predicting the long-term popularity of online content from early measurements of user accesses. For two content sharing sites, YouTube and Digg, they attempted to model the accumulation of views and votes on content offered by these sites based on initial data. In the case of Digg, the authors claimed that it is possible to forecast content popularity 30 days ahead with good accuracy by measuring accesses to the content during its first two hours after submission to the site. On the other hand, YouTube videos need to be measured for 10 days to attain the same performance. These different times are needed to predict their popularity due to differences in content consumption rate between the two portals. Digg news stories quickly become outdated, typically within two days, while YouTube videos can remain popular for long periods of time after they are initially uploaded. The authors used a simple logarithmic prediction method and the number of votes/views to measure the correlation between the interest shown early in the content lifetime and the final popularity of the content. They found that predictions are more accurate for content for which attention decays quickly, whereas predictions for content with long-lived popularity will be prone to larger errors.

Blogs are another source of interesting online content. Blogs provide commentary, news, or content on a particular subject. Many blogs have an interactive format, and some blogs become very popular. In order to find out the factors behind blog article popularity, Kim *et al.* [20] first collected 816 articles from March 15, 2011 to March 20, 2011. Using this data, they derived a popularity prediction model. Then they tested their model with 1157 articles collected from May 4 to May 10, 2011. Their article collection source was "SEOPRISE", a well-known political discussion blog in Korea, which at the time of data collection had more than 50,000 visitors per day.

Kim *et al.* defined the "saturation point" of an article as the time at which its popularity growth levels off. They derived a regression model to predict the popularity "temperature" of an article at its saturation point. Here "temperature" refers to one of four classes defined according to the range in which the hit count falls. The authors could predict the popularity temperature of over 86% of the articles in the "explosive", "hot", "warm", and "cold" categories from the first 30 minutes to 60 minutes of their lifetime and over 90% after 70 minutes. For the most popular "explosive articles" popularity was very difficult to predict.

Asur *et al.* [1] showed the application of social media for forecasting box office revenues of movies before their publishing. They developed a linear regression model using 3 million tweets from Twitter which could predict the box-office revenues of movies in advance. Their model outperformed the results of the Hollywood Stock Exchange. They found that there is a strong correlation between the attention given to a forthcoming movie and its upcoming future popularity. They also studied the initial reactions of the users through Twitter about a released movie and analyzed how those reactions could affect its upcoming popularity.

Online comments can also be used for predicting the popularity of their online content. Analysis on the possible use of comments for predicting popularity of weblogs was provided by Mishne *et al.* [15]. They studied the relationship between the weblog comments and the posts and showed that the comments provide

a good indication of the popularity for a weblog.

Tatar *et al.* [43] showed how users' comments over a short observation period could be used for prediction of the long term popularity of newspaper articles. They conducted their analysis on several data sets obtained from the Web site for 20Minutes, a free daily newspaper published in the main cities in France. At the time of the authors' work, 20Minutes was the most popular daily journal in France with an average of 2,675,000 readers per day, and their website was the 4th most popular online press site in France with more than 4 million unique visitors and 73 million site visitors per month [43]. Their data sets included more than 4 years of articles and associated comments. After removing articles for which comments had been disabled, their data sets contained approximately 260,000 articles and 2,500,000 comments. They used a simple linear regression model applying the sample initial data to predict the ultimate popularity of the news article. They found that their approach could rank articles with respect to their future popularity based on their total number of comments achieved during their first day after publication.

Lee *et al.* [23] also proposed a method for predicting the popularity of online content. Their work related the popularity of an online content item to explanatory (risk) factors, and used survival analysis to predict its popularity. They used a Cox proportional hazard regression model with explanatory variables that divides the distribution function of the observable popularity metric into two components. First, one can be described by the given set of explanatory factors; the other is a baseline distribution function that integrates all the factors not taken into account. They validated their approach using two different online discussion forums: Dpreview and MySpace. Dpreview is one of the largest online discussion groups providing news and discussion forums about many models of digital cameras. MySpace is an online social networking service.

They modeled two different popularity metrics, the discussion thread's lifetime and the number of comments per thread, and showed that their approach can predict the lifetime of threads from Dpreview and MySpace after an initial 5-6 days observation. On the other hand the number of comments could be successfully predicted after observing a thread during its first 2-3 days.

2.4 Summary

Chapter 2 talks about the papers that deals with the Twitter data characterization, Tweet propagation and the factors that impacting the popularity of online content. This chapter gives close look on different content agnostic factors like social influence; rich get richer behavior; recommendation system used in various on line content providing sites and does comparisons between the factors. This chapter also deals with the factors which could be used to predict the popularity of online content. From all of those analyses it is found that popularity of the online content varied upon the types of content. Some of the online content propagate very fast through the social networking site like Twitter, get popularity with quick succession, and besides, their popularity falls quickly i.e. Twitter tweets, Digg post etc. Some of the content's popularity resilient long

period of time like YouTube video. It is found that, in most of the cases, to make one online content popular other than the content's inherent quality some other content agnostic factors play a vital role.

General trend of the related works in this chapter gives the idea about the characterization of distinctive kinds of online content as well their popularity measurement. Thus lead us to characterizing and investigating the content agnostic factors of the on line news content in Twitterverse.

Chapter 3

Data Collection

This chapter describes the methodology used in the thesis for collecting data from Twitter, as well as basic properties of the collected data. Section 3.1 outlines Twitter APIs that were evaluated for possible use, while Section 3.2 presents the adopted data collection methodology. The basic properties of the collected data are described in Section 3.3.

3.1 Twitter Data Collection APIs

Different APIs are available for collecting data from Twitter. The APIs that were investigated for possible use in the data collection for this thesis are described in the following subsections.

3.1.1 Search API

The Twitter Search API[1] returns tweets matching the provided search parameter. A URL can be provided in this parameter. With the version of the Search API available from January 2011 through the time of the thesis data collection, at most 1500 tweets can be obtained in response to a query [38]. Results are returned in the form of multiple pages, where one can access a particular page using the 'page' parameter. The 'rpp' parameter can be used for controlling the number of outputs per page. At most 100 results can be obtained per page, and a maximum of 15 pages are returned. This version of the Search API was retired on May 7, 2013.[2]

Every tweet has a unique ID number. One can restrict search results to tweets whose IDs fall within a range of interest by applying parameters 'since_id' and/or 'max_id'. It is also possible to filter search results according to the date of the tweet, using the 'until' parameter, as well as according to the language or to the tweeter location. An example Search API query, which could be submitted using any browser, is as follows:[3]

`http://search.twitter.com/search.json?q=blue%20angels&rpp=5&include_entities=true&result_type=mixed`

This query will return a single page containing at most five tweets, each with "blue angels" in the text portion of the tweet as requested with the query parameter 'q'. The 'result_type' parameter is used to specify

[1] https://dev.twitter.com/docs/api/1/get/search; accessed on 10 February - 2013
[2] https://dev.twitter.com/blog/api-v1-retirement-final-dates; accessed on 10 April - 2013
[3] https://dev.twitter.com/docs/api/1/get/search; accessed on 10 February - 2013

the type of tweets that are of interest. Options for this parameter are 'mixed', 'popular' and 'recent', where 'mixed' is set as default. The 'include_entities' parameter has default value 'true', in which case the output for each tweet includes a structure with various metadata, including for example the shortened form of any URL referenced in the tweet as well as the expanded form. A summary of Search API parameters follows:[4]

- geocode: This parameter can be used to restrict the search to tweets from tweeters located within a given radius of the specified latitude/longitude. The location is preferentially taken from the tweet itself (if the tweet is "geotagged" with the tweeter location), but will fall back to the location specified in the Twitter profile of the tweeter. The parameter value can be given as "latitude, longitude, radius".
- include_entities: When set to 'true' the output for each tweet includes a structure with various metadata concerning the tweet.
- lang: This parameter can be used to restrict the search to tweets in a given language.
- max_id: This parameter can be used to limit the returned results to tweets with an ID less than or equal to the specified ID (i.e., tweets that are older than the tweet with the specified ID, as well as that tweet itself).
- page: It defines the page number of results one wants to access. The Search API returns up to 15 pages of results, numbered from 1 to 15.
- q: This parameter gives the search query that tweets will be matched against. It can take a URL link or any other keyword. It supports UTF-8, URL-encoded search queries of maximum length 1,000 characters, including characters for operators such as logical "and" and "or". This is the only mandatory parameter.
- rpp: This parameter specifies the number of tweets per page, up to a maximum of 100 tweets.
- result_type: This parameter specifies the type of tweets of interest, and can be set to 'popular', 'recent' or 'mixed'.
- since_id: This parameter can be used to limit the returned results to tweets with an ID greater than the specified ID (i.e., tweets that are more recent than the tweet with the specified ID).
- until: When this parameter is used, only tweets generated before the given date will be returned. The date is specified in the form YYYY-MM-DD.

The Search API returns data in a raw JSON (JavaScript Object Notation) format. JSON is a lightweight data interchange format. It is based on a subset of the JavaScript Programming Language.[5] Python has a built in functionality for accessing JSON objects.

[4] https://dev.twitter.com/docs/api/1/get/search; accessed on 10 February - 2013
[5] http://www.json.org/

```
[
 {
  "iso_language_code": "it",
  "profile_image_url_https": "https://si0.twimg.com/profile_images/2567853384/2062_normal.jpg",
  "from_user_id_str": "796183574",
  "text": "RT @decobupiwy: http://t.co/zRKGSKwzuV Argentina's Bergoglio elected Pope",
  "from_user_name": "Peggy Whisler",
  "profile_image_url": "http://a0.twimg.com/profile_images/2567853384/2062_normal.jpg",
  "id": 312915706394054657,
  "source": "&lt;a href=&quot;http://twifarm14.ru&quot;&gt;Servers blitz&lt;/a&gt;",
  "id_str": "312915706394054657",
  "from_user": "ehusagyt",
  "from_user_id": 796183574,
  "geo": null,
  "created_at": "Sat, 16 Mar 2013 13:18:12 +0000",
  "metadata": {
   "result_type": "recent"
  }
 }

{
  "iso_language_code": "en",
  "profile_image_url_https":
"https://si0.twimg.com/sticky/default_profile_images/default_profile_6_normal.png",
  "from_user_id_str": "1052191447",
  "text": "BBC News - Argentina's Bergoglio elected Pope http://t.co/2YUrnncXGq",
  "from_user_name": "Macfalad1",
  "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_6_normal.png",
  "id": 312063938143604738,
  "source": "&lt;a href=&quot;http://twitter.com/tweetbutton&quot;&gt;Tweet Button&lt;/a&gt;",
  "id_str": "312063938143604738",
  "from_user": "Macfalad1",
  "from_user_id": 1052191447,
  "geo": null,
  "created_at": "Thu, 14 Mar 2013 04:53:35 +0000",
  "metadata": {
   "result_type": "recent"
  }
 }
...
```


Figure 3.1: Sample Output from the Twitter Search API

Figure 3.1 shows some sample output from the Twitter Search API. Information for two tweets is shown. Here, 'iso_language_code' gives the language in which the tweet is made, while 'text' gives the text portion of the tweet which can include shortened URLs as well as retweet and mention notifications. Every tweet has a unique ID, which is given in the 'id' field. 'From_user' and 'from_user_id' give the screen name and ID of the tweeter, respectively. The 'created_at' field gives the time when that tweet was made. Although tweets can be made from different parts of the world in different time zones, the times reported in the 'created_at' field are consistently reported in UTC, as indicated by '+0000' at the end of the field value.

3.1.2 Streaming API

With the Streaming API, a long-lived HTTP connection is maintained with a Twitter server, on which matching a specified 'track' parameter are returned in real-time as they are made. The client sends a single HTTP request, which the Twitter server responds to incrementally as each matching tweet occurs. The most common use of the Streaming API is for tracking tweets containing a particular hashtag. The Streaming API can be conveniently accessed using cURL,[6] a command line tool that can be used to transfer data using HTTP (as well as for other purposes). Below is an example of using the Streaming API via cURL to obtain tweets containing a reference to the URL of a Yahoo News story concerning an explosion at a Texas fertilizer

[6] http://curl.haxx.se/

plant in April 2013.

News link: http://news.yahoo.com/official-12-bodies-recovered-texas-blast-135334480.html

cURL command:

curl -u nazmul26:nazmul "https://stream.twitter.com/1/statuses/filter.json" -d "track=yahoo 135334480" >>yahoo26.txt

The '-u' option allows a user name and password to be given, which is used for server authentication. In this case the valid user name and password of a Twitter user are provided. The URL given is that of the Twitter Streaming API server. With the '-d' option, the specified data, in this case the setting of the 'track' parameter, is sent in an HTTP POST request to the server. The track parameter 'yahoo 135334480' matches tweets containing both 'yahoo' and '135334480', which will correspond to those tweets including a reference to the Yahoo News story URL. Note that Twitter shortened URLs inside tweets are expanded before matching against the 'track' parameter.

Some news sites do not use identification numbers for their news stories as in the above example. The URL for the Guardian's news story for the same event, for example, is as follows:

http://www.guardian.co.uk/world/2013/apr/18/texas-explosion-fertiliser-plant-live

In this case, track = "guardian texas explosion live" could be one option for the tracking parameter. Note that a space acts as the AND operator, so this would match tweets containing all four words, which would, with high likelihood, be tweets containing a reference to the Guardian news story URL.

Figure A.1 in appendix shows an example of the data returned for a single tweet from the Streaming API. JavaScript beautifier software[7] is used to make the output more easily readable. From the figure it is seen that the Streaming API returns data such as the tweet's creation time, tweet's ID, text of the tweet, hashtag contained in the tweet, and number of retweets. The Streaming API also returns much information regarding the tweeter, including the tweeter ID, user name, screen name, number of friends (other users that the user is following), number of followers, listed count (the number of Twitter public lists on which the user appears), time zone, language, and other detailed profile information.

3.1.3 Get Users/Show API

Detailed information on a Twitter user can be found by using the Get Users/Show API. The user can be specified using either the user's Twitter ID, or the user's screen name. An example of using the Get Users/Show API to obtain information on a Twitter user with screen name of 'BBCBreaking' is as follows:[8]

`https://api.twitter.com/1/users/show.json?screen_name=BBCBreaking&include_entities=true`

This API has a rate limitation. At most 150 requests are permitted per hour for unauthenticated calls from a single IP address. By using authenticated OAuth calls at most 350 requests per hour can be served

[7] http://jsbeautifier.org/

[8] https://dev.twitter.com/docs/api/1/get/users/show

using a single oauth_token in the requests.[9] Unauthenticated calls were used for the data collection carried out for this thesis.

The parameters that can be used in Get Users/Show API queries include the following:

- user_id: This parameter specifies the ID of the user one wants information for.
- screen_name: This parameter specifies the screen name of the user one wants information for. Either user_id or screen_name is sufficient for this method.
- include_entities: When set to 'true', the output includes a structure giving the URLs that appear in the user's profile.

Figure A.2 in appendix shows the output from the Get Users/Show API query given above for the Twitter user 'BBCBreaking'. The information returned includes the date and time at which the user created their Twitter account, the user's number of friends, the user's number of followers, the number of tweets that the user has "favourited", retweet count, language, listed count, ID of the user, location of the user, as well as other information such as profile details.

3.2 Adopted Data Collection Methodology

3.2.1 Finding Clone Sets

The first step in creating a new clone set was to search for a popular news story on a well-known news site. Such stories might be featured on the home page of the news site, or might have a substantial number of initial tweets as reported by the site. Only stories that acquired a significant number of tweets during the initial hours after publication were considered further. Once such a news item was found, a search was then made for clones of that news story version, i.e. identical or nearly-identical stories published on different news sites. In order to find such clones, two approaches were followed. The first approach was to carry out a Google search on the title of the news story as used on the initial news site on which the story was first found. The second approach was to manually search for the same news content on other top-ranked news sites. A listing of top 20 news sites was used for this purpose.[10]

A backtracking approach was also applied when searching for clone set members. After getting results from the first search, some of the web links were explored to see the news content. When quite different titles were found for the same news content, these different titles were used in searches to once again find other related links. The reason for this strategy is that when picking an initial news story version and searching for related content using the title of that news story version, the search results were strongly biased towards results with an exact match of the title. Some news story versions with distinctive titles could therefore be missed. So the backtracking approach was applied to address that problem.

[9] https://dev.twitter.com/docs/rate-limiting/1
[10] http://news.nettop20.com/

3.2.2 Use of Twitter APIs

Initially an attempt was made to use the Streaming API, but problems were encountered with the connections timing out. The most common use of the Streaming API is for tracking tweets containing a particular popular hashtag, where there is a steady stream of such tweets. For the data collection needed for this thesis, however, it was necessary to track tweets containing a reference to a news story URL. There can be long periods of time when no such tweets are made, particularly after many hours have elapsed since the publication of the story. When attempting to use the Streaming API for tracking, initially there was sufficient feedback from the server to keep the connection open, but later, the amount of feedback dropped causing an HTTP connection timeout error. A clone member could get a significant number of additional tweets after its connection is closed. Since these connection timeouts were caused by the Twitter server, and the length of time a connection stayed open until timeout varied for different clone members according to their popularity, consistent data collection could not be achieved. Due to this connection timeout problem, the data collection methodology was changed.

In the adopted methodology, only the Search API was used for collecting tweets. The main challenge in using this API is the 1500 tweet limit on its returned results. This was addressed using the Search API 'until' and 'since_id' parameters. Four days after the publication of a news story, the Search API was used to determine whether there had been less than 1500 tweets referencing that URL. If so, than a sequence of Search API calls, varying the 'page' parameter value from 1 to 15, was used to obtain the information for all of these tweets. If a clone member received 1500 or more tweets, the 'until' keyword was used to collect the tweet information up to a precise date. The maximum tweet id was obtained from this collected data, which was used for the 'since_id' parameter value for a subsequent Search API call. This subsequent API call returns up to 1500 tweets from that precise point. If 1500 tweets were returned, the process could be repeated. By this approach, it was hoped that all tweets referencing each clone member could be collected. Therefore, it was not necessary to use the Streaming API. Later, all the tweet information for a particular clone set member was accumulated into a single file.

Using the Search API, information regarding tweets is collected, but not any details concerning the users that made those tweets. In order to collect detailed user information concerning the tweeters, the Get Users/Show API was used. A script was made that reads accumulated results from Search API calls, extracts the screen name of each tweeter one at a time in order, and passes that name as an argument in calling the Get Users/Show API. An appropriate time delay of 25 seconds was used between successive calls to the Get Users/Show API in order to overcome the rate limitation problem of that API.

Part way through the data collection it was discovered that sometimes a Search API query does not return 1500 tweets, even though there are more than 1500 tweets that match the search parameter. In such cases, it was observed that more than 1450 but less than 1500 tweets are returned. This creates confusion regarding whether or not further Search API calls using 'until' and 'since_id' are needed. This problem was addressed by making at least two consecutive Search API queries with the same search parameter. It was found that if

these consecutive queries returned the same results, then all tweets were being correctly returned. If not, then the total number of tweets was really more than 1500, and the different results arise from sampling performed by the Twitter server. In that case, the procedure described above for clone set members with greater than 1500 tweets was applied. The adopted data collection methodology appeared to be successful, but because of the lack of documentation concerning the precise behaviour of the Search API, it is not absolutely certain that there is no missing data.

3.2.3 Data Parsing

The results from the two APIs used for data collection, Search API and Get Users/Show API, yield raw data files that have different formats and contain different information. Parsers were built using Python for extracting the most relevant information from the raw files and putting it into comma separated value (CSV) files. The following fields were extracted from the Search API raw data files:

- created_at: The date and time at which the tweet was made.
- from_user_name: The user name of the tweeter.
- id_str: Universal ID of the tweet.

The following fields were extracted from the Get Users/Show API raw data files:

- created_at: The date and time at which the user opened their account.
- favourites_count: The number of tweets that the user has favourited.
- friends_count: The number of other users that the user is following.
- followers_count: The number of followers of the user, i.e. number of other users who subscribe to the user's tweets.
- screen_name: The screen name of the user.
- lang: The language used by the user.
- listed_count: The number of Twitter public lists that the user appears on.
- retweet_count: The total number of retweets made by the user.
- statuses_count: The total number of tweets and retweets that have been made by the user.
- time_zone: The time zone of the user.
- location: The location of the user.

These fields were identified as being potentially of use for later analysis, but not all were in fact used.

3.3 Collected Data Set

In total, data for 25 clone sets was collected, each clone set corresponding to a different news story. Each clone set has from 2 to 16 clone set members, corresponding to identical or nearly identical versions of the news story published on different news sites.

Table 3.1 lists some basic information concerning the clone sets for which data was collected. It gives the number of clone set members for each clone set, the publication date of the news story and the title of the story as given on one of the news sites. Although there were often variations in title among the clone set members, here only one representative title is given. From the news titles given for the different clone sets, it can be observed that the 25 clone sets cover a wide range of types of news including political, economic, technical, scientific and environmental. All stories were published within a time period from late January to the end of April, 2013.

Table 3.2 lists all of the news sites with at least one clone set member. In addition to the news site name and URL, the table also gives the name by which the news site is referred to later in the thesis, if needed. Note that a wide variety of sites are represented. Table 3.3 gives the number of tweets of various types that were collected for each clone set member. In particular, the table reports the total number of tweets, the total number of unique Twitter users making tweets ("unique tweeter tweets"), and the number of tweets from Twitter accounts that appeared to be associated with the news sites itself ("promotional tweets").

Table 3.1: Clone Sets for which Data was Collected

Clone Set	**Number of Members**	**Publication Date**	**Title**
1	6	January 25, 2013	Rob Ford wins appeal to remain Toronto Mayor; Supreme Court challenge promised.
2	6	January 31, 2013	Student shot at Atlanta middle school; suspect in custody.
3	6	February 11, 2013	Pope Benedict says he will resign February 28
4	5	February 19, 2013	Microsoft made mistakes in early mobile strategy: Bill Gates
5	4	February 20, 2013	Agency checks water after body found in hotel tank
6	6	February 27, 2013	Iran upbeat on nuclear talks, West still wary
7	8	February 28, 2013	Bangladesh sentences Jamaat leader to death for 1971 war crimes
8	8	March 2, 2013	Cuts in place, Obama and GOP brace for next fight
9	11	March 3, 2013	Queen Elizabeth II hospitalized with stomach bug
10	15	March 5, 2013	Hugo Chavez, fiery Venezuelan leader, dies at 58
11	8	March 7, 2013	Russian scientists may have found new life under Antarctic ice
12	7	March 10, 2013	Afghan president accuses U.S., Taliban of collusion
13	16	March 13, 2013	New Pontiff Is Pope Francis of Argentina
14	14	March 16, 2013	Swiss tourist gang-raped in central India
15	2	March 16, 2013	13 dead, dozens hurt in Mexico fireworks explosion
16	13	March 23, 2013	Senate passes $3.7 trillion budget, its first in 4 years
17	14	March 23, 2013	Pope Francis tells Benedict: “we’re brothers”
18	9	April 2, 2013	North Korea to restart nuclear reactor in weapons bid
19	5	April 3, 2013	$700m missing in Katrina
20	16	April 4, 2013	South Korea: North moved missile to East Coast
21	6	April 9, 2013	US Navy shows laser shooting down a drone
22	3	April 11, 2013	Pentagon: NKorea could launch nuclear missile
23	13	April 18, 2013	Deadly explosion hits Texas fertiliser plant
24	13	April 20, 2013	Strong quake hits China; 156 dead, more than 5,500 injured
25	4	April 22, 2013	Canada thwarts “Al Qaeda-supported” passenger train plot

Table 3.2: Clone Member News Sites

Name on Website	**Name Used in Thesis**	**URL**
ABC Local	ABCLocal	www.abc.net.au/local
ABCNews.com	ABC	www.abcnews.go.com
Aljazeera	Aljazeera	www.aljazeera.com
Bakersfield.com	Bakersfield	www.bakersfield.com
BBC News	BBC	www.bbc.co.uk/news
Businessweek	Businessweek	www.businessweek.com
CBC News	CBC	www.cbc.ca/news
CBS Local Sites	CBSLocal	www.cbslocal.com
CBS News	CBS	www.cbsnews.com
Channel NewsAsia	Channelnewsasia	www.channelnewsasia.com
Chicago Tribune	Chicagotribune	www.chicagotribune.com
CNBC	CNBC	www.cnbc.com
CNN.com	CNN	www.cnn.com
CNN Political Ticker	Politicalticker	politicalticker.blogs.cnn.com
Dallasnews	Dallasnews	www.dallasnews.com
Daily News America	Nydailynews	www.nydailynews.com
Daily Record	Dailyrecord	www.dailyrecord.co.uk
Dawn	Dawn	www.dawn.com
Examiner	Examiner	www.examiner.com
Fox Business	Foxbusiness	www.foxbusiness.com
Fox News	Fox	www.foxnews.com
GlobalPost	Globalpost	www.globalpost.com
Indiatimes	Indiatimes	www.indiatimes.com
IBTIMES.com	IBTimes	www.ibtimes.com
Las Vegas Sun News	Lasvegassun	www.lasvegassun.com
Los Angeles Times	LATimes	www.latimes.com
MiamiHerald.com	Miamiherald	www.miamiherald.com
Milwaukee Journal Sentinel	Jsonline	www.jsonline.com
MSN	MSN	www.msn.com

Name on Website	**Name Used in Thesis**	**URL**
National Post	Nationalpost	www.nationalpost.com
NBC Los Angeles	NBCLosangeles	www.nbclosangeles.com
NBC News	NBC	www.nbcnews.com
NBC Politics	NBCPolitics	www.nbcpolitics.nbcnews.com
New York Post	NYPost	www.nypost.com
New York Times	NYTimes	www.nytimes.com
News	News	www.news.com.au
Newsday	Newsday	www.newsday.com
NJ.com	NJ	www.nj.com
NPR : National Public Radio	NPR	www.npr.org
Pat Dollard	Patdollard	www.patdollard.com
PBS : Public Broadcasting Service	PBS	www.pbs.org
PCWorld	PCWorld	www.pcworld.com
PennLive.com	Pennlive	www.pennlive.com
People.com	People	www.people.com
POLITICO.com	Politico	www.politico.com
Prothom Alo	Prothomalo	www.prothomalo.com
Reuters Group PLC	Reuters	www.reuters.com
San Diego News, Local, California and National News	Utsandiego	www.utsandiego.com
Sky News	Sky	www.news.sky.com
SMH News Online & World News	SMH	www.smh.com.au
Stuff.co.nz & World News	Stuff	www.stuff.co.nz
Swampland	Swampland	swampland.time.com
The Associated Press	AP	www.ap.org
The Baltimore Sun	Baltimore	www.baltimoresun.com
The Christian Post	Christianpost	www.christianpost.com
The Christian Science Monitor	CSMonitor	www.csmonitor.com
The Daily Caller	Dailycaller	www.dailycaller.com
The Dallas Morning News	Dallasnews	www.dallasnews.com

Name on Website	**Name Used in Thesis**	**URL**
The Economic Times	Economictimes	www.economictimes.indiatimes.com
The Guardian	Guardian	www.theguardian.com
The Hill	Thehill	www.thehill.com
The Huffington Post	Huffington	www.huffingtonpost.com
The Irish Times	Irishtimes	www.irishtimes.com
The Raw Story	Rawstory	www.rawstory.com
The Smirking Chimp	Smirkingchimp	www.smirkingchimp.com
The Straits Times	Straitstimes	www.straitstimes.com
The Telegraph	Telegraph	www.telegraphindia.com
The Toronto Star	Thestar	www.thestar.com
Toronto Sun	Torontosun	www.torontosun.com
The Wall Street Journal	WSJ	www.online.wsj.com
TODAYonline	Todayonline	www.todayonline.com
TV Guide	TVGuide	www.tvguide.com
US News & World Report	USNews	www.usnews.com
USA Today	USAToday	www.usatoday.com
Vancouver Sun	Vancouversun	www.vancouversun.com
VentureBeat	Venturebeat	www.venturebeat.com
Washington Post	Washingtonpost	www.washingtonpost.com
Washington Times	Washingtontimes	www.washingtontimes.com
WJLA.com & Virginia News	WJLA	www.wjla.com
WND	WND	www.wnd.com
WNYC	Wnyc	www.wnyc.org
Yahoo News	Yahoo	www.news.yahoo.com

Table 3.3: Number of Collected Tweets

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
1	Torontosun	167	145	11
	Thestar	105	103	0
	CBC	96	94	0
	Globaltoronto	28	22	13
	Yahoo	18	18	0
	Globalmail	1	1	0
2	USBTV	425	399	5
	CNN	279	264	6
	USAToday	183	178	1
	Yahoo	63	60	1
	MSN	15	14	0
	Newsday	2	2	0
3	CBS	130	127	2
	Yahoo	75	71	0
	WJLA	41	36	0
	Dallasnews	23	22	4
	Bakersfield	2	2	0
	CSMonitor	2	2	0
4	Venturebeat	312	303	2
	PCWorld	170	154	2
	Reuters	125	125	1
	Yahoo	72	69	0
	Economictimes	30	29	1
5	ABC	115	108	1
	NBClosangeles	86	82	4
	NBC	66	65	2
	Yahoo	65	60	
6	Yahoo	335	290	0
	Aljazeera	205	194	7
	Reuters	101	94	0
	NBC	9	9	0
	CBS	8	8	0
	ABC	6	6	0

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
7	BBC	682	634	12
	Aljazeera	245	230	6
	Reuters	114	107	1
	Guardian	35	35	0
	Timesofindia	28	29	0
	Dawn	16	20	0
	Prothomalo	12	11	1
	IBTimes	11	8	3
8	Yahoo	858	278	0
	Huffington	92	89	0
	Swampland	42	41	0
	Washignton post	40	40	0
	Thestar	14	14	0
	MSN	10	8	1
	Baltimore	4	4	2
	Recordonline	1	1	0
9	CNN	849	800	14
	ABC	174	171	5
	Reuters	83	83	1
	NBC	75	65	1
	CBS	33	31	2
	People	17	17	1
	Stuff	8	5	4
	Irishtimes	6	6	0
	TVGuide	5	5	1
	Examiner	4	1	0
	UK_Reuters	4		
10	Guardian	1487	1468	2
	BBC	1484	1425	6
	CNN	1459	1415	12
	Yahoo	1132	1016	2
	Fox	934	900	5
	Aljazeera	706	688	9
	CBC	342	340	5

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
	Reuters	196	169	6
	MSN	27	27	0
	ABC	17	17	0
	Dailycaller	13	13	1
	Lassvegassun	6	6	0
	Miamiherald	4	4	0
	Businessweek	2	2	0
	Jsonline	2	2	0
11	BBC	948	916	0
	Guardian	160	160	1
	Reuters	76	76	1
	CBS	74	70	3
	Yahoo	74	73	0
	Telegraph	28	28	0
	Globalpost	8	8	1
	Nationalgeographic	3	3	0
12	BBC	629	583	14
	Fox	335	319	3
	Aljazeera	202	195	4
	CBS	154	143	4
	USAToday	104	102	1
	AP	66	58	0
	ABC	32	31	1
13	Reuters	1466	788	9
	Guardian	1273	1172	16
	USAToday	911	853	6
	BBC	823	780	2
	CNN	569	535	10
	Aljazeera	398	385	7
	ABC	234	222	3
	Yahoo	191	183	0
	Straitstimes	168	165	0
	NJ	83	79	4

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
	CBS	39	39	3
	Timesofindia	38	35	0
	NBC	34	31	0
	Thetimes	18	17	0
	Channelnewsasia	16	16	1
	PBS	16	16	0
14	ABC	1499	81	1
	Yahoo	1472	81	0
	Aljazeera	734	714	6
	CNN	657	633	7
	BBC	417	403	2
	Reuters	275	263	1
	CBC	101	85	0
	Rawstory	96	91	1
	Washingtonpost	44	43	0
	CSMonitor	39	39	0
	Dawn	34	33	0
	SMH	33	32	1
	MSN	11	10	0
	NYTimes	1	1	0
15	Yahoo	187	180	1
	Fox	176	173	0
16	CNBC	513	486	0
	Thehill	488	464	2
	NBCPolitics	196	187	0
	Reuters	145	140	1
	Yahoo	143	139	0
	Guardian	79	78	0
	Politicalticker	77	74	1
	Fox	61	59	0
	Patdollard	34	29	1

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
	Sky	30	30	2
	Foxbusiness	12	11	0
	LATimes	6	6	0
	Pennlive	4	4	1
17	BBC	1538	1378	46
	Yahoo	472	419	16
	Fox	245	220	17
	ABC	119	113	1
	CBS	116	99	19
	Huffington	101	95	0
	NYPost	43	43	0
	Telegraph	32	32	0
	Newsmax	19	18	1
	Nydailynews	18	18	1
	CBC	10	10	0
	Dailyrecord	6	6	1
	Newsday	2	2	0
	Politico	1	1	0
18	CNN	1520	64	37
	Reuters	1492	225	17
	BBC	1066	958	30
	NYTimes	582	540	19
	WSJ	492	434	38
	Yahoo	331	297	0
	Guardian	69	68	0
	CBC	66	64	1
	NBC	15	12	

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
19	Yahoo	875	807	16
	ABC	663	607	38
	Foxnews	53	38	16
	Christianpost	15	15	0
20	BBC	1532	1444	40
	NYTimes	781	745	18
	Yahoo	437	389	0
	Guardian	373	368	1
	Reuters	245	224	16
	Telegraph	210	161	33
	CBS	192	187	2
	USAToday	115	113	1
	NBC	112	109	0
	CBC	42	42	0
	Nationalpost	38	37	1
	Washingtontimes	29	29	0
	Fox	18	18	0
	NPR	10	9	0
	ABC	9	9	0
	Timesofindia	1	1	0
21	Guardian	598	591	0
	BBC	590	556	12
	ABC	256	214	35
	Sky	176	160	17
	Timesofisrael	25	24	1
	Nydailynews	19	19	0
22	Yahoo	465	429	0
	Timesofisrael	25	24	1
	Fox	13	13	0

Clone Set	Member	Total Tweets	Unique Tweeter Tweets	Promotional Tweets
23	Fox	1722	1571	28
	Yahoo	1076	1005	15
	Reuters	892	747	38
	BBC	744	671	30
	Aljazeera	661	478	29
	Guardian	463	414	11
	USAToday	451	416	22
	Huffington	312	301	1
	CBS	156	85	3
	ABC	149	141	3
	CBC	100	90	30
	CNN	42	42	0
	News	30	21	6
	Vancouversun	29	21	4
24	Reuters	931	860	52
	Yahoo	682	618	33
	BBC	448	424	9
	Huffington	186	184	0
	Chicagotribune	80	63	18
	NYTimes	47	39	0
	CBSnews	40	39	0
	LATimes	35	31	1
	Washingtonpost	32	31	0
	Thestar	15	15	0
	Todayonline	13	13	0
	CNN	12	9	0
	ABCLocal	3	3	0
25	BBC	1395	1323	0
	CNN	1528	1414	29
	Reuters	625	558	34
	Yahoo	435	418	1

Chapter 4

Data Characterization

In this chapter, data for all of the clone sets is used for studying the distribution of the total number of tweets for a clone set member, total unique tweeter tweets, total self-promotional tweets, total redundant tweets, tweet timing, and the maximum number of tweeter followers. News site popularity and relative publication timing are also characterized. Four top tweeted clone sets 13, 14, 18 and 23 are chosen for detailed individual analysis.

Section 4.1 describes the clone sets chosen for individual analysis. Section 4.2 examines the distribution of the number of tweets, both in total and of various types. Section 4.3 concerns news site popularity. Section 4.4 characterizes the relative publication times of the clone set members. Section 4.5 looks at tweet timing, while Section 4.6 examines the distribution of the maximum number of tweeter followers.

4.1 Clone Sets Analyzed Individually

Table 4.1: Basic Characteristics of Clone Sets Chosen for Individual Analysis

Clone Set	Number of Members	Total Tweets	Maximum Tweets	Minimum Tweets	Mean Tweets	Median Tweets
13	16	6277	1466	16	392	180
14	14	5413	1499	1	368	99
18	9	5633	1520	15	626	492
23	14	6827	1722	29	488	382

Table 4.1 gives some basic characteristics of the four clone sets chosen for individual analysis. These four clone sets were chosen since they had a relatively large number of clone members as well as total tweets. In addition to the total (across all clone set members) number of tweets, the table also reports the maximum, minimum, mean and median number of tweets received by the individual clone set members. Note that for all of these clone sets, the variability in the number of tweets received for the various clone set members is quite high. The total tweet count and the maximum number of tweets across the clone set members are similar for the four clone sets, with the total tweets in the range of 5200-6800, and a maximum tweets count

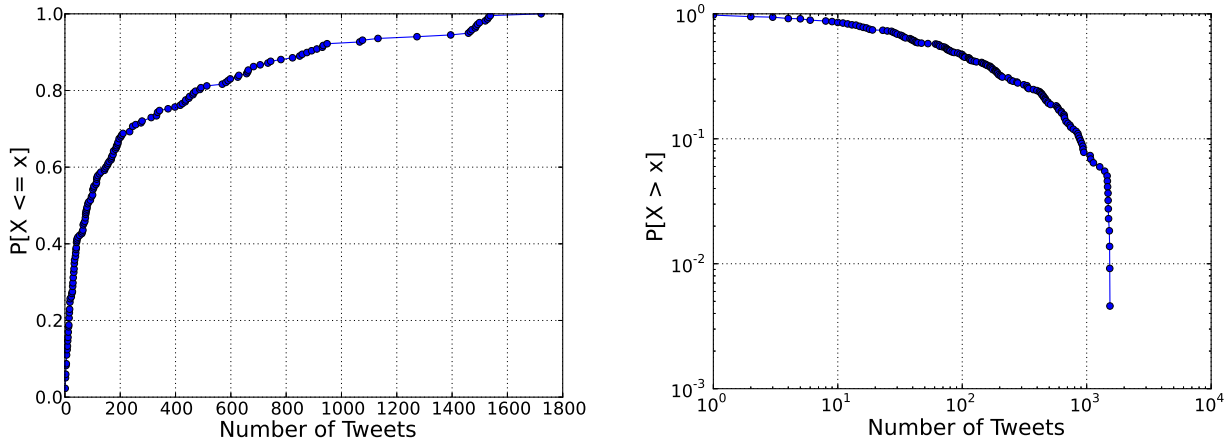
in the range of 1450 - 1750. The clone sets differ somewhat more with respect to the mean, median, and minimum tweet counts.

4.2 Number of Tweets

This section analyzes the distribution of the total number of tweets received by a clone set member. Then, the number of tweets of various types is considered, specifically the number of tweets from distinct Twitter users (unique tweeter tweets), the number of self-promotional tweets, and the number of redundant tweets. Both the cumulative and complementary cumulative distribution functions are considered, as well as the proportion of tweets in the different categories for the four clone sets chosen for individual analysis. In total there were 60973 tweets received by the 218 clone set members of the 25 clone sets.

Section 4.2.1 describes the distribution of the total number of tweets, Section 4.2.2 considers the distribution of the number of unique tweeter tweets, Section 4.2.3 considers the distribution of the number of self-promotional tweets, Section 4.2.4 examines the distribution of the number of redundant tweets and finally Section 4.2.5 considers the number of tweets of the different categories for the four specific clone sets.

4.2.1 Distribution of Total Number of Tweets



(a) CDF

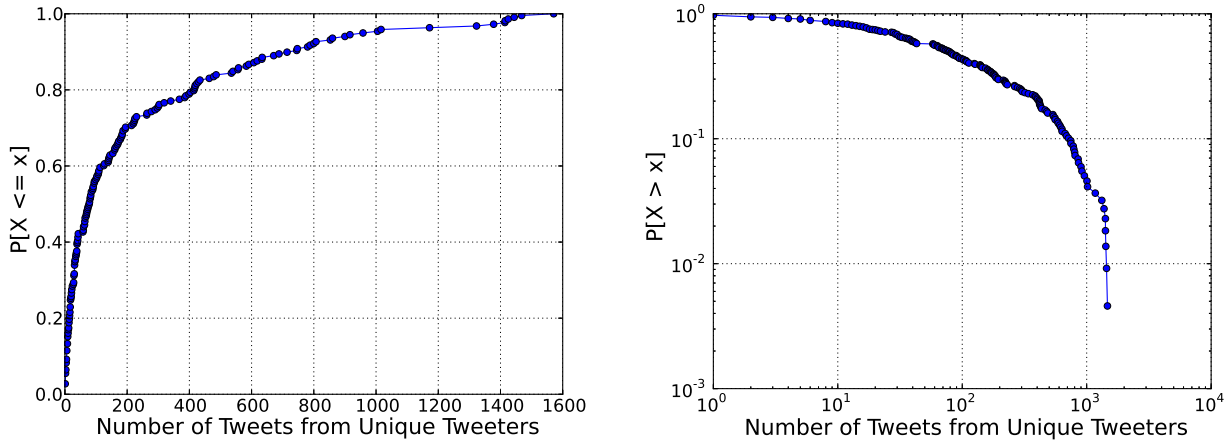
(b) CCDF

Figure 4.1: Distribution of Number of Tweets (All Clone Sets)

In order to obtain a more complete view of the distribution, both the cumulative distribution function (CDF) and the complementary cumulative distribution function (CCDF) are shown, with the former shown with linear x-axis and y-axis scales and the latter shown with logarithmic x-axis and y-axis scales, as is typical in the literature.

Figure 4.1(a) shows the CDF of the total number of tweets for each clone set member, across all 25 clone sets. Note that about 80% of the news site stories received less than 500 tweets. About 10% of the news site stories received between 500 and 900 tweets, and about 10% received between 900 and 1800 tweets. From the CCDF graph shown in Figure 4.1(b), the distribution is clearly light-tailed.

4.2.2 Distribution of Number of Tweets from Unique Tweeters



(a) CDF

(b) CCDF

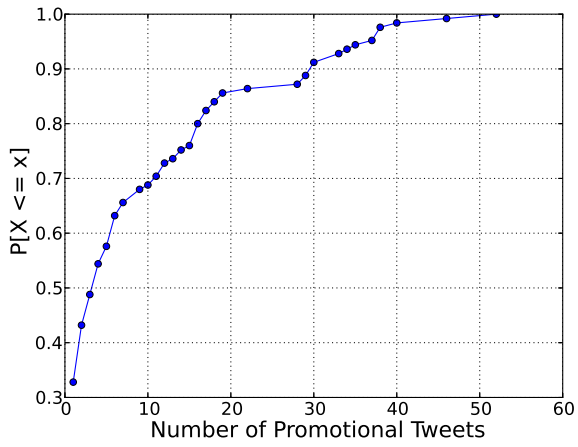
Figure 4.2: Distribution of Number of Tweets from Unique Tweeters (All Clone Sets)

The number of unique tweeter tweets could be a better measure of the popularity of online content. Sometimes the total number of tweets is substantially inflated by many tweets coming from the same tweeters. These "redundant" tweets could be promotional tweets made from a Twitter account affiliated with the news site, or could correspond to conversational exchanges among the tweeters. The number of unique tweeter tweets represents the number of different users that tweeted regarding the news story version.

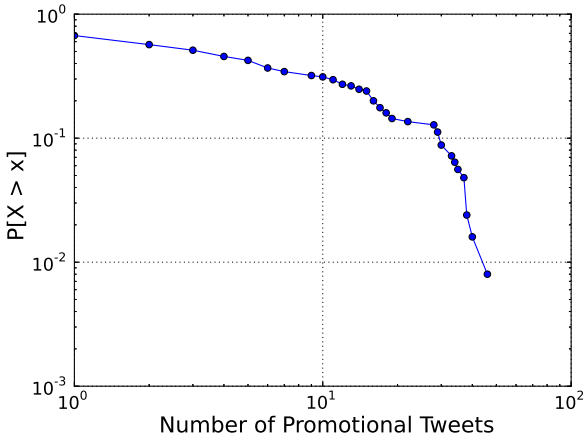
Figure 4.2(a) shows the CDF of the unique tweeter tweets received by a clone set member. It is found that about 80% of the clone set members got less than 400 unique tweeter tweets, about 10% of them got between 400 and 800, and about 10% got more than 800 unique tweeter tweets. As with the total number of tweets, the CCDF graph in Figure 4.2(b) shows that the distribution is clearly light-tailed.

4.2.3 Distribution of Number of Self-Promotional Tweets

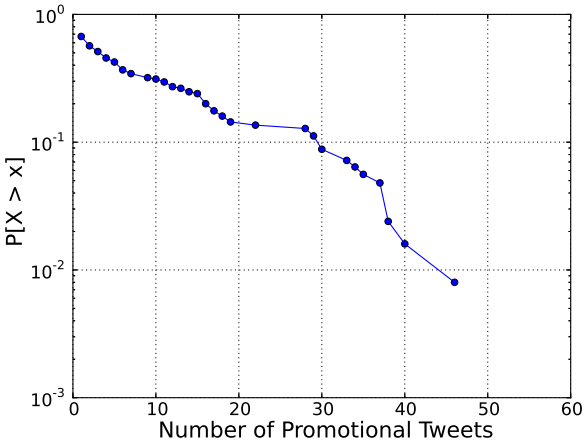
Sometimes it is observed that significant numbers of tweets come from a Twitter account that appeared to be affiliated with the news site itself, following some kind of pattern such as posting one or two tweets every hour or at some other regular interval. In order to identify such a self-promotional tweeter, each of the tweeter names was compared to the news site name to check for a match in either suffix, prefix, or any portion,



(a) CDF



(b) CCDF



(c) CCDF with x-axis on linear scale

Figure 4.3: Distribution of Number of Self-Promotional Tweets (All Clone Sets)

either in full name or as an abbreviation. When a match was found those tweeter's tweets were treated as self-promotional tweets for that particular clone set member.

This kind of activity from a news site could be quite effective in promoting their news stories. Many popular news sites maintain active social networking accounts having thousands of followers. News sites tweet their important news stories in Twitter via their official Twitter account, so their followers can reach their posted news stories in their Twitter news feed, without surfing those news sites or reading a newspaper. It is also observed that news sites tweet the same news story redundantly to reach the attention of their users. In Figure 4.3, the distribution of the self-promotional tweets received by a clone set member is shown. From the CDF shown in Figure 4.3(a), it is found that about 90% of the clone set members received fewer than 30 self-promotional tweets. The highest number of promotional tweets received by any clone set member was 52 (for a Reuter's news story, in clone set 24). Note that not all the clone set members received promotional tweets; only 125 out of the 218 clone set members had at least one promotional tweet. The CCDF is shown in both Figure 4.3(b) (with a log scale on both axes), and in Figure 4.3(c) (with a linear scale on the x-axis). When graphed with a linear scale on the x-axis, the plot has a roughly linear form. The exponential distribution gives a linear plot when its CCDF is graphed with a log scale on the y-axis and a linear scale on the x-axis, suggesting that the distribution of self-promotional tweets may resemble an exponential distribution.

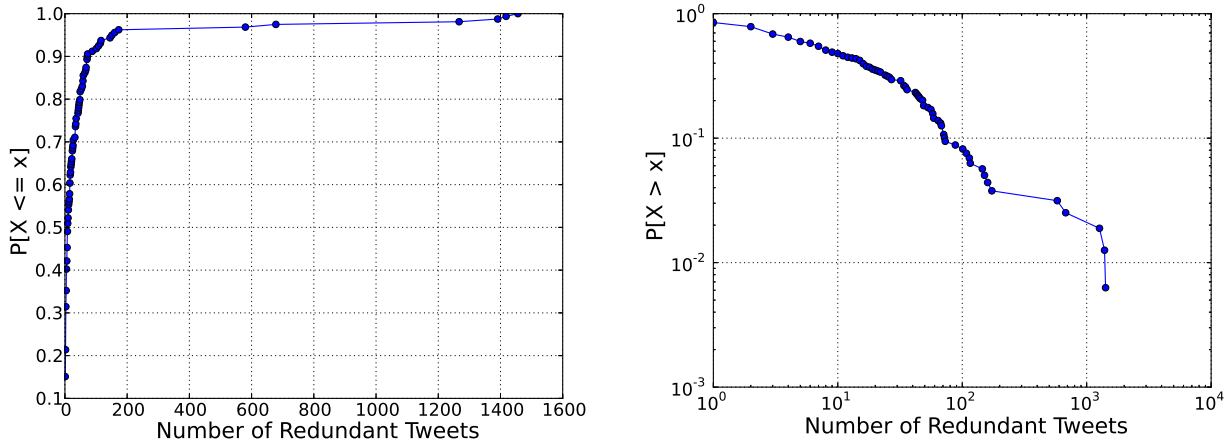
4.2.4 Distribution of Number of Redundant Tweets

Redundant tweets are defined here as tweets made by tweeters that have already tweeted at least once regarding the news site story. Sometimes many redundant tweets come from tweeters not affiliated with the news site. It is also observed that only some of the clone set members received redundant tweets. In particular, 159 out of the 218 clone set members received at least one redundant tweet. Figure 4.4 shows the distribution of the number of redundant tweets received by a clone set member. From the CDF shown in Figure 4.4(a), it is found that around 96% of the clone set members received fewer than 200 redundant tweets. Only 2 of them received between 200 and 1000 redundant tweets, while four of them received between 1000 and 1500 redundant tweets. The CCDF shown in Figure 4.4(b) suggests a more heavy-tailed form than seen earlier in Figures 4.1(b), 4.2(b), and 4.3(b).

4.2.5 Number of Tweets in Different Categories for Example Clone Sets

In this section, a detailed look is taken at the proportion of tweets in the different categories for the four clone sets chosen for individual analysis. This will show how the proportion of tweets in the different categories can widely vary across clone set members.

Figure 4.5 shows the number of received tweets in different categories for each member of clone set 13. For this figure as well as the subsequent figures in the section, tweets are categorized in a mutually-exclusive manner. One category consists of the self-promotional tweets ("self"), which are identified as described in



(a) CDF

(b) CCDF

Figure 4.4: Distribution of Number of Redundant Tweets (All Clone Sets)

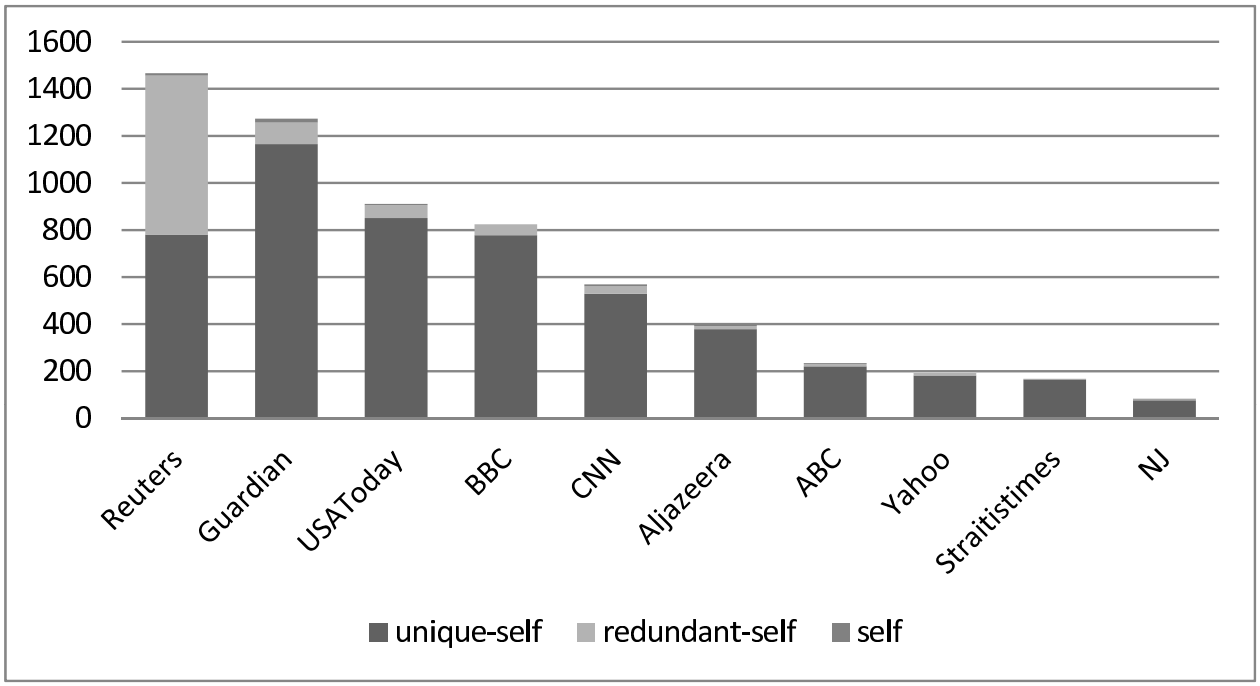


Figure 4.5: Number of Tweets in Different Categories for Members of Clone Set 13

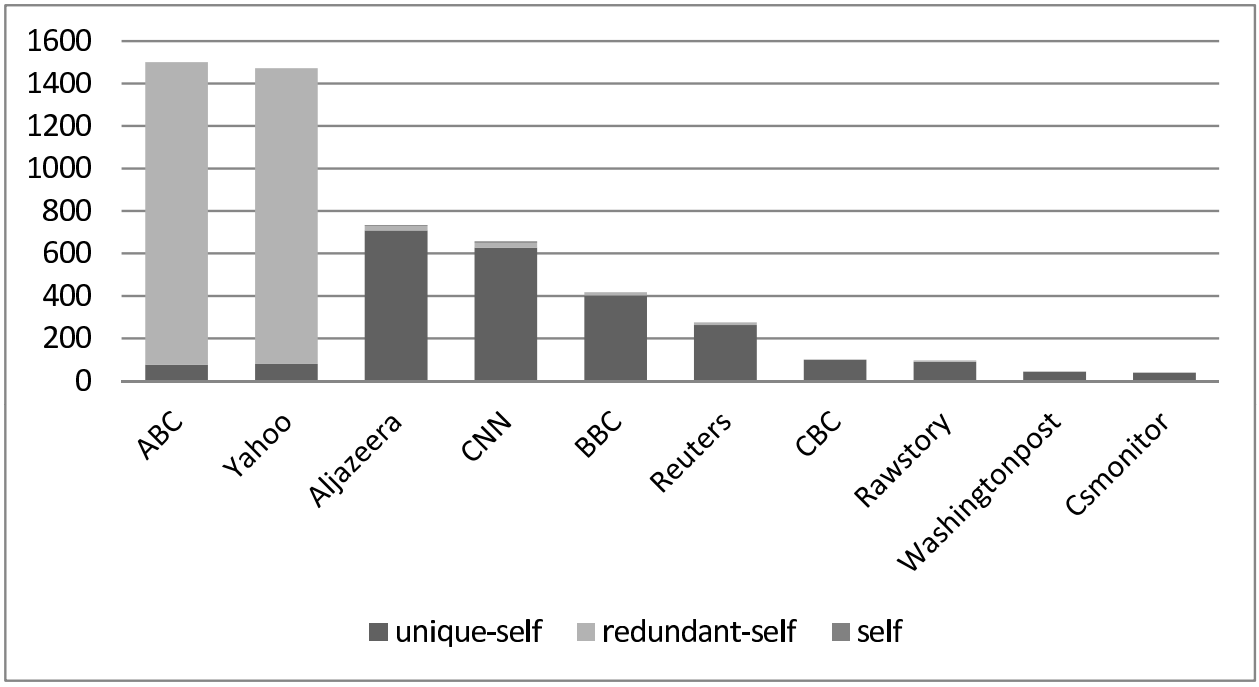


Figure 4.6: Number of Tweets in Different Categories for Members of Clone Set 14

Section 4.2.3. A second category consists of redundant tweets excluding self-promotional tweets ("redundant-self"). The final category consists of the unique tweeter tweets, i.e. the number of tweets from different Twitter users, excluding self-promotional tweeters ("unique-self"). As seen in Figure 4.5, for clone set 13 the highest number of total tweets was received by the version of the news story on the Reuters site, but a substantial number of these tweets were redundant, non-self-promotional tweets. In particular, more than 40% of the total tweets were of that type. A small number of self-promotional tweets came from Reuter's Twitter account. For all the other clone set members, the considerable majority of tweets were unique tweeter (excluding self-promotional tweeters) tweets.

Figure 4.6 shows the results for clone set 14. In this case, almost all of the tweets for the Reuter's clone set member fall into the unique-self category. However, for the two clone set members with the most tweets, from ABC and Yahoo, most tweets were redundant, non-self-promotional tweets. If only unique tweeter tweets were considered, these clone set members would be ranked seventh and eighth with respect to popularity within their clone set. All other clone set members received mostly unique tweeter tweets. This figure illustrates that redundant tweets for a clone set member can give a high total tweet count and relative popularity rank within the clone set, even without considering self-promotional tweets.

Figure 4.7 shows clone set 18's different categories of tweets. For the top tweeted site (CNN), most tweets were unique tweeter tweets, whereas for the news story version with the second largest number of total tweets (Reuters) most tweets were redundant, non-self-promotional tweets. The high popularity rank of the Reuters clone set member, when measuring popularity by the total number of tweets, again illustrates the potential impact of redundant tweets, even when excluding self-promotional tweets. For all other clone set members,

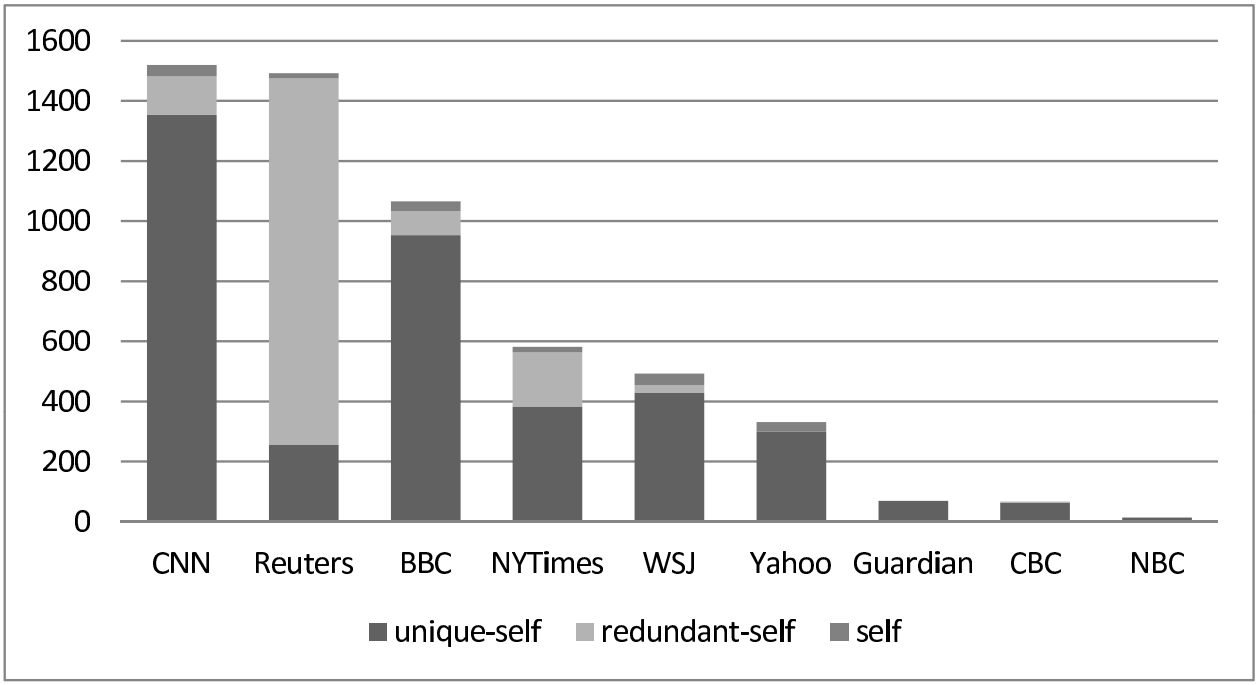


Figure 4.7: Number of Tweets in Different Categories for Members of Clone Set 18

most tweets were unique tweeter tweets. The WSJ clone set member illustrates a case where the number of self-promotional tweets exceeds the number of redundant, non-self-promotional tweets.

Figure 4.8 shows the results for clone set 24's different categories of tweets. For this clone set, most tweets for all clone set members were unique tweeter tweets. Some clone set members also received significant, but relatively small, numbers of redundant, non-self-promotional tweets, and most received a few self-promotional tweets.

4.3 News Site Popularity

In order to measure the overall popularity of the news sites, according to the number of collected tweets, both the average number of total tweets and the average number of unique tweeter tweets were considered. News sites that had a clone set member in fewer than 3 of the 25 clone sets were not considered. Figure 4.9 shows a bar chart of the average number of total tweets received by clone set members on different news sites, considering only the top 13 news sites with respect to this metric. From the bar chart, it is found that other than the top two and bottom three news sites, the average number of total tweets for a clone set member on each of the sites is between 250 and 475. According to this average number of total tweets metric, the BBC site is the most popular, whereas CNN is the next most popular.

Figure 4.10 shows a bar chart of the average number of unique tweeter tweets received by clone set members on different news sites, considering only the top 13 news sites with respect to this metric. Comparing Figure 4.10 and Figure 4.9, it can be seen that twelve news sites are in the top 13 with respect to both metrics, but the

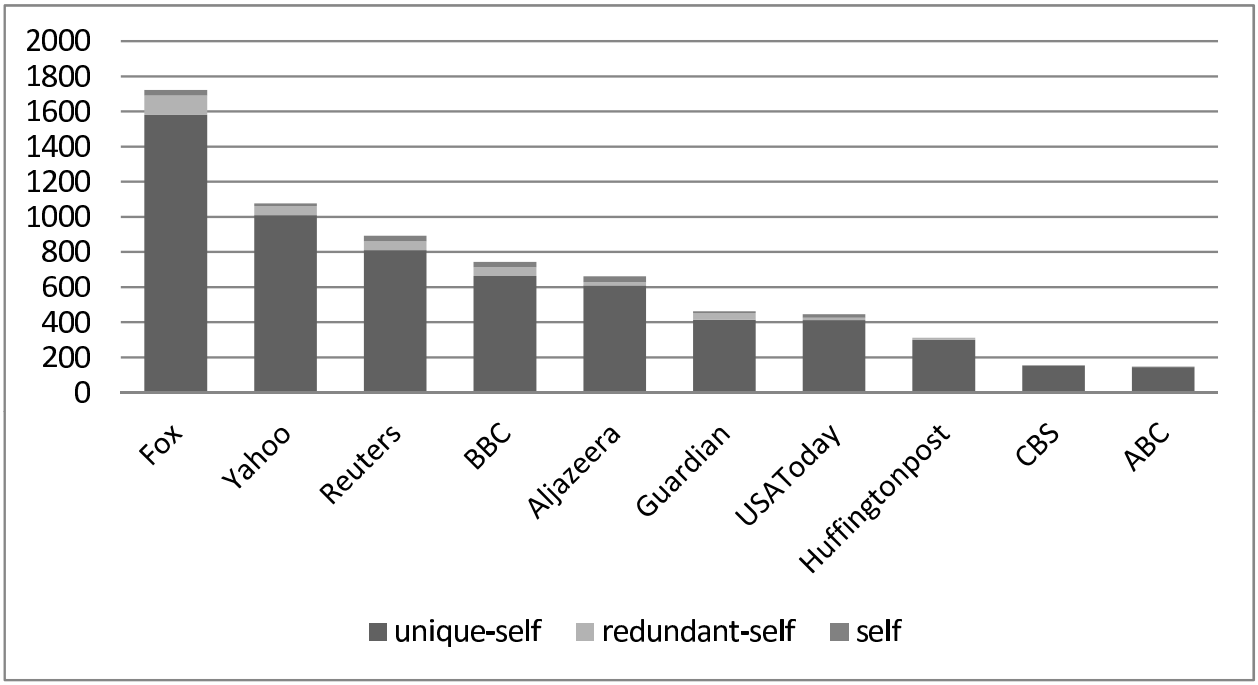


Figure 4.8: Number of Tweets in Different Categories for Members of Clone Set 23

relative rankings of these sites differ between the two figures in many cases. The CBC news site is not present in the top 13 when the average number of unique tweeter tweets is considered, whereas Washingtonpost, which is middle-ranked in Figure 4.10, is not present in the top 13 when the average number of total tweets is considered. In both cases the top two sites are the same, i.e. BBC and CNN, but for third place Guardian replaces Yahoo when considering the average number of unique tweeter tweets. The sites with ranks 6 to 10 have a very similar average number of unique tweeter tweets, while the top-ranked site and bottom three sites differ substantially from the others with respect to this metric.

4.4 Relative Publication Time

News sites publish their stories at varying times, reflecting in part time zone differences. Some of the news sites do not give the story publication time, so it is difficult to consistently get publication times for publication time analysis. The time of each tweet, however, is obtained from the Twitter Search API and reported in universal (UTC) time. In this thesis, the time of the first tweet referencing a clone set member is used as an estimate of that member's publication time.

This section concerns the relative publication times of clone set members. For each clone set, the publication time of each clone set member is estimated using the time of that clone set member's earliest tweet. The clone set member with the earliest estimated publication time is assigned a relative publication time of zero. Each other member of the clone set is assigned a relative publication time given by the difference between its estimated publication time and that of the earliest clone set member. Figure 4.11(a) shows the CDF of the

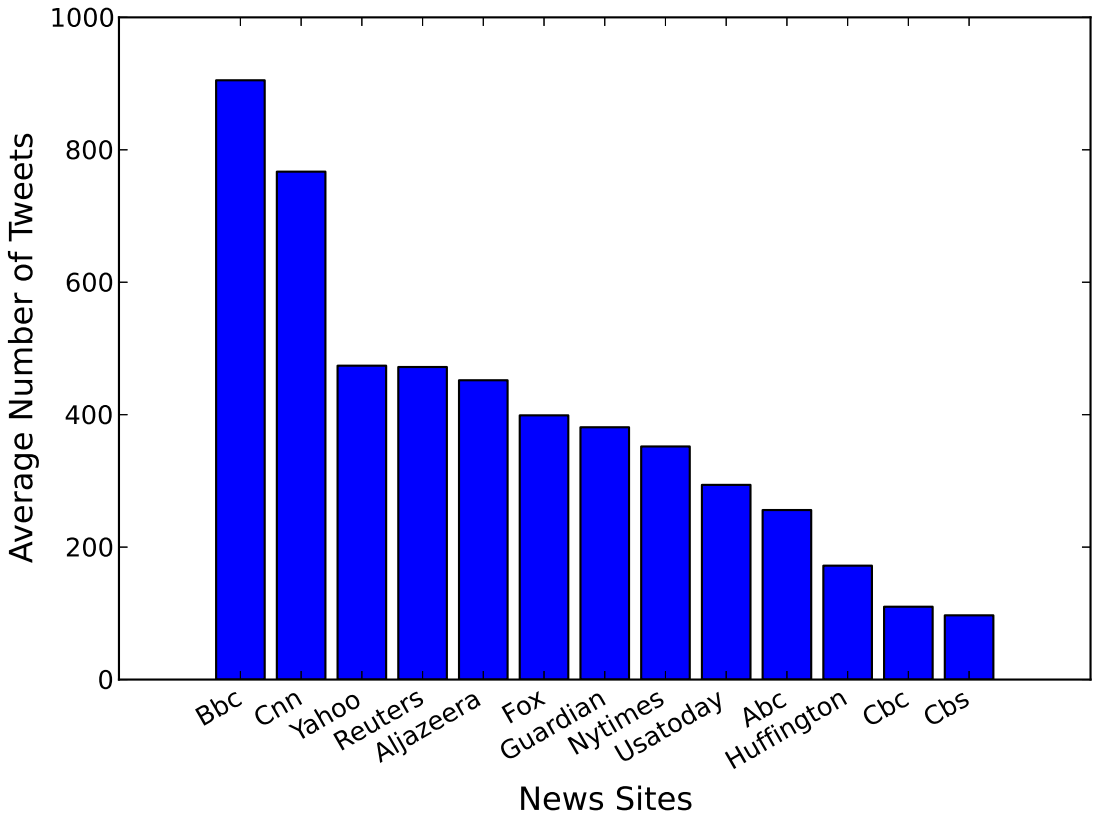


Figure 4.9: Average Number of Tweets for Clone Set Members on Different Sites (All Clone Sets, All News Sites in at least 3 Clone Sets)

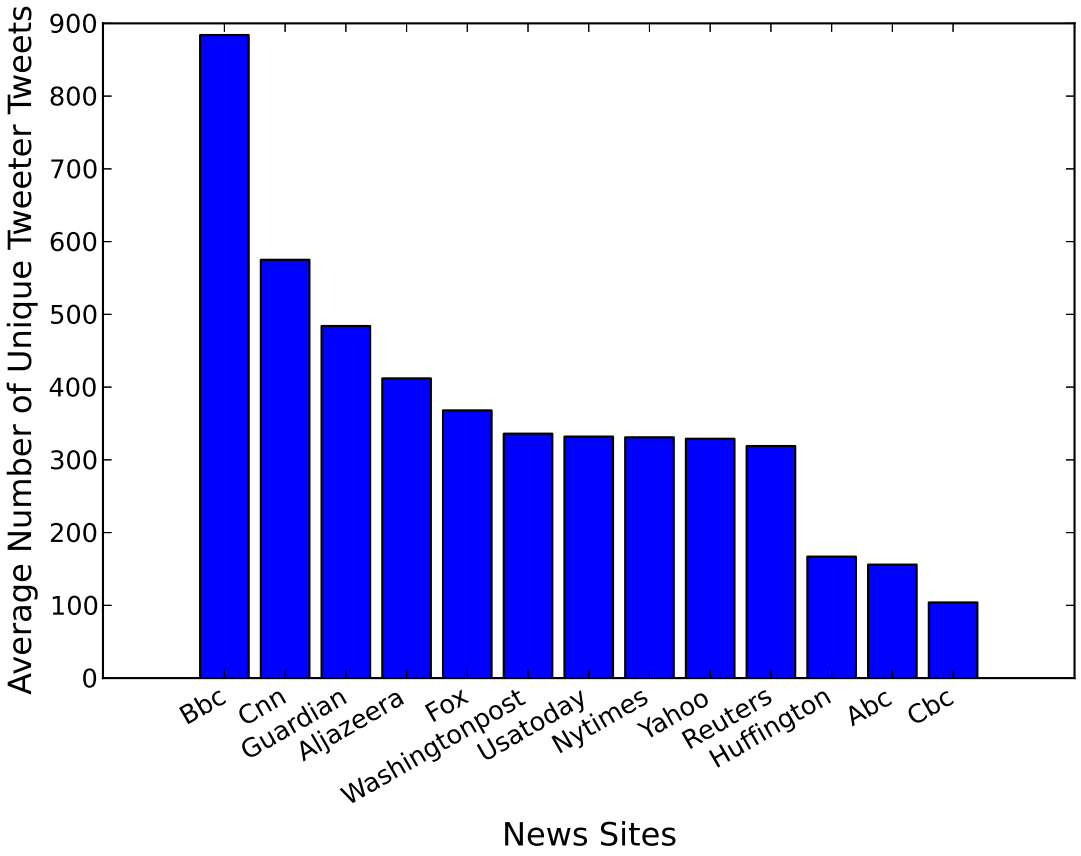
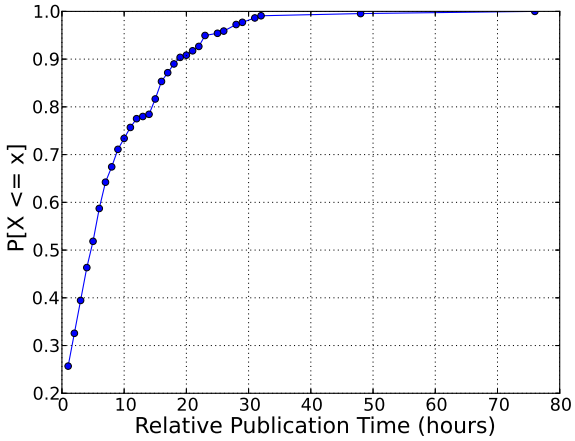
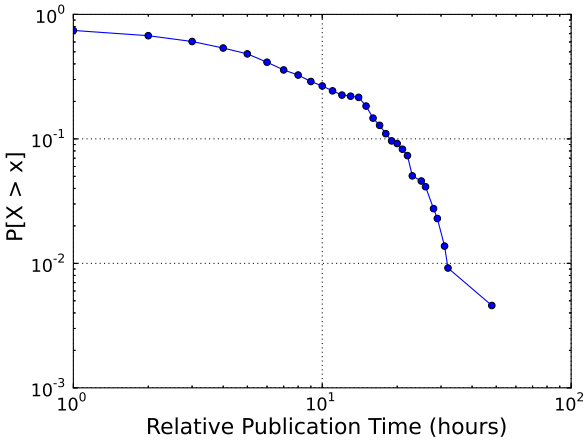


Figure 4.10: Average Number of Unique Tweeter Tweets for Clone Set Members on Different Sites (All Clone Sets, All News Sites in at least 3 Clone Sets)

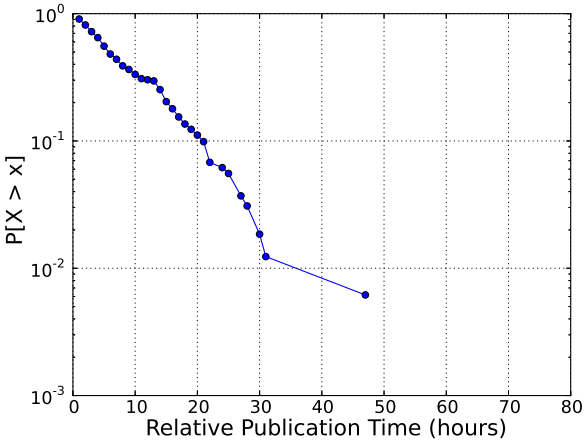
resulting distribution of relative clone set member publication times. It is found that almost 98% of clone set members have a publication time within 30 hours of that of the earliest member of their clone set. However, one clone set member has a publication time more than 75 hours later than that of the earliest member of its clone set. Figures 4.11(b) and 4.11(c) show the CCDF using a log scale on the y-axis and either a log scale on the x-axis (4.11(b)) or a linear scale (4.11(c)). The roughly linear form of the plot in Figure 4.11(c) suggests that the relative publication time distribution may resemble an exponential distribution over much of its range.



(a) CDF



(b) CCDF



(c) CCDF with x-axis on linear scale

Figure 4.11: Distribution of Relative Publication Time for the Clone Set Members within a Clone Set (All Clone Sets)

Table 4.2 gives the relative publication times of the members of the clone sets chosen for individual analysis. Note that for each clone set, the earliest published clone set member has a relative publication time of 0:00, and the times shown for later published clone set members are calculated relative to the publication

Table 4.2: Relative Publication Time of the Example Clone Set Members

Clone Set 13	hh:mm	Clone Set 14	hh:mm	Clone Set 18	hh:mm	Clone Set 23	hh:mm
USAToday	0:00	BBC	0:00	Yahoo	0:00	Reuters	0:00
Guardian	3:56	Dawn	0:07	BBC	3:58	USAToday	0:07
Reuters	6:41	SMH	0:57	NBCNews	3:59	Yahoo	0:08
BBC	14:05	Reuters	1:19	WSJ	4:30	Aljazeera	1:10
Straitstimes	14:37	CNN	1:49	Guardian	5:15	Vancouversun	2:03
NJ	15:08	Rawstory	2:09	NYTimes	5:38	CNN	2:17
Aljazeera	15:20	Aljazeera	2:19	CBC	6:28	Huffington	2:33
THETimes	15:26	CBC	2:33	CNN	7:09	Guardian	3:33
Channelnewsasia	15:30	MSN	5:02	Reuters	9:39	Fox	3:42
PBS	15:35	CSMonitor	5:12			ABC	8:06
Timesofindia	15:58	NYTimes	5:56			CBC	9:33
CBSNews	16:19	Washigntonpost	6:59			BBC	14:27
CNN	16:23	Yahoo	30:05			News	18:40
ABC	17:07	ABC	30:58			CBS	19:19
NBCNews	17:57						
Yahoo	90:53						

time of the earliest. It is interesting to note that in some cases, bunching of the publication time is apparent.

4.5 Tweet Timing

This section examines the timing of tweets. For most of the figures in this section times of tweets are calculated relative to the time of the earliest tweet for the respective clone set member. Specifically, the earliest tweet received by a clone set member is considered to occur at time 0. The times of all other tweets for that clone set member are calculated relative to the time of that first tweet. Figure 4.12(a) shows the CDF of the relative tweet times, aggregating the relative tweet times for all the members of all of the clone sets. Note that more than 95% of all tweets in the entire data set were tweeted within 35 hours of the earliest tweet of their respective clone set member. Figures 4.12(b) and 4.12(c) show the CCDF using a log scale on the y-axis and either a log scale on the x-axis (4.12(b)) and a linear scale (4.12(c)) respectively. The roughly linear form of the plot in Figure 4.12(c) suggests that the relative tweet time distribution may resemble an exponential distribution over much of its range.

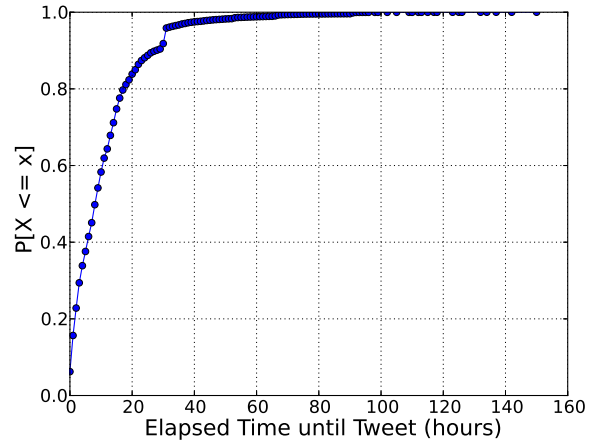
Figure 4.13 shows the CDF of the relative tweet times for the five most tweeted clone set members for each of the four clone sets chosen for individual analysis. Figure 4.13(a) shows the results for clone set 13. Note that for the five most tweeted clone set members, almost all tweets occurred within the first day. For two of the clone set members, about 80% of their tweets occurred within the first five hours. For the other three clone set members, tweets were more spread out in time, and there appear to be spikes in tweeting rate that may reflect time-of-day effects or cases where one tweet had sparked many other tweets.

Figure 4.13(b) shows the CDF of the relative tweet times for clone set 14. For two of the clone set members, almost all tweets occurred within the first few hours. The rest of the clone set members show variable behaviour, with one of them acquiring all its tweets by the first 24 hours, while the other two took around 55 hours to acquire most of their tweets.

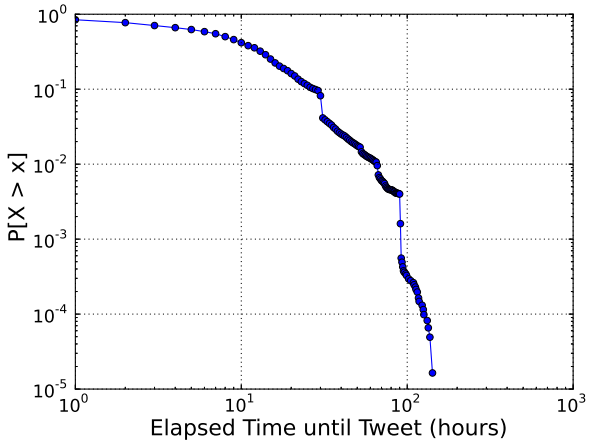
As seen in Figure 4.13(c), two of the members of clone set 18 received more than 85% of their total tweets within the first five hours, and almost all of their tweets within the first day. Another two of the clone set members took around 20 hours for acquiring 85% of its total tweets and had acquired almost all of their tweets within 60 hours. One of the clone set members took almost 50 hours for acquiring 85% of its total tweets. Again, there appear to be prominent spikes in tweeting rate that may reflect time of day effects or cases where one tweet has triggered many others.

Finally, as seen in Figure 4.13(d), the members of clone set 23 are more similar in their tweet timing than observed for the other clone sets. All five members received at least 80% of their total tweets within the first 10 hours.

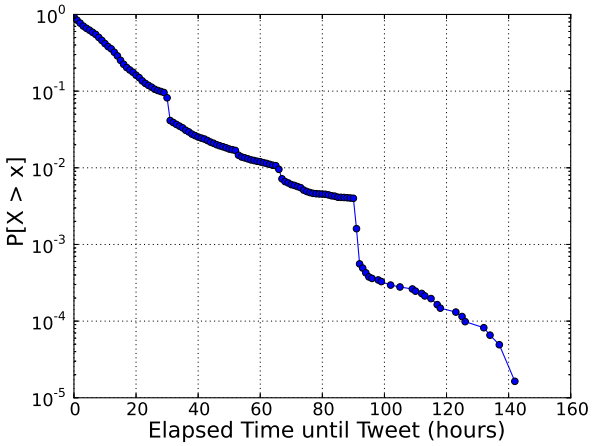
Figures 4.14 and 4.15 show the CCDF of the relative tweet times for the same five members of each of four clone sets as considered for Figure 4.13. Figure 4.14 uses a log scale on the x-axis of each plot, while Figure 4.15 uses a linear scale. Although the roughly linear form of the plot in Figure 4.12(c) suggests that



(a) CDF

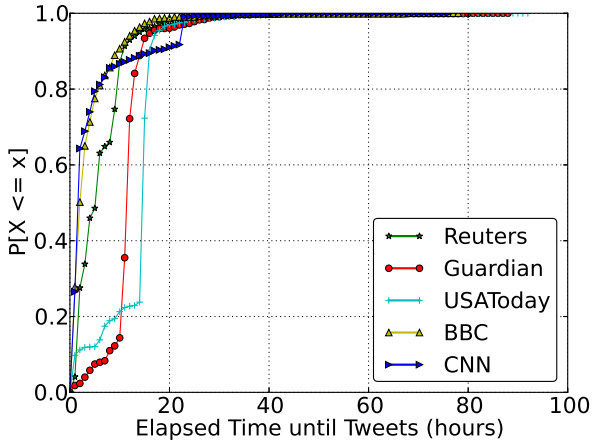


(b) CCDF

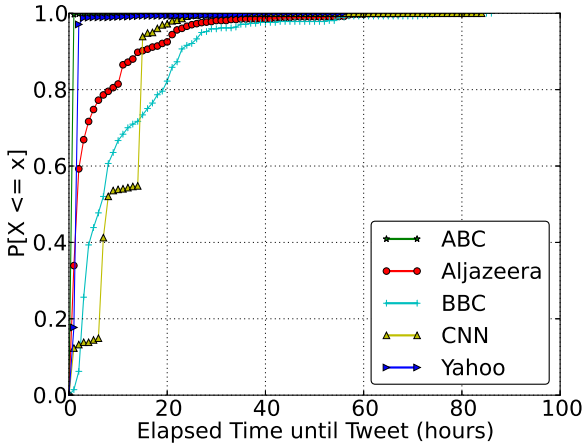


(c) CCDF with x-axis on linear scale

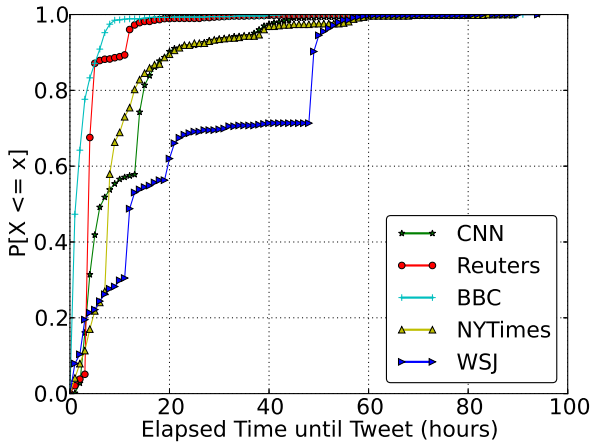
Figure 4.12: Distribution of Elapsed Time from first Tweet for a Clone Set Member, to Times of Subsequent Tweets for that Clone Set Member (All Clone Sets)



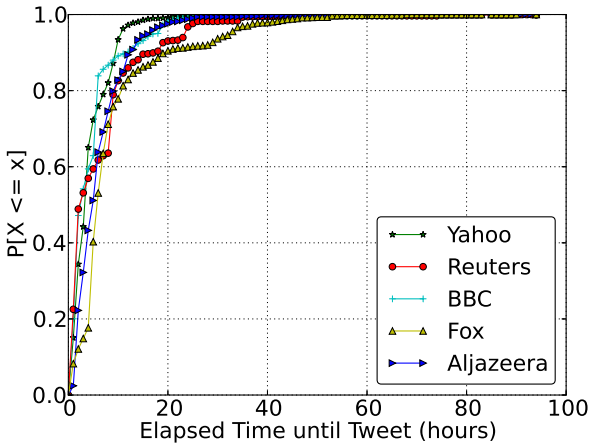
(a) Clone Set 13



(b) Clone Set 14

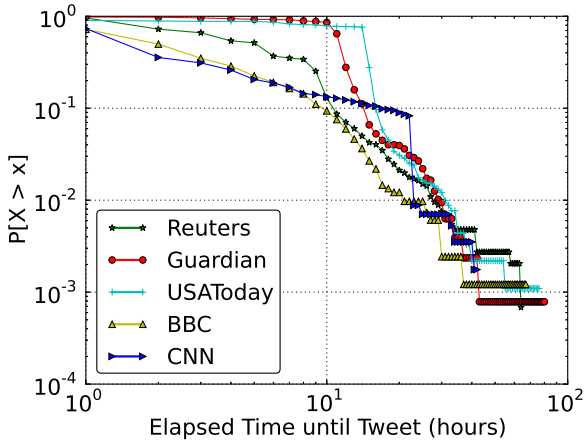


(c) Clone Set 18

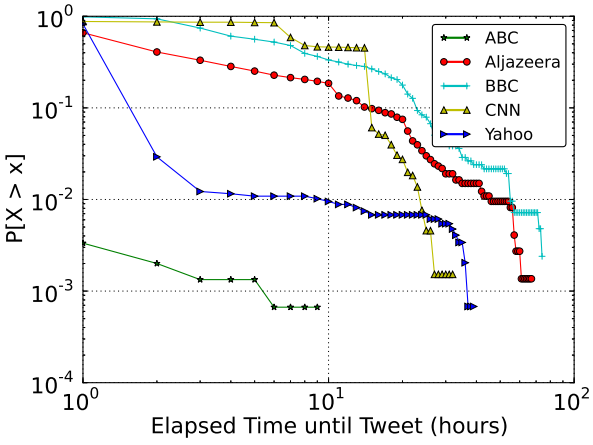


(d) Clone Set 23

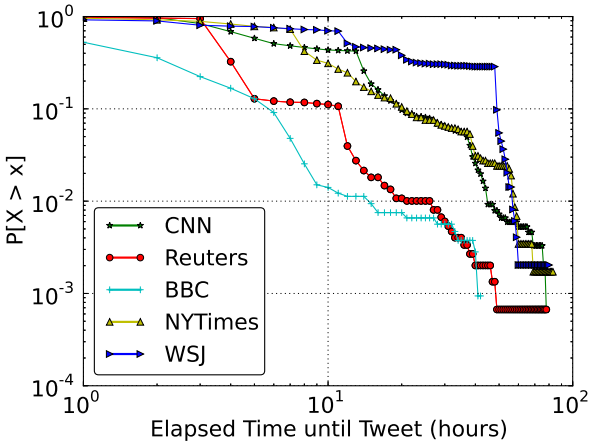
Figure 4.13: Cumulative Distribution Function of Elapsed Time from first Tweet for a Clone Set Member, to Times of Subsequent Tweets for that Clone Set Member (Example Clone Sets, Top Five Sites for each Clone Set)



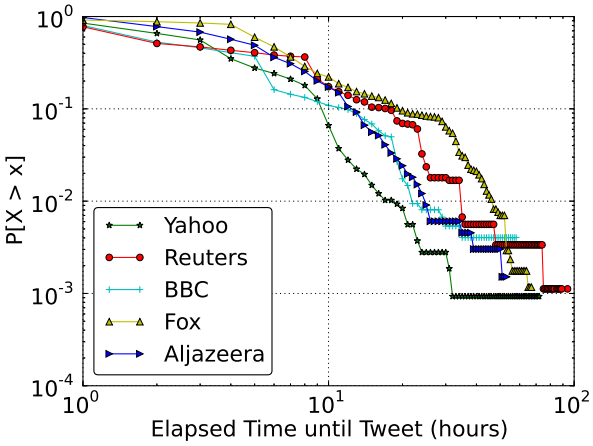
(a) Clone Set 13



(b) Clone Set 14

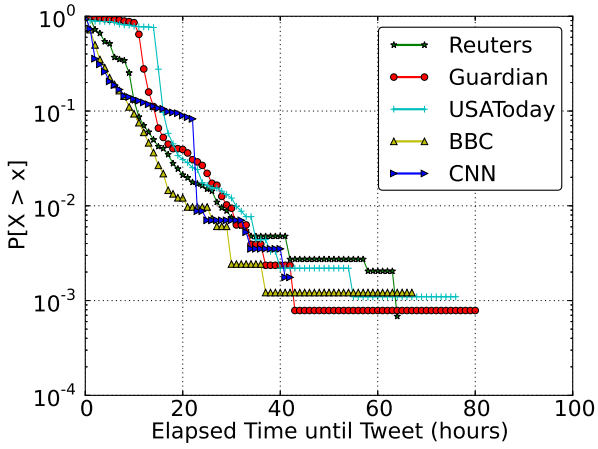


(c) Clone Set 18

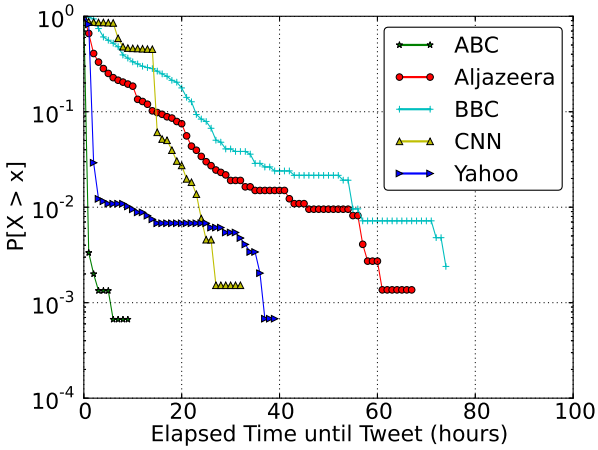


(d) Cline Set 23

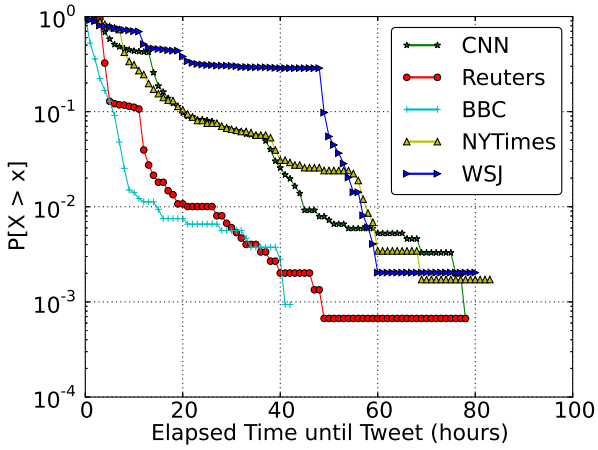
Figure 4.14: Complementary Cumulative Distribution Function of Elapsed Time from first Tweet for a Clone Set Member, to Times of Subsequent Tweets for that Clone Set Member (Example Clone Sets, Top Five Sites for each Clone Set)



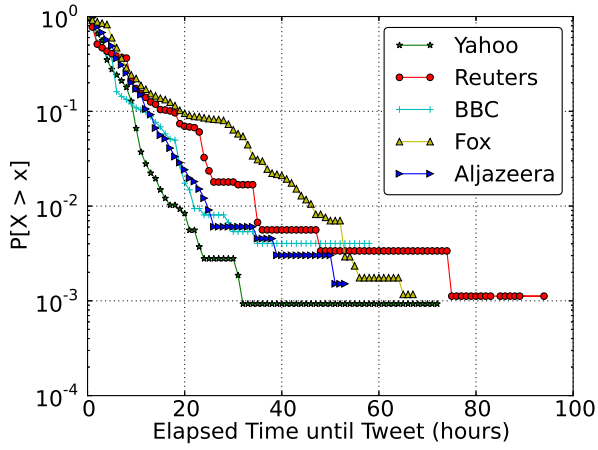
(a) Clone Set 13



(b) Clone Set 14



(c) Clone Set 18



(d) Cline Set 23

Figure 4.15: Complementary Cumulative Distribution Function of Elapsed Time from first Tweet for a Clone Set Member, to Times of Subsequent Tweets for that Clone Set Member (Example Clone Sets, Top Five Sites for each Clone Set, x-axis on Linear Scale)

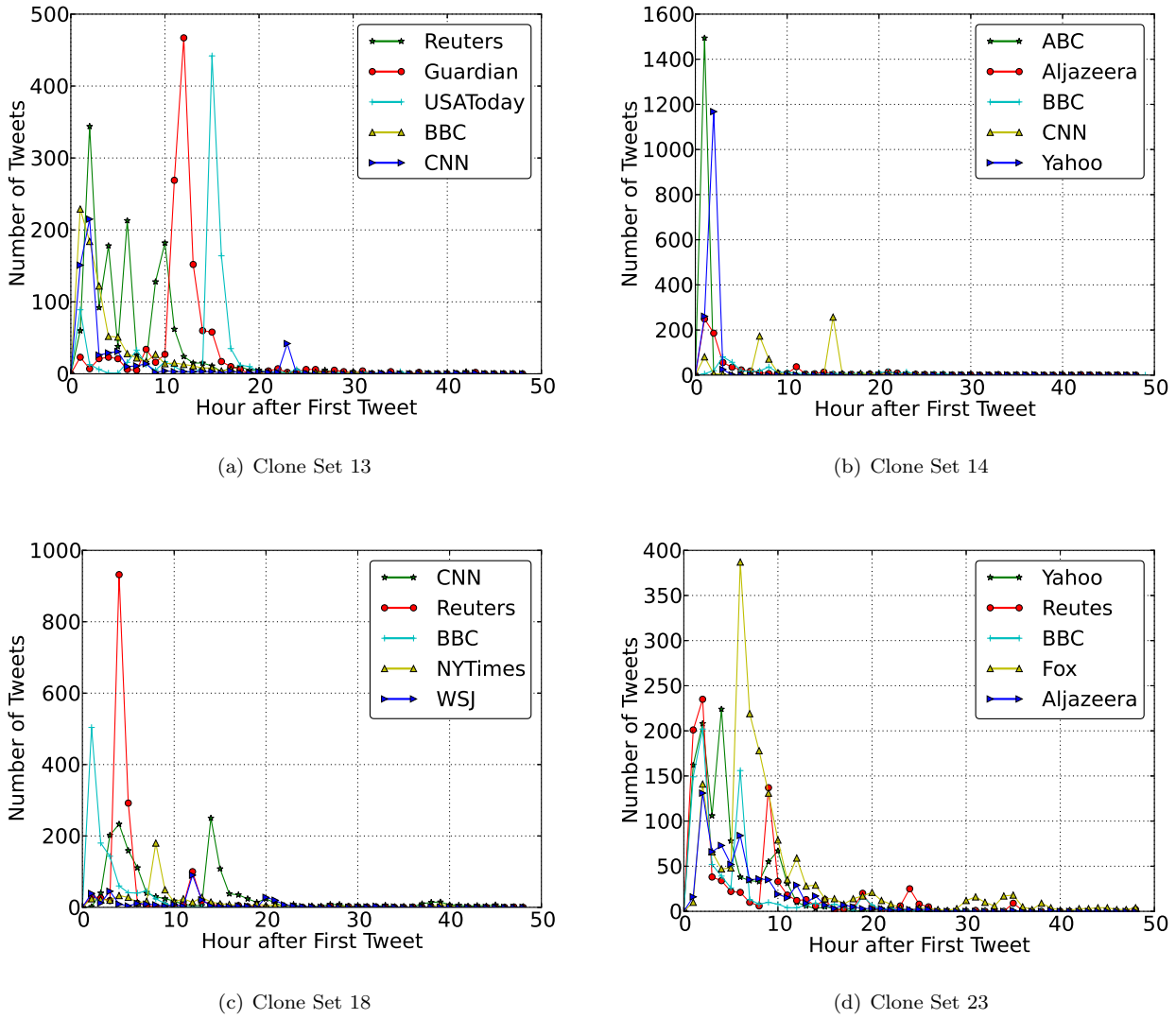
the relative tweet timing may resemble an exponential distribution, when data is aggregated from all the clone sets, such behaviour is not as clearly seen when looking at plots for the individual clone set members. Although some roughly linear forms are seen in the plots of Figure 4.15, each plot has relatively few data points (compared to the plot in Figure 4.12(c)), and the plots are highly variable.

Figure 4.16 plots the number of tweets received in each of the first 48 hours (measured from the time of the first tweet to the clone set member), for the same five members of each of four clone sets as considered for Figures 4.13-4.15. This figure clearly shows the hours with peak tweeting activity. For clone set 13 in Figure 4.16(a), all of the five members had their peak hour with respect to tweeting activity within 20 hours of their initial tweet. Three of the clone set members had their peak within the first three hours. It is interesting that the Reuters clone set member had a sequence of smaller peaks following its initial, highest peak. This may reflect multiple cascades of tweeting activity triggered by influential tweeters. A sequence of multiple peaks is also seen for the CNN member of clone set 14, in Figure 4.16(b). In other cases, for example for the ABC and Yahoo members of clone set 14, there is a single very dominant peak.

In Figure 4.17, unlike the earlier figures, times of tweets are considered relative to the time of the earliest tweet for any member of the clone set. This can illustrate the impact on tweet timing of the publication of other clone set members. Figure 4.17 shows the CDF of the relative tweet timing for the same five members of each of four clone sets as considered for Figures 4.13-4.16. As seen in Figure 4.17(a), all the members of clone set 13 were published within the first 16 hours after the earliest member was published. Note that the last and second last published clone set members received most of their tweets very quickly, whereas it took the clone set members that were published earlier a substantially longer period of time to acquire the majority of their tweets. Similar differences in the tweet timing for early and late published clone set members are seen for clone set 14 in Figure 4.17(b). Two of its members were published more than 30 hours after the earliest clone set member, and again, they received most of their tweets very quickly. Also for clone set 23, as seen in Figure 4.17(d), the clone set member published last acquired most of its tweets soon after publication. For clone set 18 this effect is also present, as seen in Figure 4.17(c), although in this case tweeting activity for some of the earlier published clone set members is spread out over a long time period than for the other clone sets.

For all the clone sets considered in Figure 4.17, the clone set members were published within a day and a half (and typically sooner) of the earliest clone set member. Differences in publication time may be due to time zone differences between different sites. On the other hand, publication times are often similar since whenever a story is published by a news agency, then other news sites get notified and may quickly cover the story. When publication times are close, it is found that often clone set members require similar periods of time to acquire most of their tweets. For clone set members published much later then the earliest clone set member, however, most tweets are acquired relatively quickly. The reason behind this could be due to people learning of events from the earlier published clone set members or from social networks, but waiting until they can read the news from their favorite news source and then quickly tweeting. Another reason could be

that the news story is nearing the end of its lifetime (time period over which it is of interest).



(a) Clone Set 13

(b) Clone Set 14

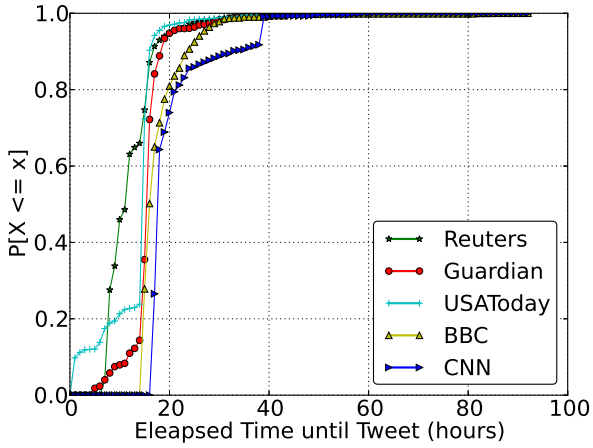
(c) Clone Set 18

(d) Clone Set 23

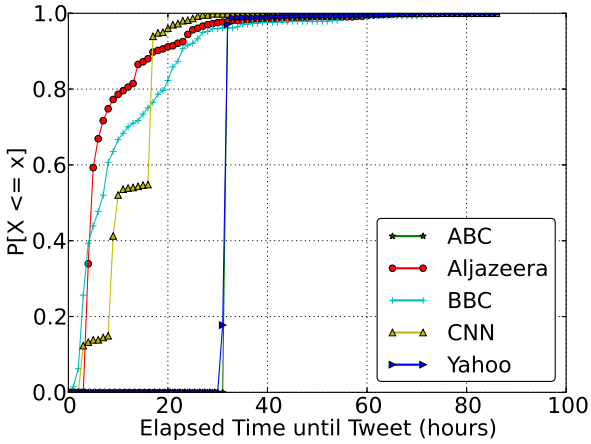
Figure 4.16: Number of Hourly Tweets for a Clone Set Member, for each Hour after the First Tweet for that Clone Set Member (Example Clone Sets, Top Five Sites for each Clone Set)

4.6 Number of Tweeter Followers

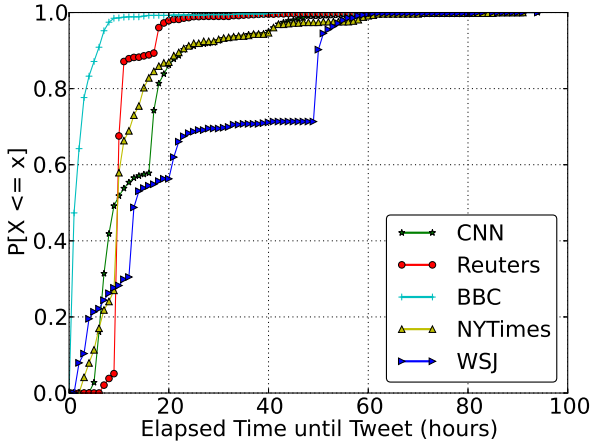
Some users, such as celebrities, have a huge number of followers, which could have a large impact on making a news site story popular. When such a user posts a link to a story in their status, their followers find that update in their news feed which may lead them to read the story and share/retweet it. Figure 4.18 shows the distribution of the maximum number of tweeter followers, among the tweeters for a clone set member. Note that, here instead of all 218 clone members only 158 clone members are used. The reason behind this is, all followers information is collected way later than the publication time of the clone members, by that



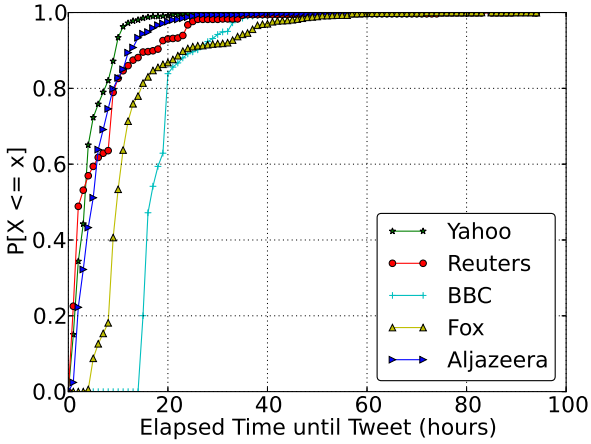
(a) Clone Set 13



(b) Clone Set 14

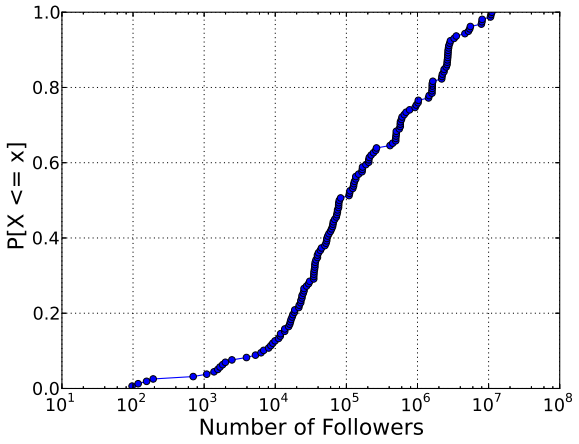


(c) Clone Set 18

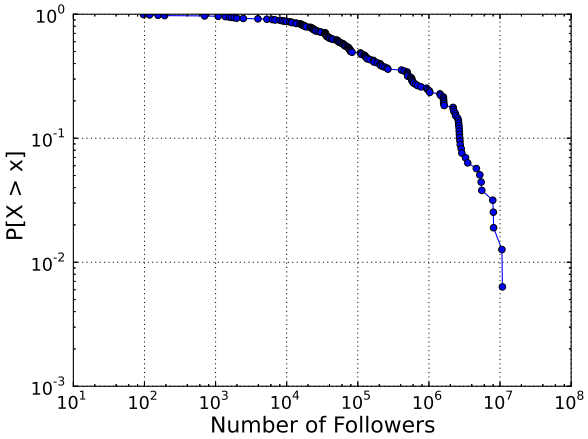


(d) Clone Set 23

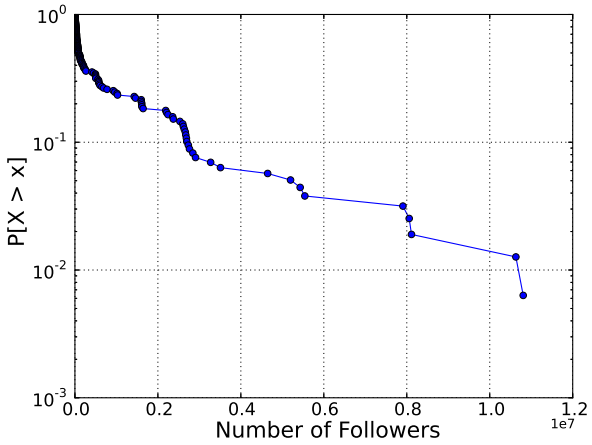
Figure 4.17: Cumulative Distribution Function of Elapsed Time from first Tweet for any Member of the Clone Set, to Times of Clone Set Member Tweets (Example Clone Sets, Top Five Sites for each Clone Set)



(a) CDF



(b) CCDF



(c) CCDF, with x-axis on linear scale

Figure 4.18: Distribution of Maximum Number of Tweeter Followers, among the Tweeters for a Clone Set Member (158 out of 218 Clone Members)

time some of the tweeters of clone members left the Twitterverse. Here only 158 clone members' all of the tweeters exist during the followers information crawling, that are been used for maximum number of followers distribution. From the CDF shown in Figure 4.18(a), it is found that for about 22% of the clone set members, the tweeter with the most followers has more than 1,000,000 followers. Plotting the CCDF with a linear scale on the x-axis, in Figure 4.18(c), reveals a roughly linear form except for the head of the distribution (small numbers of followers). This suggests that the distribution of maximum number of followers may resemble an exponential distribution over much of its range.

Some basic characteristics of the tweeter with the most followers, and of the tweeter with the second most followers, are shown for the members of the four clone sets chosen for individual analysis in Tables 4.3 through 4.10. Each table gives, for each clone set member, the tweeter name, the number of followers of that tweeter, the number of friends (number of other Twitter users that the tweeter is following), the percentage of the total tweets for that clone set member that were made by the tweeter, and the percentage of the total redundant tweets for that clone set member that were made by the tweeter. The last two statistics give some indication of the extent to which redundant tweets are due to the tweeter, and to what extent the tweeter's tweets direct impact the total tweet count.

Table 4.3: Characteristics of the Clone Set Member Tweeter with the Most Followers (Clone Set 13)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
Reuters	digg	1429419	181	0.07	46
Guardian	guardian	1027493	1065	0.23	7.9
USAToday	USATODAY	503493	549	0.44	6.36
BBC	debsylee	135561	89050	0.12	5.22
CNN	cnnbrk	10806555	77	0.18	5.98
Aljazeera	AJEnglish	1610552	152	0.25	3.23
ABC	ABC	2208410	641	0.85	5.1
Yahoo	YahooNews	572519	1209	0.52	4.2
Straitstimes	YahooNews	572517	1209	0.61	1.79
NJ	njdotcom	36089	335	2.4	4.8
CBSNews	CBSThisMorning	73335	327	2.5	-
Timesofindia	tinucherian	222767	102554	2.6	7.89
NBC	DrEkarin	6873	1975	2.9	8.82
Thetimes	thetimes	64813	372	5.55	5.55
Channelnewsasia	ChannelNewsAsia	126099	75	6.25	-
PBS	1WilliamGold	21735	21651	6.25	-

As seen in these tables, most of the members of each of the four clone sets have tweeters with large

Table 4.4: Characteristics of the Clone Set Member Tweeter with the Second Most Followers (Clone Set 13)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
Reuters	Tomhall	140585	142261	0.06	46
Guardian	YahooNews	611364	1766	0.08	7.9
USAToday	BODIESOFLIGHT	148234	151075	0.11	6.36
BBC	Marcome	84702	33317	0.12	5.22
CNN	KamalFaridi	166002	1066	0.17	5.98
Aljazeera	AJELive	378489	220	0.75	3.23
ABC	GMA	1973692	44	0.85	5.1
Yahoo	kwikermoney	52643	46608	0.52	4.2
Straitstimes	Esquiremag	161609	2782	0.6	1.79
NJ	starledger	26571	243	1.2	4.8
CBS	PorteableCom	68800	1301	0.02	-
Timesofindia	DiptiKantaMohap	14679	14693	2.63	7.89
NBC	AAR_FreightRail	4120	3419	2.9	8.82
Thetimes	muradahmed	4621	719	5.55	5.55
Channelnewsasia	AlenKarabegovic	998	-	6.25	-
PBC	NewsHourWorld	1888	81	6.25	-

Table 4.5: Characteristics of the Clone Set Member Tweeter with the Most Followers (Clone Set 14)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
ABC	UnivisionNews	74693	475	0.07	95
Yahoo	jewelrymandave	25394	16057	0.07	95
Aljazeera	AJEnglish	1614042	152	0.13	2.7
CNN	KamalFaridi	165841	1067	0.14	3.65
BBC	Gomazed	36040	76	0.24	3.35
Reuters	Reuters	2672584	1021	0.36	4.36
CBC	HowardRoper	34754	32159	0.95	15.84
Rawstory	CynthiaY29	77571	75406	1.04	5.2
Washingtonpost	HowardRoper	34753	32143	2.27	2.27
Csmonitor	jaketapper	258862	2024	2.56	-
Dawn	emkwan	35593	1097	5.88	2.94
SMH	twowisegals	23698	21043	6.06	3.03
MSN	tapferkeit	2456	2184	9.09	9.09
NYTimes	somegibberish	118	310	1	-

Table 4.6: Characteristics of the Clone Set Member Tweeter with the Second Most Followers (Clone Set 14)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
ABCNews	bossbev	5574	5521	0.07	95
Yahoo	soniamossri	19570	3274	0.07	95
Aljazeera	NickKristof	1402537	561	0.14	2.7
CNN	breakingnews_90	76356	28227	0.15	3.65
BBC	PeteLinforth	16022	16218	0.24	3.35
Reuters	nycjim	84058	1154	0.36	4.36
CBC	realestatefeeds	13311	12130	0.99	15.84
Rawstory	RawStory	31619	1740	1.04	5.2
Washingtonpost	profsubramanian	12334	11839	1.04	2.27
Csmonitor	globalreportorg	63437	134	2.56	-
Dawn	yasmeen_9	4386	501	2.94	2.94
SMH	smhonline	2764	1	3.03	3.03
MSN	4ahealthyhabit	603	1789	9.09	9.09

Table 4.7: Characteristics of the Clone Set Member Tweeter with the Most Followers (Clone Set 18)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
CNN	cnnbrk	11109098	77	2.10	95.79
Reuters	Reuters	2761795	1055	1.07	84.9
BBC	BBCNews	1460198	88	1.5	10.6
NYTimes	nytimes	8113386	742	2.92	7.21
WSJ	WSJ	2696313	678	6.7	11.78
Yahoo	kwikermoney	53202	47639	0.30	10.27
Guardian	USRealityCheck	79591	74996	1.45	1.45
CBC	GhostRiderRadio	18574	859	1.51	3.30
NBC	GlobalNewsDaily	9226	2108	6.67	20

Table 4.8: Characteristics of the Clone Set Member Tweeter with the Second Most Followers (Clone Set 18)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
CNN	uhprensagrafica	12258	826	0.06	95.79
Reuters	nycjim	86988	1172	1	84.9
BBC	sondakika_haber	106650	2424	1	10.6
NYTimes	Politicbody	249953	2936	0.17	7.21
WSJ	HamzeiAnalytics	130987	88	0.20	11.78
Yahoo	TrendingReport	379801	1132	0.30	10.27
Guardian	lovesey	23727	18048	1.45	1.45
CBC	Canada_Business	15923	17516	1.44	3.30
NBC	Truthbuster	6852	5990	6.67	20

Table 4.9: Characteristics of the Clone Set Member Tweeter with the Most Followers (Clone Set 23)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
Fox	foxnews	2661279	289	1.22	8.76
Yahoo	YahooNews	595382	1515	1.39	6.59
Reuters	Reuters	2907406	1060	1.39	16.25
BBC	BBCWorld	3508863	58	0.13	9.81
Aljazeera	AjEnglish	1643496	153	0.15	26.57
Guardian	guaraian	960794	1059	0.22	10.58
USAToday	Usatoday	554439	569	.22	7.98
Huffington	peoplesearches	265131	24281	0.32	3.52
CBS	Marapolsa	30502	6832	1.28	45.5
ABC	frmheadtotoe	49874	243	0.67	5.37
CNN	HowardRoper	22961	24693	2.38	-
News	newscomauHQ	54751	2476	10	30

Table 4.10: Characteristics of the Clone Set Member Tweeter with the Second Most Followers (Clone Set 23)

Clone Member	Tweeter Name	#Followers	#Friends	% of Tweets	% of Redundant Tweets
Fox	DRUDGE_REPORT	512558	2	0.06	8.76
Yahoo	cesdrilon	317721	1226	0.09	6.59
Reuters	copano	523950	16158	0.11	16.25
BBC	InternetGuru798	74521	56088	0.13	9.81
Aljazeera	AJELive	398945	224	3.69	26.57
Guardian	guardianworld	105781	320	0.43	10.58
USAToday	RickSanchezTV	136194	12641	0.22	7.98
Huffington	gamesdotcom	21761	1832	0.32	3.52
CBS	kgbt	16710	1635	1.28	45.5
ABC	Ink_Garage	26004	12477	0.67	5.37
CNN	laurabergerol	7829	8602	2.38	-
News	AntNom	50126	49651	3.33	30

numbers of followers. In some cases, the tweeter with the most or the second most followers is a Twitter account associated with the news site. Most of the popular news sites maintain their own social networking account, and they use that account in order to propagate their important news to their followers so as to increase the traffic to their site.

It is also found that the tweeters with the most or the second most followers have a very limited number of friends compared to their total number of followers, and make only a small fraction of the total tweets for the clone set member. Also, usually (but not always) the number of redundant tweets made by the tweeter with the most or second most followers is very low compared to the total number of redundant tweets. This suggests that most of the redundant tweets are usually not made by highly influential users, but possibly instead correspond to conversational exchanges among ordinary users regarding these news items, where they tag the news line while posting their tweets.

Chapter 5

Factors Impacting Popularity

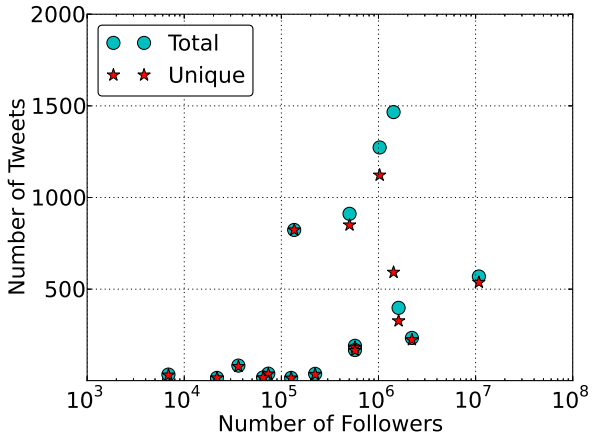
In this chapter, several factors are considered which might impact the ultimate popularity of the clones of online news content. Whenever one important incident happens, a number of news sites cover that incident by publishing their own report. Sometimes news publishers share the same news content from popular news agencies such as Reuters and Associated Press. Though different clone set members share identical or nearly identical news stories, their level of popularity in terms of the total number of tweets or unique tweeter tweets they receive for their stories are sometimes very different. This motivates investigating the factors that make an impact on the popularity of that online content. In this chapter, several factors are considered to investigate their impact on the data set, and their contribution in making some online news clones more popular than others. Specifically, the maximum number of tweeter followers, the content provider's popularity, early publishing, news site promotional tweets, and number of redundant tweets are considered to see their impact on making an online news clone popular.

Section 5.1 concerns the impact of the maximum number of tweeter followers in making online content popular. Section 5.2 considers the content provider's i.e. news site's popularity. Section 5.3 studies the impact of publication time. Section 5.4 analyzes the impact of the news site's promotional tweets, and finally Section 5.5 deals with the factor of redundant tweets.

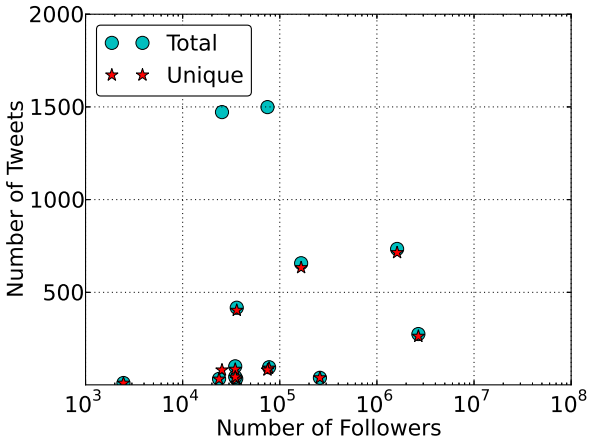
5.1 Maximum Number of Tweeter Followers

In this section, the impact of the maximum number of tweeter followers in making one clone popular among the clone set members is investigated. Twitter is a unidirectional social networking site where people can follow anyone as they wish without the relationship necessarily being reciprocated. Various organizations use Twitter as the fastest communication medium to reach interested users about their product. Twitter users can follow any other Twitter users, including ordinary people, celebrities, and users representing organizations, to get up-to-date information from these individuals and organizations. When a tweet is made, all of the tweeter followers get that tweet in their news feed. So if the tweeter of a news link has a large number of followers, that might result in a positive impact for that news link's ultimate popularity.

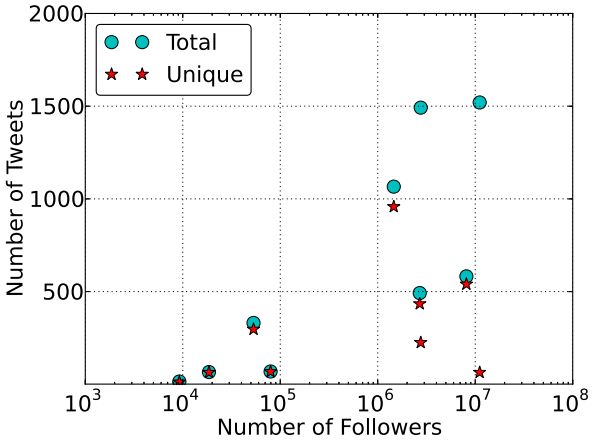
Figure 5.1 shows the relationship between the maximum number of tweeter followers, among the tweeters for a clone set member, and both the total number of tweets and the number of tweets from unique tweeters,



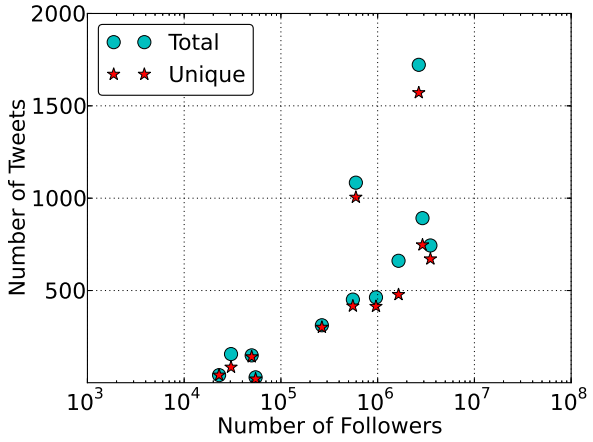
(a) Clone Set 13



(b) Clone Set 14



(c) Clone Set 18



(d) Clone Set 23

Figure 5.1: Relationship between Maximum Number of Followers of Tweeters and Content Popularity (Example Clone Sets)

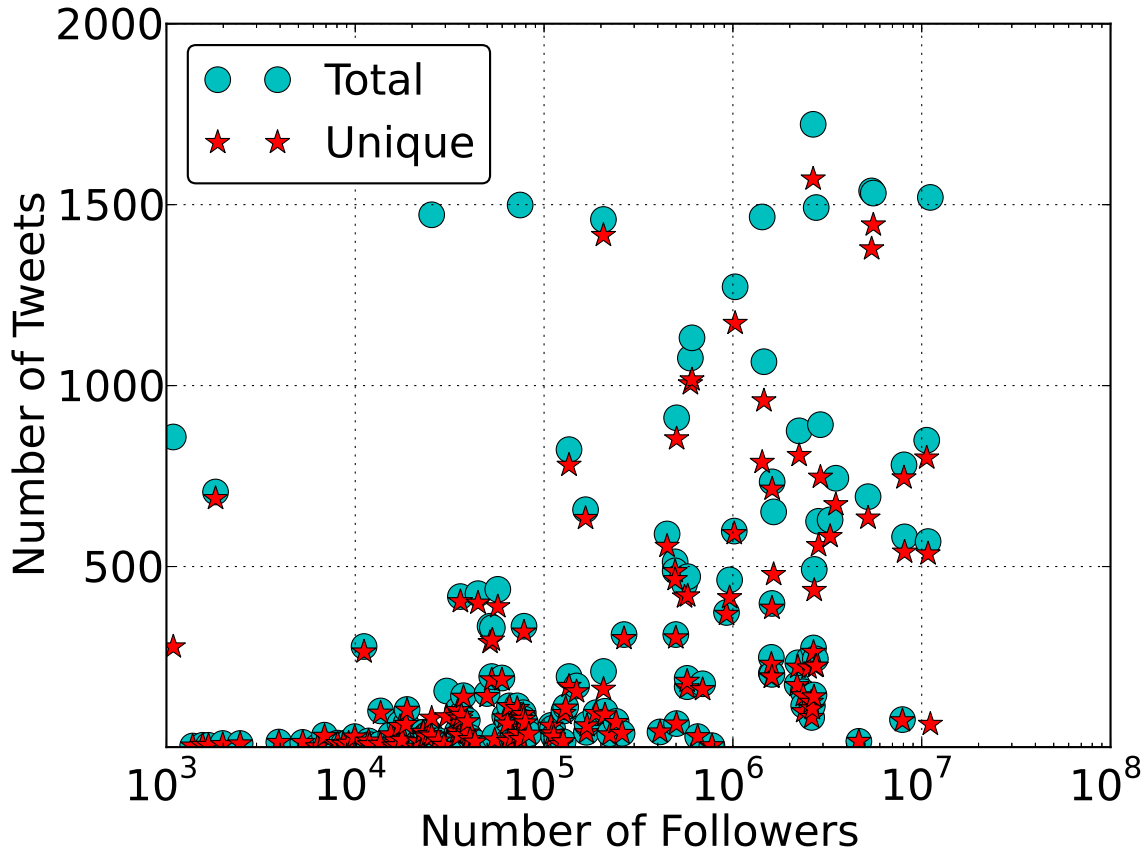


Figure 5.2: Relationship between Maximum Number of Followers of Tweeters and Content Popularity (158 out of 218 Clone Members)

for the members of the four clone sets chosen for individual analysis. For each clone set member, two points are plotted, with the x coordinate of both points being the maximum number of tweeter followers, among the tweeters for that clone set member, and the y coordinate being either the total number of tweets or the total number of unique tweeter tweets. Although there is much variability, there appears to be a tendency for an increasing maximum number of tweeter followers to correspond to increasing numbers of total and unique tweeter tweets.

Figure 5.2 has the same form as in Figure 5.1, but now points are plotted for 158 of the members of all of the clone sets. Here only 158 out of 218 clone members are used because of the unavailability of all followers data, discussed in section 4.6. As in Figure 5.1, there is much variability, but again an increasing maximum number of tweeter followers often corresponds to increasing numbers of total and unique tweeter tweets.

Table 5.1 quantifies these relationships by giving correlation coefficients between the maximum number of tweeter followers, and the numbers of total and unique tweeter tweets. The table gives results for the four clone sets chosen for individual analysis, as well as results calculated over all clone sets (158 out of 218 clone

Table 5.1: Correlation Coefficients for Maximum Number of Followers of Tweeters Versus Content Popularity

	Pearson Correlation Coefficient		Spearman Correlation Coefficient	
Clone Set	Total Tweets	Unique Tweets	Total Tweets	Unique Tweets
13	0.21	0.24	0.72	0.68
14	0.05	0.46	0.54	0.75
18	0.63	0.05	0.9	0.39
23	0.71	0.68	0.84	0.83
All(158 out of 218 clone members)	0.43	0.385	0.77	0.78

set members). For the Pearson correlation coefficients, raw data values are used, while for the Spearman correlation coefficients the rank of each variable within the clone set is used. Denoting the maximum number of tweeter followers for a clone set member by random variable m, the total number of tweets for a clone set member by random variable t, the number of unique tweeter tweets by random variable u, the rank of a clone set member within its clone set with respect to maximum number of tweeter followers by random variable r_m, the rank with respect to total tweets by random variable r_t, and the rank with respect to unique tweeter tweets by another random variable r_u, the Pearson correlation coefficients are calculated using the value of m and t or u, whereas the Spearman correlation coefficients are calculated between the values of r_m and r_t or r_u. The Spearman correlation coefficient is more suitable for this analysis for several reasons. First, the relationships between the maximum number of tweeter followers and the numbers of total and unique tweeter tweets may not be linear. Second, the Spearman correlation coefficient is less sensitive than the Pearson correlation coefficient to outliers. Finally, and perhaps most importantly, by ranking within clone sets (rather than across clone sets or using the raw data values), in the overall analysis, differences in content are controlled for.

As seen in Table 5.1, positive correlations are observed between the maximum number of tweeter followers, and the numbers of total and unique tweeter tweets. The Spearman correlation coefficients are larger than the Pearson correlation coefficients, reflecting the Spearman correlation coefficient's lower sensitivity to outliers and the fact that it measures the extent to which a monotone relationship exists, rather than the more restrictive condition of a linear relationship.

The correlation coefficients for total tweets are similar to those for unique tweeter tweets for clone sets 13 and 23, and when calculated over all clone sets. For clone sets 14 and 18, however, the correlation coefficients for total tweets and those for unique tweeter tweets are quite different. This can be explained by the fact that for each of these clone sets, the ranking of the clone set members with respect to total tweets is quite different than that with respect to unique tweeter tweets, as can be seen from Figures 4.6 and 4.7. Interestingly, the highest correlation coefficients of each type, with the exception of the Spearman correlation coefficient for total tweets for clone set 18, are for clone set 23. As can be seen from Figures 4.5 - 4.8, clone set 23 has the

highest numbers of unique tweeter tweets among the four clone sets chosen for individual analysis.

5.2 News Site Popularity

The prominence of the online content provider could have a great impact on content popularity. A popular news site has regular readers whose tweets could result in higher popularity in the Twitterverse for its news stories. On the other hand, if news content is frequently found in Twitter's news feed or by any other means than surfing the actual news sites, the relative numbers of users regularly surfing particular news sites may have only a weak impact on popularity in the Twitterverse. The collected clone sets include both popular and relatively unpopular news sites. In the four clone sets chosen for individual analysis, four sites are common: BBC, Yahoo, Reuters and CNN. A fifth clone set, clone set 10, also includes these four sites. An analysis is carried out using the four clone set members from the BBC, Yahoo, Reuters and CNN sites, within each of the five clone sets 10, 13, 14, 18, and 23, so as to attempt to quantify the impact of overall site popularity on clone set member popularity.

Table 5.2 shows the relative ranking of BBC, CNN, Reuters and Yahoo clone set members in the different clone sets, with respect to the number of total tweets. The columns with headings 10, 13, 14, 18 and 23 give the relative rank of each news site's member within the respective clone set, according to the total number of tweets received. Next to each of these columns is a column giving the average rank of each news site's clone set members in the other four clone sets. For example, in the "Average without 10" column, the entry for CNN is 2.5, which is the average of 3 (rank in clone set 13), 2 (rank in clone set 14), 1 (rank in clone set 18), and 4 (rank in clone set 23). If overall site popularity was a major determinant of clone set member popularity, the rank of each news site's clone set member in clone set 10, for example, would be expected to correspond to the average rank given in the "Average without 10" column. Also shown, in the last column of the table, is a ranking of the news sites with respect to their relative overall popularity, derived from their respective Alexa[1] ranks. Table 5.3 includes the same columns, but uses the number of unique tweeter tweets rather than the number of total tweets when ranking the clone set members of the four news sites within each clone set.

Table 5.4 gives the correlation coefficient between the relative Alexa ranking of a news site and the relative ranking of its clone set member, for both ranking according to total number of tweets and ranking according to number of unique tweeter tweets. The table also gives these correlation coefficients with the average ranking of a news site's members in the other clone sets used in place of its relative Alexa ranking. As seen in the table, all correlation coefficients are very small except for one large negative correlation, between the average ranking of a news site's members in the other clone sets and the total number of tweets. This evidence does not support site popularity being a major determinant of clone set member popularity. However, note that all of the competitor news sites in this analysis are widely known, highly popular sites. It would seem likely

[1] http://www.alexa.com/topsites/category/Top/News; accessed on September 8, 2013

Table 5.2: Relationship between Content Popularity Rank, Average Popularity Rank of the Hosting Site and Relative Alexa Rank (Example Clone Sets and Four Common News Sites, Total Tweets)

News Site	10	Average without 10	13	Average without 13	14	Average without 14	18	Average without 18	23	Average without 23	Relative Alexa Rank
BBC	1	2.75	2	2.5	3	2.25	3	2.25	3	2.25	3
CNN	2	2.5	3	2.25	2	2.5	1	2.75	4	2	2
Reuters	4	2.25	1	3	4	2.25	2	2.75	2	2.75	4
Yahoo	3	2.5	4	2.25	1	3	4	2.25	1	3	1

Table 5.3: Relationship between Content Popularity Rank, Average Popularity Rank of the Hosting Site, and Relative Alexa Rank (Example Clone Sets and Four Common News Sites, Tweets from Unique Tweeters)

News Site	10	Average without 10	13	Average without 13	14	Average without 14	18	Average without 18	23	Average without 23	Relative Alexa Rank
BBC	1	2	2	1.75	2	1.5	1	2	3	1.5	3
CNN	2	3	3	2.75	1	3.25	4	2.5	4	2.5	2
Reuters	4	2.25	1	3	3	2.5	3	2.5	2	2.75	4
Yahoo	3	2.75	4	2.5	4	2.5	2	3	1	3.25	1

Table 5.4: Correlation Coefficients for Relative Alexa Rank and Average News Site Rank in Other Clone Sets Versus Clone Set Member Rank (Example Clone Sets and Four Common News Sites, Total Tweets and Tweets from Unique Tweeters)

	Correlation Coefficient using Average News Site Rank in Other Clone Sets	Correlation Coefficient using Relative Alexa Rank
All Tweets	-0.91	0
Tweets from Unique Tweeters	-0.18	-0.16

that when there are great differences in site popularity, there is more impact on clone set member popularity.

To test this hypothesis, a different group of clone sets was selected, with a somewhat different set of common news sites: BBC, Yahoo, ABC and CBC. The BBC and Yahoo news site are substantially more popular news sites than the ABC and CBC sites. Tables 5.5 and 5.6 give the same ranking information as Tables 5.2 and 5.3, respectively, but for this second selection of clone sets and common news sites.

Table 5.7 gives the same correlation coefficients, but for the new choices of news clone sets and news sites, as those given earlier in Table 5.4. With respect to correlation with relative Alexa rank, a large positive correlation of 0.68 is found in the case of all tweets and a correlation coefficient of 0.57 is found in the case of unique tweeter tweets. Large positive correlations are also found when considering the average rank of a news site's members in the other four clone sets rather than the relative Alexa rank. This analysis suggests that news site popularity can be an important factor in determining the relative popularities of news story clones on sites with widely-differing overall popularities. Among similarly-prominent sites, however, other factors may become much more important.

Table 5.5: Relationship between Content Popularity Rank, Average Popularity Rank of the Hosting Site and Relative Alexa Rank (Second Selection of Example Clone Sets and Four Common News Sites, Total Tweets)

News Site	10	Average without 10	14	Average without 14	17	Average without 17	20	Average without 20	23	Average without 23	Relative Alexa Rank
ABC	4	2.75	1	3.5	3	3	4	2.75	3	3	3
BBC	1	1.75	3	1.25	1	1.75	1	1.75	2	1.5	2
CBC	3	3.75	4	3.5	4	3.5	3	3.75	4	3.75	4
Yahoo	2	1.75	2	1.75	2	1.75	2	1.75	1	2	1

Table 5.6: Relationship between Content Popularity Rank, Average Popularity Rank of the Hosting Site and Relative Alexa Rank (Second Selection of Example Clone Sets and Four Common News Sites, Tweets from Unique Tweeters)

News Site	10	Average without 10	14	Average without 14	17	Average without 17	20	Average without 20	23	Average without 23	Relative Alexa Rank
ABC	4	3.375	3.5	3.5	3	3.625	4	3.375	3	3.625	3
BBC	1	1.25	1	1.25	1	1.25	1	1.25	2	1	2
CBC	3	3.25	2	3.5	4	3	3	3.25	4	3	4
Yahoo	2	2.125	3.5	1.75	2	2.125	2	2.125	1	2.375	1

Table 5.7: Correlation Coefficients for Relative Alexa Rank and Average Site Rank in Other Clone Sets versus Clone Set Member Rank (Second Selection of Example Clone Sets and four Common News Sites, Total Tweets and Tweets from Unique Tweeters)

	Correlation Coefficient using Average News Site Rank in Other Clone sets	Correlation Coefficient using Relative Alexa Rank
All Tweets	0.58	0.68
Tweets from Unique Tweeters	0.72	0.57

5.3 Time of Publication

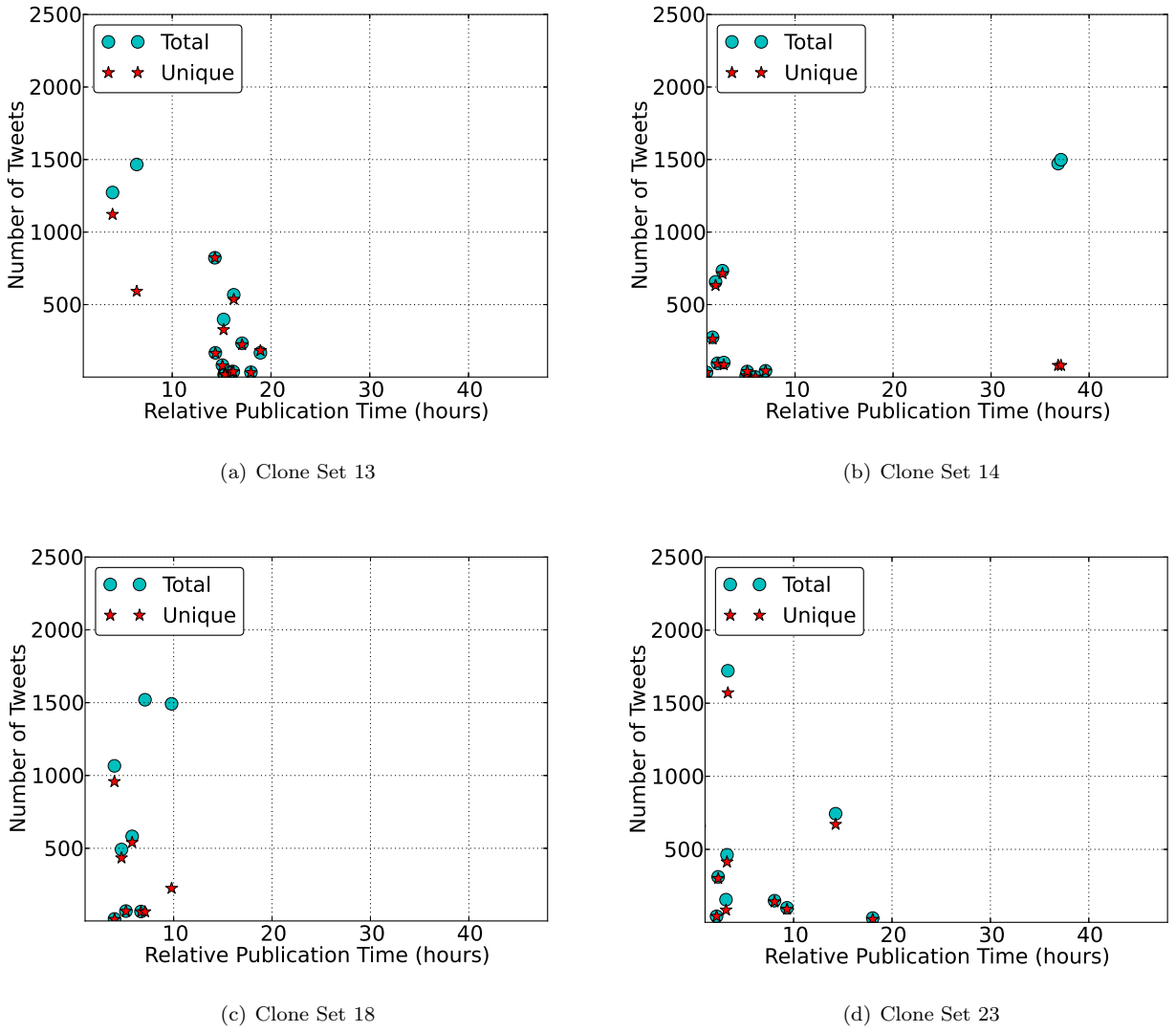
The popularity of online content could be dependent on its publication time. Early published content may get viewed earlier and recommended as well as propagated through social networks. However, online news content has a short lifetime, and when a news event occurs the stories reporting it are all published within a short time period. As described in Chapter 4, for the clone sets in the collected data it was found that different news providers took a maximum of two days to publish their own clone of a news story and around 95% of them published within a day of the first publisher of the story. The delay until different clone set members were published is very short compared to other online content types such as user-generated video.

In order to find the impact of early publication of online news content the relative publication times of all of the members of each clone set are estimated using the relative timings of the first tweet to each clone set member. Figure 5.3 shows the relationship between the relative publication times estimated in this manner and both the total number of tweets and the number of tweets from unique tweeters, for the members of the four clone sets chosen for individual analysis. For each clone set member, two points are plotted, with the x coordinate of both points being the relative publication time of that clone set member within its clone set, and the y coordinate being either the total number of tweets or the total number of unique tweeter tweets. For clone set 13, a correlation is observed between getting a higher number of tweets and having a relatively early publication time. The first three published clone set members received more tweets and unique tweeter tweets than any of the clone set members that were published later.

For the other clone sets, however, the results shown in Figure 5.3 do not suggest that relative publication time has a large impact on relative popularity. For example, for both clone sets 14 and 18 the most total tweets were received by the two clone set members with the latest relative publication times (although note that many of these were redundant tweets). In the case of clone set 23, the most total and unique tweeter tweets were received by the clone set member that was the 10^{th} to be published.

Figure 5.4 has the same form as in Figure 5.3, but now points are plotted for all of the members of all of the clone sets. This figure does suggest some correlation between early relative publication time and numbers of total and unique tweeter tweets, but from comparison with Figure 5.2 the relationship seems somewhat

weaker than that between maximum number of tweeter followers and numbers of total and unique tweeter tweets.



(a) Clone Set 13

(b) Clone Set 14

(c) Clone Set 18

(d) Clone Set 23

Figure 5.3: Relationship between Relative Publication Time and Content Popularity (Example Clone Sets)

Table 5.8 gives correlation coefficients between the relative publication time and the numbers of total and unique tweeter tweets. The table gives results for the four clones sets chosen for individual analysis, as well as results calculated over all clone sets. As in Table 5.1, both Pearson correlation coefficients are given, which use the raw data values, and Spearman correlation coefficients, which use the rank of each variable within the respective clone set. For the same reasons as described for Table 5.1, the Spearman correlation coefficient is more suitable for this analysis. Note that if relative publication time has substantial impact on clone set member relative popularity, one would expect to see large magnitude, negative correlations, since a greater relative publication time would tend to imply smaller numbers of tweets. As seen in the table, some

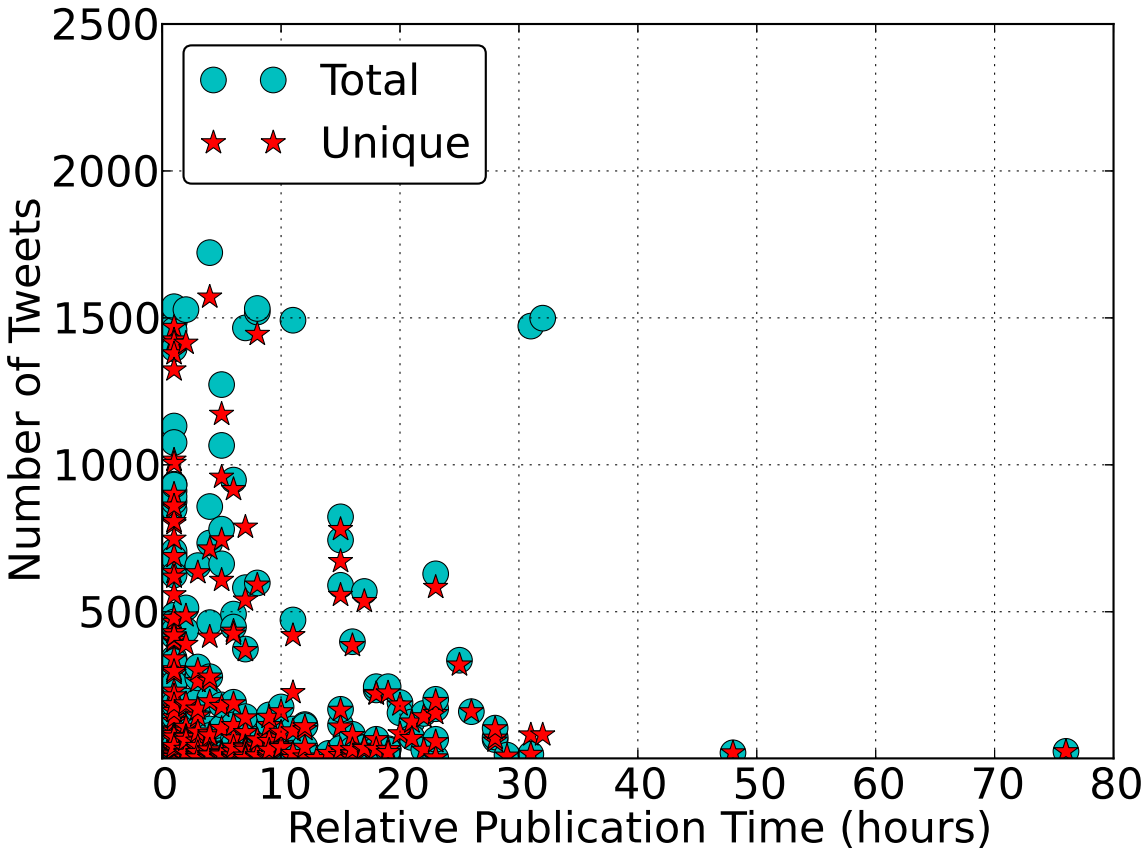


Figure 5.4: Relationship between Relative Publication Time and Content Popularity (All Clone Sets)

Table 5.8: Correlation Coefficients for Relative Publication Time Versus Content Popularity

	Pearson Correlation Coefficient		Spearman Correlation Coefficient	
Clone Set	All Tweets	Unique Tweeter Tweets	All Tweets	Unique Tweeter Tweets
13	-0.79	-0.79	-0.47	-0.48
14	0.72	-0.21	-0.02	-0.3
18	0.46	-0.15	0.3	-0.13
23	-0.38	-0.37	-0.45	-0.51
All	-0.11	-0.21	- 0.51	-0.55

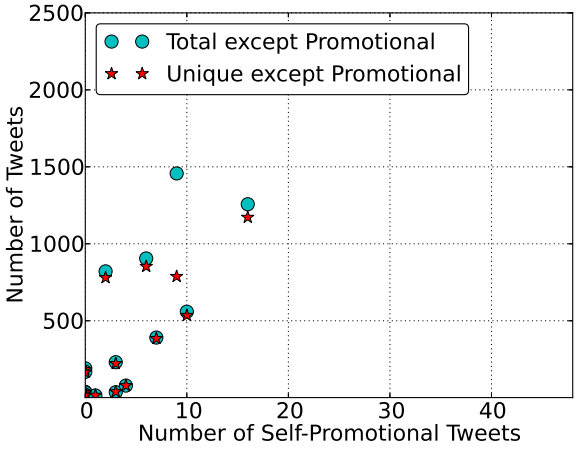
positive correlation coefficients are found for clone sets 14 (Pearson coefficient for all tweets) and clone set 18 (both the Pearson and Spearman coefficients for all tweets). This can be explained by the fact that for both of these clone sets, the most total tweets were received by the two clone set members with the latest publication times, as observed earlier when discussing Figure 5.3. All other correlation coefficients presented in the table are negative, however, as expected. The Spearman correlation coefficients calculated over all clone sets, in particular, suggest a significant correlation between relative publication time and clone set member popularity. However, the magnitudes of these correlation coefficients are significantly smaller than those observed for maximum number of tweeter followers in Table 5.1.

5.4 Self-Promotional Tweets

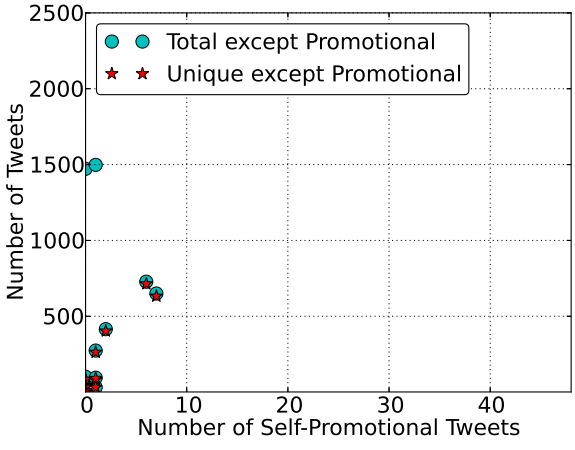
Self-promotional tweets could be an important factor in making a news site's stories more popular. Many news sites maintain their own social networking accounts, through which they publicize their important news stories. In particular, in the collected data set it is observed that some news sites maintain a Twitter account that is used for tweeting their news stories. Sometimes the same story is tweeted multiple times, sometimes in a regular pattern such as every hour. In this section, self-promotional tweets are removed from the counts of total tweets and unique tweeter tweets, and the impact of self-promotional tweets on content popularity is investigated.

Figure 5.5 shows the relationship between the number of self-promotional tweets and both the total number of tweets (excluding the self-promotional tweets) and the number of tweets from unique tweeters (excluding the self-promotional tweeters), for the members of the four clone sets chosen for individual analysis. For each clone set member, two points are plotted, with the x coordinate of both points being the number of self-promotional tweets for that clone set member, and the y coordinate being either the total number of tweets or the total number of unique tweeter tweets. In the case of clone set 14, there are relatively few self-promotional tweets, making it difficult to see any relationship. For clone sets 13, 18 and 23, however, it appears that often both the total number of tweets and the number of unique tweeter tweets increase with increasing numbers of self-promotional tweets. Figure 5.6 has the same form as in Figure 5.5, but now points are plotted for all of the members of all of the clone sets. Note that for most of the clone set members, the number of self-promotional tweets is under 10. The figure suggests that there may be some correlation between number of self-promotional tweets and content popularity.

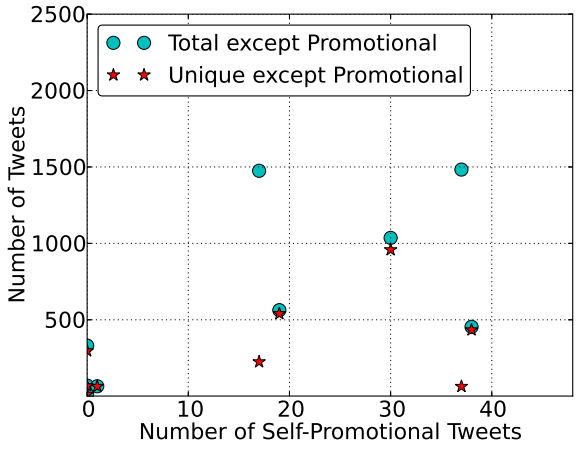
Table 5.9 gives correlation coefficients between the self-promotional tweets, and the numbers of total and unique tweeter tweets excluding self-promotional tweets. As before, both Pearson correlation coefficient are given, which use the raw data values, and Spearman correlation coefficients, which use the rank of each variable within the respective clone set. For the same reasons as described earlier, the Spearman correlation coefficient is more suitable for this analysis. As seen in the table, substantial positive correlation are observed, of greater magnitude than those seen in Table 5.8.



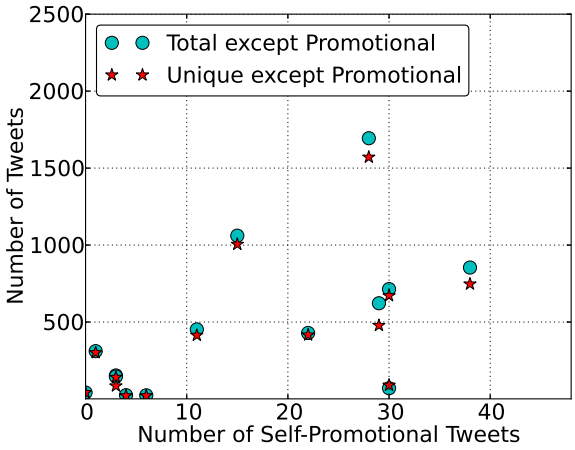
(a) Clone Set 13



(b) Clone Set 14



(c) Clone Set 18



(d) Clone Set 23

Figure 5.5: Relationship between Number of Self-Promotional Tweets and Content Popularity (Example Clone Sets)

Table 5.9: Correlation Coefficients for Number of Self-Promotional Tweets Versus Content Popularity

	Pearson Correlation Coefficient		Spearman Correlation Coefficient	
Clone Set	All Tweets except Promotional	Unique Tweeter Tweets	All Tweets except Promotional	Unique Tweeter Tweets
13	0.78	0.79	0.73	0.73
14	0.28	0.94	0.53	0.76
18	0.69	0.75	0.76	0.69
23	0.59	0.58	0.58	0.61
All	0.54	0.54	0.70	0.72

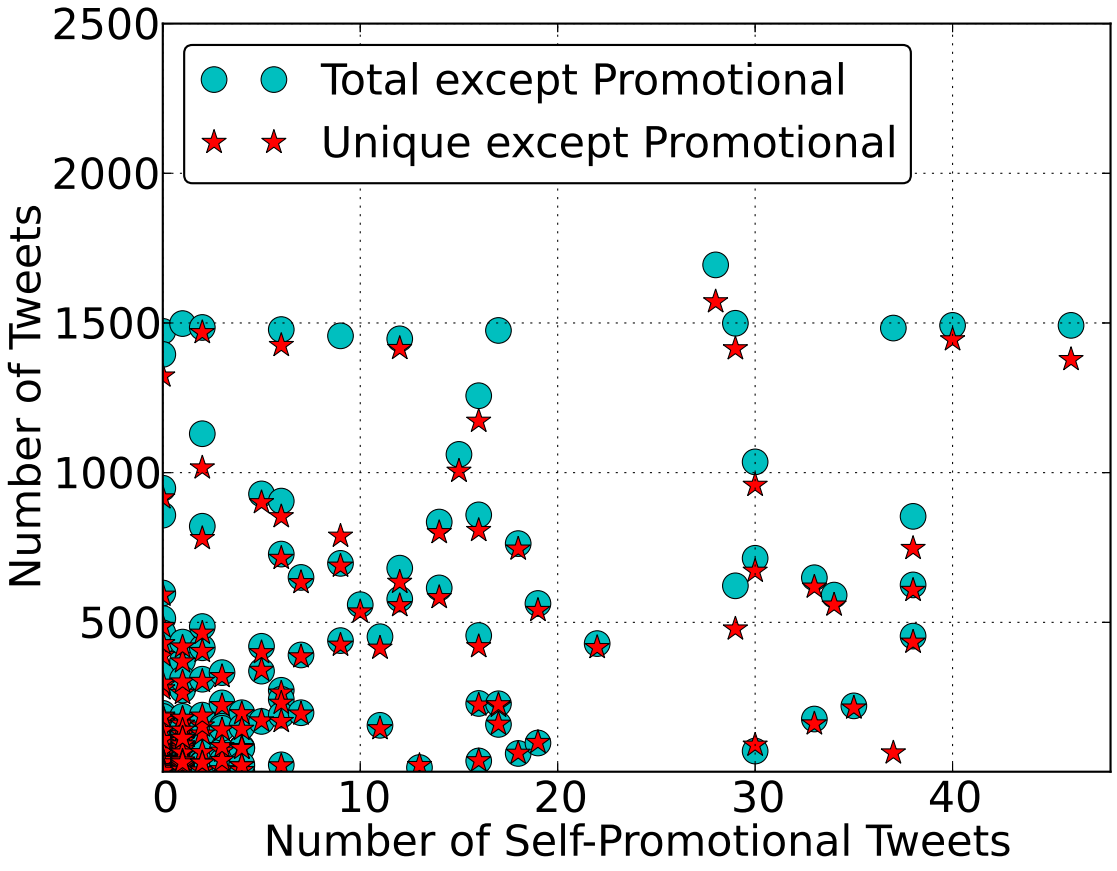


Figure 5.6: Relationship between Number of Self-Promotional Tweets and Content Popularity (All Clone Sets)

5.5 Redundant Tweets

This section consider possible correlations between the number of redundant tweets and the number of unique tweeter tweets. Note that the number of total tweets cannot be used as another measure of popularity in this analysis, since the number of total tweets is simply the number of unique tweeter tweets plus the number of redundant tweets. Figure 5.7 shows the relationship between the number of redundant tweets and the number of unique tweeter tweets for the members of the four clone sets chosen for individual analysis.

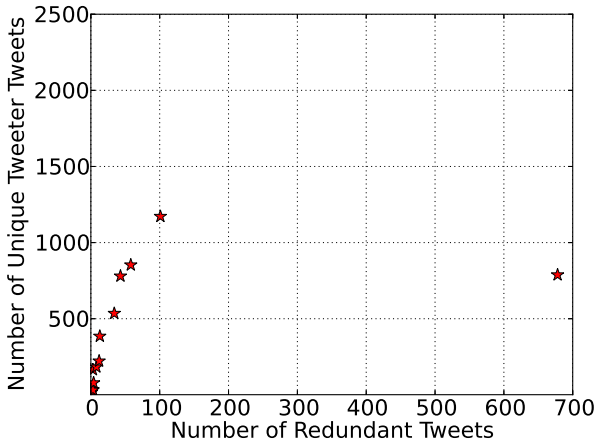
For clone sets 13 and 23 it appears that often there is a higher number of unique tweeter tweets when there is a higher number of redundant tweets. For clone sets 14 and 18, note that two clone set members in each case have a very high number of redundant tweets and yet a small number of unique tweeter tweets.

Figure 5.8 has the same form as in Figure 5.7, but now points are plotted for all of the members of all of the clone sets. Note that for numbers of redundant tweets of at most 200, there seems to be a clear trend of increasing numbers of redundant tweets being correlated with increasing numbers of unique tweeter tweets. Most clone set members have a number of redundant tweets falling into this range, with only 6 clone set members getting more than 200 redundant tweets.

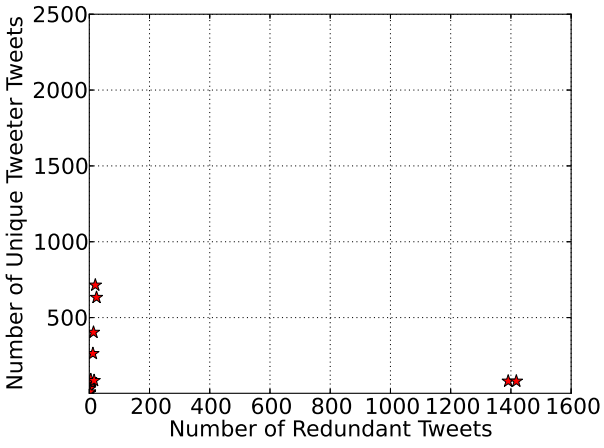
Table 5.10 gives correlation coefficients between the number of redundant tweets and the number of unique tweeter tweets for the four clone sets chosen for individual analysis as well as over all clone sets. As before, both Pearson correlation coefficients are given, which use raw data values, and Spearman correlation coefficients, which use the rank of each variable with the respective clone set. For the same reasons described earlier, the Spearman correlation coefficient is more suitable for this analysis. In all cases except for clone set 13, the Spearman correlation coefficient is positive and of substantial magnitude.

Table 5.10: Correlation Coefficients for Number of Redundant Tweets Versus Number of Unique Tweeter Tweets

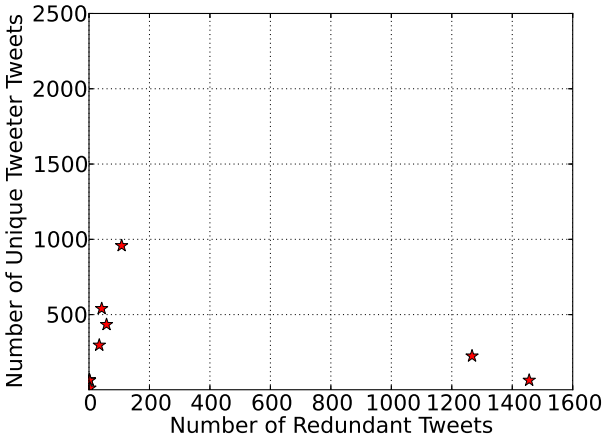
Clone Set	Pearson Correlation Coefficient	Spearman Correlation Coefficient
13	0.47	0.95
14	-0.16	0.70
18	-0.23	0.31
23	0.71	0.81
All	0.10	0.86



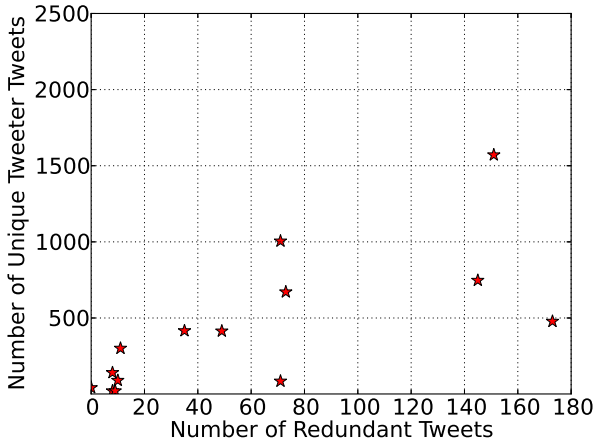
(a) Clone Set 13



(b) Clone Set 14



(c) Clone Set 18



(d) Clone Set 23

Figure 5.7: Relationship between Number of Redundant Tweets and Number of Unique Tweeter Tweets (Example Clone Sets)

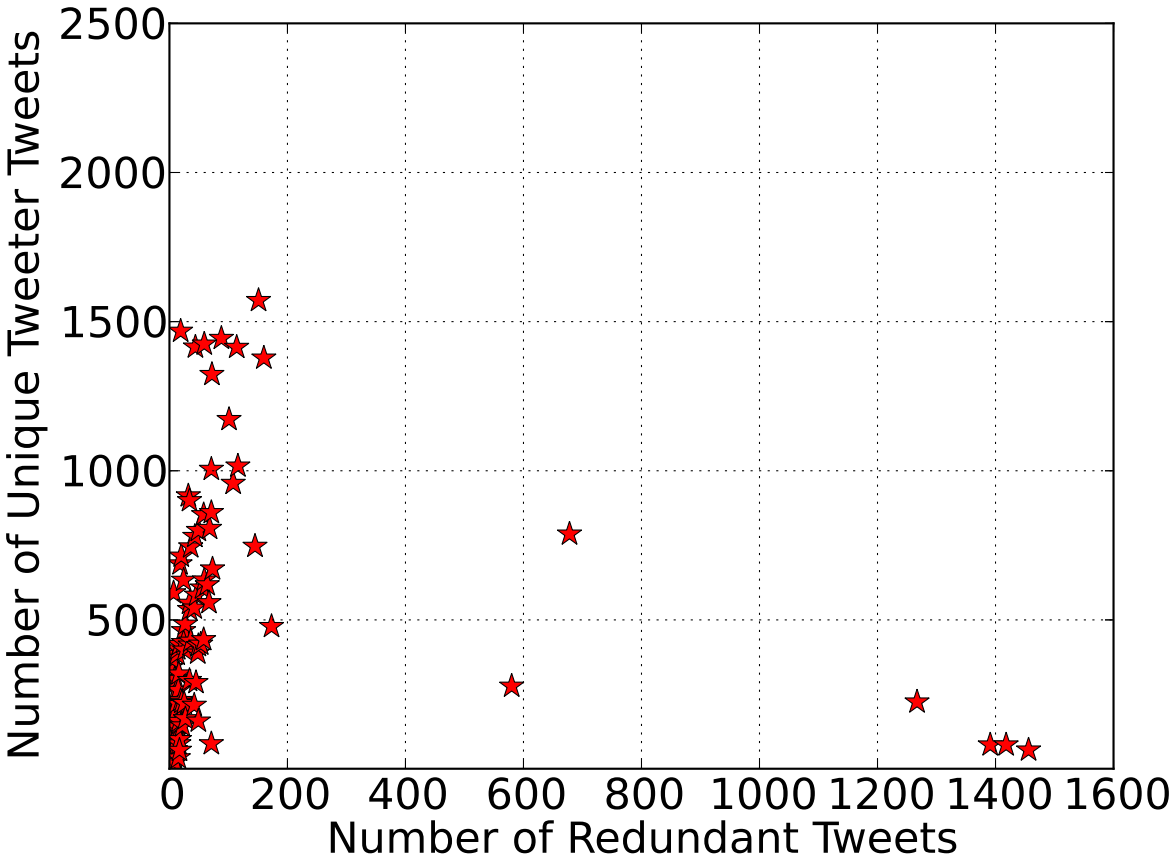


Figure 5.8: Relationship between Number of Redundant Tweets and Number of Unique Tweeter Tweets (All Clone Sets)

Chapter 6

Conclusions and Future Work

With the advancement of technology people like to get updated news content via the Internet and also like to inform their friends and followers about the interesting news and pass their opinion of that news via social networking sites like Twitter. Though identical or nearly identical news items are being published by a number of news site, their popularity varies due to many factors. In this thesis research, a dataset was collected containing the tweets made for the 218 members of 25 distinct sets of news story clones. The collected data was analyzed with respect to basic popularity characteristics concerning number of tweets of various types, relative publication times of clone set members, tweet timing, and numbers of tweeter followers. This data set was then used in an investigation of the factors that might make some news story clones more popular than others. It was found that multiple content-agnostic factors may impact news site story popularity, and a first step was taken at quantifying their relative importance. Section 6.1 gives a summary of the thesis. Section 6.2 presents the contributions of the thesis, and Section 6.3 describes directions for the future work.

6.1 Summary

In this thesis, online news content popularity is measured through the associated tweeting activity in the popular social networking site Twitter. Identical or nearly identical news story versions are termed "clones" in this thesis. In this thesis 25 different news stories were chosen, and a total of 218 clones were collected. Each of the clone sets contains 2 to 16 clone set members. The news stories cover different types of news, including political, economic, technical, scientific and environmental.

Chapter 3 described the data collection methodology. News stories were chosen according to their initial popularity, and clones were found using Google search and by searching top 20 news sites. The Twitter Search API was then applied for each clone set member to get its fundamental tweet information, including tweet publication time, tweeter name, text of the tweet, language used and tweet ID. Later, a parser written in python was used to obtain the screen name of each tweeter from the Search API returned raw data file. Each user screen name was passed as the argument while calling the Get Users/Show API to collect detailed tweeter information. Some basic information from the raw data files was parsed and output to comma separated value (CSV) files that were used in further analysis.

In Chapter 4, data from all of the clone sets was used to study the distribution of the total number of tweets

for a clone set member, total unique tweeter tweets, total self-promotional tweets, total redundant tweets, tweet timing, and maximum number of tweeter followers. Relative publication time and news site popularity were also characterized in that chapter. Some of the top-tweeted clone sets were chosen for detailed analysis. It was found that the total numbers of tweets and unique tweeter tweets follow light-tail distributions. The distribution of the number of redundant tweets appeared to be heavy-tailed. The distribution of the number of self-promotional tweets was found to resemble the exponential distribution.

Chapter 4 also characterized the proportion of tweets in different categories for the clone sets chosen for individual analysis. It was found that for most (but not all) clone set members, most of the total tweets were unique tweeter tweets. Second highest portion of the total tweets were typically redundant tweets excluding the self-promotional tweets, whereas only a small portion of the total tweets were typically self-promotional tweets. Overall news site popularity was also investigated. Among the top 13 ranked sites with respect to the total number of tweets, and the top ranked 13 sites with respect to the number of unique tweeter tweets, 12 news sites were in common, but their positions in these ranking often differed. In this thesis the time of the initial tweet of a clone set member was used to estimate the publication time of that clone set member, due to the inaccessibility of the actual publication time for all clone set members. It was found that, almost 98% of clone set members were published within the first 30 hours of publication of the earliest member of the clone set. The distribution of the timing of each tweet, relative to the time of the first tweet for the respective clone set member, was found to have a form resembling that of an exponential distribution. Most tweets were made within the first day after publication. In case of the distribution of the maximum number of follower of tweeters, a form resembling that of an exponential distribution was found except for the head of the distribution corresponding to relatively small numbers of followers.

Chapter 5 deals with the investigation of some of the factors that make some online news clones more popular than others. Correlation coefficients were calculated using both the total number of tweets and the number of unique tweeter tweets as measures of popularity. Both the raw data values and the rank of each variable within the clone set were used to calculate the Pearson and Spearman correlation coefficients. A strong correlation was found between the maximum number of tweeter followers and clone set member popularity. In the case of overall news site popularity, it was found that the impact on clone set member popularity appeared to be small when considering news sites of similar prominence, whereas if considering a mix of both popular and relatively unpopular sites, then there appeared to be a substantial impact.

Relatively early publication of a clone set member appeared to have some positive impact on clone set member popularity, although the correlation was weaker than that found for maximum number of tweeter followers. Self-promotional tweets appeared to often have significant positive impact in making a clone set member more popular than others, while redundant tweets also showed some positive correlation with popularity. Overall, it was found that multiple content-agnostic factors may impact news site story popularity, and a first step was taken at quantifying their relative importance.

6.2 Thesis Contributions

The contributions of this thesis are listed below.

- A dataset was collected containing information for 25 clone sets each with 2 - 15 clone set members, with 218 members in total. Each clone set corresponds to a different news story, with the clone set members corresponding to identical or nearly identical versions of that news story posted on various news sites. For each news story version (clone set member) the tweets referencing that news story version were collected, as well as information concerning the tweeters.

- The collected data was analyzed with respect to its basic characteristics. It was found that both the number of tweets for a clone set member (version of a news story published on some Web site), and the elapsed time from its first tweet to the times of subsequent tweets, appear to have light-tailed distributions, reflecting the ephemeral popularity of news stories. The elapsed time distribution appears to have a similar form as the exponential distribution. With respect to the characteristics of the tweeters for a clone set member, it was found that the distribution of their maximum number of followers also appears to be light-tailed, with a similar form as the exponential distribution over most of its range.

- The possible impacts of various content-agnostic factors on clone set member popularity were investigated. Both the total number of tweets and the total number of unique tweeters were investigated as possible measures of popularity. A strong correlation (as measured by the Spearman correlation coefficient) was found between the popularity of a clone set member and the maximum number of followers of its tweeters, as well as with the number of self-promotional tweets (from a Twitter account associated with the news site). Relative publication time, and overall site popularity (at least among similarly-prominent sites), in contrast, appeared to have weaker correlations with popularity. These results suggest that in the Twitterverse, popularity of online news content is possibly strongly impacted by having influential tweeters, and less so by factors such as the "first mover advantage" and the overall news site popularity. More detailed analysis requires a larger data set.

- In previous work, Borghol et. al. [6] collect a dataset of YouTube video clones and find several factors which are responsible for varying popularity of online video clones. YouTube video popularity can be long lasting [19] whereas the popularity of news articles in the Twitterverse is short lasting [3]. By analyzing the factors impacting the popularity of online news articles it is now possible to compare the factors which are responsible for differences in clone popularity for these two very different types of content.

6.3 Future Work

More could be learned about online news popularity in the Twitterverse by some additional research work. Some possible directions of future work are as follows.

- In this thesis, some factors were investigated with respect to their possible impact on the total number of tweets and total number of unique tweeter tweets. Each factor was investigated separately. Investigating the correlations among these factors by studying the correlation matrix and using regression analysis could be a fruitful direction for future work.

- News stories could be divided by category, and the methodology used in this thesis could be applied to see whether different factors play major roles in promoting unlike kinds of news stories or not. For example, news stories could be divided into political, economic, natural disaster, science, and sports categories. The methodology of this thesis could be applied to investigate whether the popularity impacts of the various factors remains equivalent across the different categories.

- The methodology of this thesis could be applied for other types of content, or other social networking sites to investigate how content gets disseminated via their media compared to in Twitter.

- A characterization could be done based on the redundant tweets. It was found that the top followed tweeters were not the prominent redundant tweeters. Many ordinary tweeters make redundant tweets, which can have a huge impact in making some of the online content popular when considering the total number of tweets. Characterization of the redundant tweets of all of the clone set members could give a further view on who makes large numbers of redundant tweets and their impact.

References

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proc. IEEE/WIC/ACM'10 Vol. 01*, pages 492–499, Toronto, Canada, Aug. 2010.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proc. ACM WSDM'11*, pages 65–74, Hong Kong, Feb. 2011.

[3] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. In *Proc. ICWSM'12*, Ireland, Dublin, June 2012.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[5] M. Boanjak, E. Oliveira, J. Martins, E. Mendes, and L. Sarmento. Twitterecho: a distributed focused crawler to support open research with Twitter data. In *Proc. WWW'12*, pages 1233–1240, Lyon, France, Apr. 2012.

[6] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. The untold story of the clones: content-agnostic factors that impact Youtube video popularity. In *Proc. ACM SIGKDD'12*, pages 1186–1194, Beijing, China, Aug. 2012.

[7] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037–1055, Nov. 2011.

[8] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *Proc. INFOCOM'99*, pages 126–134, New York, NY, USA, Mar. 1999.

[9] M. Cha, H. Haddadi, F. Benevenuto, and P K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. *ICWSM*, 10:10–17, May 2010.

[10] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *Proc. WWW'09*, pages 721–730, Madrid, Spain, Apr. 2009.

[11] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, Feb. 2011.

[12] D. Fleder and K. Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712, Mar. 2009.

[13] D. M. Fleder and K. Hosanagar. Recommender systems and their impact on sales diversity. In *Proc. ACM Electronic Commerce '07*, pages 192–199, San Diego, California, Jun. 2007.

[14] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proc. WOSN'10*, Boston, MA, Jun. 2010.

[15] M. Gilad and G. Natalie. Leave a reply: An analysis of weblog comments. In *Proc. Third annual workshop on the Weblogging ecosystem*, May 2006.

[16] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW'04*, pages 491–501, New York, NY, May 2004.

[17] C. Hui, Y. Tyshchuk, W. A. Wallace, M. Magdon-Ismail, and M. Goldberg. Information cascades in social media in response to a crisis: a preliminary model and a case study. In *Proc. WWW'12*, pages 653–656, Lyon, France, Apr. 2012.

[18] S. Ihm and V. S. Pai. Towards understanding modern web traffic. In *Proc. ACM SIGCOMM'11*, pages 295–312, Toronto, ON, Canada, Aug. 2011.

[19] M. A. Islam, D. L. Eager, N. Carlsson, and A. Mahant. Revisiting popularity characterization and modeling of user-generated videos. In *Proc. MASCOTS'13*, San Francisco, CA, Aug. 2013.

[20] S. Kim, S.-H. Kim, and H.-G. Cho. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *Proc. IEEE CIT'11*, pages 449–454, Aug. 2011.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. WWW'10*, pages 591–600, Raleigh, NC, Apr. 2010.

[22] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert. The influentials: New approaches for analyzing influence on twitter. *Web Ecology Project*, 29, Sep. 2009.

[23] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Proc. IEEE/WIC/ACM WI-IAT'10*, pages 623–630, Toronto, Canada, Aug. 2010.

[24] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *ICWSM*, 10:90–97, May 2010.

[25] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. WWW'10*, pages 621–630, Raleigh, NC, Apr. 2010.

[26] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti. Characterizing the file hosting ecosystem: A view from the edge. *Performance Evaluation*, 68(11):1085–1102, Nov. 2011.

[27] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message Q & A behavior. In *Proc. CHI'10*, pages 1739–1748, Atlanta, GA, Apr. 2010.

[28] M. R. Morris, J. Teevan, and K. Panovich. A comparison of information seeking using search engines and social networks. *ICWSM*, 10:23–26, May 2010.

[29] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep. 2003.

[30] V. N. Padmanabhan and L. Qiu. The content and access dynamics of a busy web site: findings and implications. In *Proc. SIGCOM'00*, pages 111–123, Stockholm, Sweden, Aug. 2000.

[31] S. A. Paul, L. Hong, and E. H. Chi. Is twitter a good place for asking questions? a characterization study. In *Proc ICWSM '11*, Barcelona, Spain, July 2011.

[32] R.D.W. Perera, S. Anand, K.P. Subbalakshmi, and R. Chandramouli. Twitter analytics: Architecture, tools and analysis. In *Proc. MILCOM'10*, pages 2186–2191, San Jose, CA, Nov. 2010.

[33] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes. Do all birds tweet the same?: Characterizing Twitter around the world. In *Proc. ACM CIKM'11*, pages 1025–1030, Glasgow, Scotland, UK, Oct. 2011.

[34] G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. Analyzing patterns of information cascades based on users' influence and posting behaviors. In *Proc. TempWeb'12*, pages 1–8, Lyon, France, Apr. 2012.

[35] G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. Characterizing topic-specific hashtag cascade in Twitter based on distributions of user influence. In *Proc. Web Technologies and Applications*, pages 735–742. Kunming, China, Apr. 2012.

[36] G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. A study on characteristics of topic-specific information cascade in Twitter. *IEICE Technical Report*, 111(361):65–70, Dec. 2011.

[37] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *Proc. IMC'11*, pages 381–396, Berlin, Germany, Nov. 2011.

[38] M. Russell. *21 Recipes for Mining Twitter*. O'Reilly Media, 2011.

[39] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. WWW '10*, pages 851–860, Raleigh, NC, Apr. 2010.

[40] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, Feb. 2006.

[41] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proc. SIGSPATIAL GIS'09*, pages 42–51, Seattle, WA, Nov. 2009.

[42] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, Aug. 2010.

[43] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida. Predicting the popularity of online articles based on user comments. In *Proc. WIMS'11*, pages 1–8, Sogndal, Norway, May 2011.

[44] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, Dec. 2007.

[45] R. Zhou, S. Khemmarat, and L. Gao. The impact of YouTube recommendation system on video views. In *Proc. ACM SIGCOMM'10*, pages 404–410, Melbourne, Australia, Nov. 2010.

[46] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *National Academy of Sciences*, 107(10):4511–4515, Feb. 2010.

Appendix A

Appendix

```
{
   "created_at": "Wed Mar 06 13:01:23 +0000 2013",
   "id": 309287594938748929,
   "id_str": "309287594938748929",
   "text": "From a global politics perspective, he will not be missed by many Americans -  http:\/\/t.co\/pH2JWUFspK",
   "source": "web",
   "truncated": false,
   "in_reply_to_status_id": null,
   "in_reply_to_status_id_str": null,
   "in_reply_to_user_id": null,
   "in_reply_to_user_id_str": null,
   "in_reply_to_screen_name": null,
   "user": {
                "id": 471160540,
                "id_str": "471160540",
                "name": "Tornado Spectral ",
                "screen_name": "tornadospectral",
                "location": "Toronto \/ Ithaca",
                "url": "http:\/\/www.tornado-spectral.com",
                "description": "The creators of HyperFlux U1 spectrometers and OCTANE - nanophotonics-based
                   spectrometers for optical coherence tomography (OCT) and NDT",
                "protected": false,
                "followers_count": 89,
                "friends_count": 207,
                "listed_count": 1,
                "created_at": "Sun Jan 22 15:26:38 +0000 2012",
                "favourites_count": 0,
                "utc_offset": -18000,
                "time_zone": "Eastern Time (US & Canada)",
                "geo_enabled": false,
                "verified": false,
                "statuses_count": 408,
                "lang": "en",
                "contributors_enabled": false,
                "is_translator": false,
                "profile_background_color": "E68A19",
                "profile_background_image_url":
                   "http:\/\/a0.twimg.com\/profile_background_images\/632529914\/j6z06maj3gcd0umbcy18.jpeg",
```

```
        "profile_background_image_url_https":
              "https:\/\/si0.twimg.com\/profile_background_images\/632529914\/j6z06maj3gcd0umbcy18.jpeg",
            "profile_background_tile": false,
            "profile_image_url":
              "http:\/\/a0.twimg.com\/profile_images\/3256530660\/3209df455eb9ea7f5f878b060b5c8769_normal.
              png",
            "profile_image_url_https":
              "https:\/\/si0.twimg.com\/profile_images\/3256530660\/3209df455eb9ea7f5f878b060b5c8769_norma
              l.png",
           "profile_banner_url": "https:\/\/si0.twimg.com\/profile_banners\/471160540\/1360936798",
            "profile_link_color": "303F45",
            "profile_sidebar_border_color": "C6E2EE",
            "profile_sidebar_fill_color": "DAECF4",
            "profile_text_color": "663B12",
            "profile_use_background_image": true,
            "default_profile": false,
            "default_profile_image": false,
            "following": null,
            "follow_request_sent": null,
            "notifications": null
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "retweet_count": 0,
  "entities": {
    "hashtags": [],
    "urls": [{
        "url": "http:\/\/t.co\/pH2JWUFspK",
        "expanded_url": "http:\/\/www.bbc.co.uk\/news\/world-latin-america-21679053",
        "display_url": "bbc.co.uk\/news\/world-lat\u2026",
        "indices": [79, 101]
      }
    ],
    "user_mentions": []
  },
  "favorited": false,
  "retweeted": false,
  "possibly_sensitive": false,
  "filter_level": "medium"
}
```


Figure A.1: Sample Output from the Twitter Streaming API

```
{
  "contributors_enabled": false,
  "created_at": "Sun Apr 22 14:42:37 +0000 2007",
  "default_profile": false,
  "default_profile_image": false,
  "description": "Breaking news alerts and updates from the BBC. For news, features, analysis follow @BBCWorld (our World edition) and @BBCNews (our UK edition).",
  "favourites_count": 0,
  "follow_request_sent": null,
  "followers_count": 5536331,
  "following": null,
  "friends_count": 2,
  "geo_enabled": false,
  "id": 5402612,
  "id_str": "5402612",
  "is_translator": false,
  "lang": "en",
  "listed_count": 72906,
  "location": "London, UK",
  "name": "BBC Breaking News",
  "notifications": null,
  "profile_background_color": "FFFFFF",
  "profile_background_image_url": "http://a0.twimg.com/profile_background_images/571084338/bgjygltd2afkpc1sdnjd.jpeg",
  "profile_background_image_url_https": "https://si0.twimg.com/profile_background_images/571084338/bgjygltd2afkpc1sdnjd.jpeg",
  "profile_background_tile": false,
  "profile_banner_url": "https://si0.twimg.com/profile_banners/5402612/1357232722",
  "profile_image_url": "http://a0.twimg.com/profile_images/2186829506/128x128_twitter_bbc_breaking_normal.jpg",
  "profile_image_url_https": "https://si0.twimg.com/profile_images/2186829506/128x128_twitter_bbc_breaking_normal.jpg",
  "profile_link_color": "1F527B",
  "profile_sidebar_border_color": "CCCCCC",
  "profile_sidebar_fill_color": "FFFFFF",
  "profile_text_color": "5A5A5A",
  "profile_use_background_image": true,
  "protected": false,
  "screen_name": "BBCBreaking",
  "status": {
    "contributors": null,
    "coordinates": null,
    "created_at": "Tue Apr 09 12:34:25 +0000 2013",
    "entities": {
      "hashtags": [
        {
          "indices": [
```

```
            82,
            90
          ],
          "text": "Bushehr"
        }
      ],
      "urls": [
        {
          "display_url": "bbc.in/ZKY7Du",
          "expanded_url": "http://bbc.in/ZKY7Du",
          "indices": [
            91,
            113
          ],
          "url": "http://t.co/ku4HjlfHrs"
        }
      ],
      "user_mentions": []
    },
    "favorite_count": 30,
    "favorited": false,
    "geo": null,
    "id": 321601995754573824,
    "id_str": "321601995754573824",
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "lang": "en",
    "place": null,
    "possibly_sensitive": false,
    "retweet_count": 490,
    "retweeted": false,
    "source": "<a href=\"http://www.bbc.co.uk/news/\" rel=\"nofollow\">BBC News</a>",
    "text": "Update: Earthquake epicentre is 89km (55 miles) from Iran's nuclear power station #Bushehr
http://t.co/ku4HjlfHrs",
    "truncated": false
  },
  "statuses_count": 11768,
  "time_zone": "London",
  "url": "http://www.bbc.co.uk/news",
  "utc_offset": 0,
  "verified": true
}
```


Figure A.2: Sample Output from the Twitter Get User/Show API