2017

Social Networks Influence Analysis

Doaa Gamal

SOCIAL NETWORKS INFLUENCE ANALYSIS

by

Doaa H. Gamal

A thesis submitted to the
School of Computing
in partial fulfillment of the requirements for the degree of

Master of Science in Computing and Information Sciences

UNIVERSITY OF NORTH FLORIDA
SCHOOL OF COMPUTING

Spring, 2017

This thesis titled “Social Networks Influence Analysis” submitted by Doaa H. Gamal in partial fulfillment of the requirements for the degree of Master of Science in Computing and Information Sciences has been

Approved by the thesis committee: Date

_____ _____

Dr. Karthikeyan Umapathy
Thesis Advisor and Committee Chairperson

_____ _____

Dr. Lakshmi Goel

_____ _____

Dr. Sandeep Reddivari

Accepted for the School of Computing:

_____ _____

Dr. Sherif A. Elfayoumy
Director of the School

Accepted for the College of Computing, Engineering, and Construction:

_____ _____

Dr. Mark Tumeo
Dean of the College

Accepted for the University:

_____ _____

Dr. John Kantner
Dean of the Graduate School

ACKNOWLEDGEMENT

I would first like to thank my thesis advisor, Dr. Karthikeyan Umapathy, for his invaluable guidance and expert insights throughout this research. I would like also to thank Dr. Lakshmi Goel and Dr. Sandeep Reddivari for their insightful feedback and thorough review of this thesis. Last, I would like to thank Mr. James Littleton for his thorough review and very helpful suggestions to improve this document.

CONTENTS

FIGURES

EQUATIONS

TABLES

ABSTRACT

Pew Research Center estimates that as of 2014, 74% of the Internet Users used social media, i.e., more than 2.4 billion users. With the growing popularity of social media where Internet users exchange their opinions on many things including their daily life encounters, it is not surprising that many organizations are interested in learning what users say about their products and services. To be able to play a proactive role in steering what user's say, many organizations have engaged in efforts aiming at identifying efficient ways of marketing certain products and services, and making sure user reviews are somewhat favorable. Favorable reviews are typically achieved through identifying users on social networks who have a strong influence power over a large number of other users, i.e. influential users.

Twitter has emerged as one of the prominent social network services with 320 million monthly active users worldwide. Based on the literature, influential Twitter users have been typically analyzed using the following three models: topic-based model, topology-based model, and user characteristics-based model. The topology-based model is criticized for being static, i.e., it does not adapt to the social network changes such as user's new posts, or new relationships. The user characteristics-based model was presented as an alternative approach; however, it was criticized for discounting the impact of interactions between users, and users' interests. Lastly, the topic-based model, while sensitive to users' interests, typically suffers from ignoring the inclusion of inter-user interactions.

This thesis research introduces a dynamic, comprehensive and topic-sensitive approach for identifying social network influencers leveraging the strengths of the aforementioned models. Three separate experiments were conducted to evaluate the new approach using the information diffusion measure. In these experiments, software was developed to capture users' tweets pertinent to a topic over a period of time, and store the tweet's metadata in a relational database. A graph representing users was extracted from the database. The new approach was applied to the users' graph to compute an influence score for each user.

Results show that the new composite influence score is more accurate in comprehensively identifying true influential users, when compared to scores calculated using the characteristics-based, topic-based, and topology-based models. Also, this research shows that the new approach could leverage a variety of machine learning algorithms to accurately identify influencers.

Last, while the focus of this research was on Twitter, our approach may be applicable to other social networks and micro-blogging services.

Chapter 1

INTRODUCTION

For more than a decade, the Internet has been a major platform for conducting all aspects of business. International Data Corporation (IDC) estimates that by 2020, business transactions on the Internet will reach US $450 billion per day (Gantz & Reinsel, 2010). The Internet has become not only a marketplace where items and services are offered, but also a medium where opinions are made. Internet Live Stats (ILS) estimates that there is currently more than 3.3 billion Internet users (about 40% of the world population). Pew Research Center estimates that as of 2014, 74% of the Internet Users used social media, i.e., more than 2.4 billion users. With the growing popularity of social media where Internet users exchange their opinions on many things, including their daily life encounters, it is not surprising that many organizations are interested in learning what users say about their products and services (Internet Live Stats, 2016).

The increased amount of data created every moment through the different social media services such as Facebook, Twitter, Flickr, MySpace, and LinkedIn makes researchers eager to mine these huge datasets to glean insights and create value for users and organizations. Several of these social media services have developed tools to enable the extraction of real time data, which made this type of research on real data possible. One of the research objectives that have been sought after by many researchers is identifying influencer users on these social networks.

This thesis research focuses on analyzing the influence of users in social networks from the perspective of particular topics. The importance of analyzing user's influence and behavior on social networks arises from the huge increase of social media applications, number of users, average time spent per user, and the amount of data created and exchanged. Individuals and organizations that interact through these services create and share data with other users, typically within their groups. Pang and Lee (Pang & Lee, 2008) assert that major companies are increasingly realizing that consumer opinions expressed on social networks can wield enormous influence in shaping the opinions of other consumers. This assertion confirms the results of a previous study (comScore, 2007) that concluded that many readers of online reviews are influenced by those reviews in making their purchasing decisions. Pang and Lee (Pang & Lee, 2008) suggest that due to the economic impact of online influence companies should expend on online reputation monitoring and management.

The "Connected Consumers Are Not Created Equal: A Global Perspective" report by the A.T. Kearney management consulting firm, researches individuals' online behavior around the world (comScore, 2007). In that report, of the 10,000 "connected consumers", people who say they connect to the Internet at least once a week, as much as 28% said they are "continuously connected", and 23% use the Internet every hour. The survey shows also that about two thirds of the respondents in the age range 16-45 base their buying decisions, to varying degrees, on what they read on social media, as shown in Figure 1.

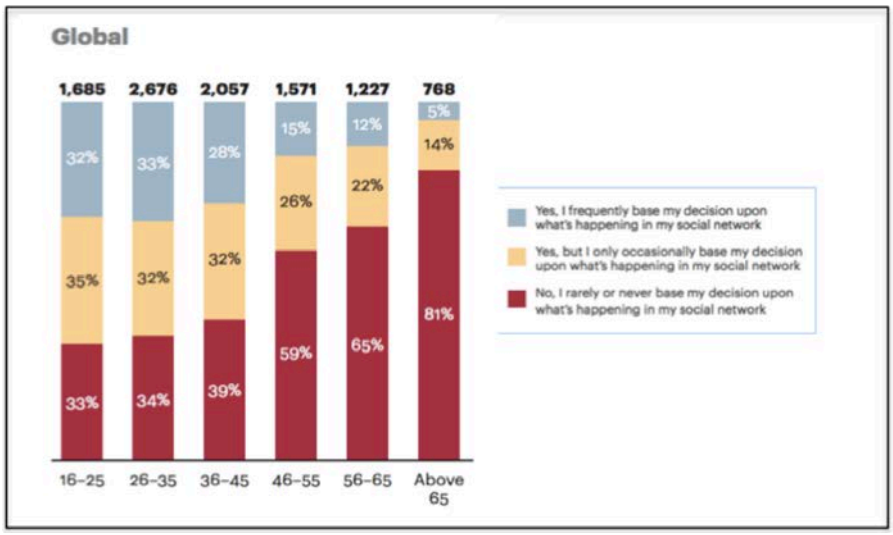

Figure 1. Influence by Social Media (comScore, 2007)

Analyzing users who have a strong ability to influence the opinion of others has developed into an interesting problem that could benefit many parties. For example, some organizations adopt this kind of market research to understand consumer behavior which could help them bias public's opinion in their favor. This emerging need led to the development of new companies that perform this type of market research using social networks. One of the earlier companies in that space is Klout (Klout, 2016).

Klout measures user's influence by using data from Twitter such as following counts, follower counts, retweets, list memberships, number of dead following accounts, the influence of people who retweet user's tweets, and unique mentions. Additionally, Klout links these data with data from a variety of social networks and other sources such as Wikipedia to compute a Klout score for every user (Parr, 2010). The Klout score is an integer in the range 0 to 100, where 0 is no influence whatsoever and 100 is most influential.

A good example of applying Klout score in market research is a project performed by Klout for American Airline. The giant airline was interested in driving its Admirals club membership. The company offered influencers, those with a Klout score of 55 or higher, with an Admirals Club One-Day Pass to access the lounges and enjoy their amenities, including: WIFI, snacks, beverages, wine, and mobile device charging stations. Those influencers were encouraged to share their experience using twitter's hash-tag #AdmiralsClub to create positive buzz (Klout, 2015a).

Another example is Motorola's market research, which aimed at increasing consideration for its Bluetooth S-11 Flex HD Headset (Klout, 2015b). The headset targets active, on the go consumers who want good quality, cordless headset to listen to music while at the gym or on the run. Klout identified influencers in certain verticals such as music, technology and sports, particularly those who were already passionate about the new headset to create a trustworthy and engaging content about the product. Motorola's product received over 62 million media impressions during the campaign. Additionally, it was found that customers who were driven to the product through Klout's influencers spent 2.5 times longer on the site compared to standard visitors. According to Nielsen, 90% of people trust a peer recommendation over an advertisement, and it seems that the authentic reviews produced by the influencers/evangelists activated through that campaign have successfully driven the consideration that Motorola was seeking (Klout, 2015b).

Clearly, there is a growing interest in performing this type of market research, as exemplified by Microsoft's announcement of its strategic investment in Klout whereby Bing (Microsoft's search service) would have access to Klout influence technology, and Klout would have access to Bing search data for its scoring algorithm (Grove, 2012). As of 2015, Klout is contracted by more than 3,500 companies to perform market research.

1.1 Problem Statement

The widespread adoption of social media as means for expressing, sharing, or even challenging opinions has increasingly attracted the interest of organizations that provide, or sell, products and services. Social media provides a platform for peer recommendations to play a greater role in adoption and purchase decisions (Wong, 2014). Many of these entities seek to identify the users who have the strongest influence on other users (i.e. influential), specifically with respect to these entities' services and products. Identifying influential users, i.e. those who have great abilities of influencing the opinion of other social media users, could help organizations effort in enhancing the reputation of their products, services, and public image in general. Influence is a crucial concept in sociology and viral marketing (Cha, Haddadi, Benevenuto, & Gummadi, 2010). Pedro Domingos asserts that the existence of network effects is acknowledged in marketing literature (Domingos, 2005). Domingos developed a social network model to identify the optimal set of customers to market to such that marketing to that set will yield the highest return on investment.

This research focuses on developing a composite score for measuring users influence on social media, particularly Twitter. The model carefully includes the aspects of topicality, reachability, and network dynamism to provide a more comprehensive and true measure of user's influence.

Chapter 2

BACKGROUND AND LITERATURE REVIEW

Social media is different than any other types of media because users have the ability to spread news, reviews and opinions, interact with others, and potentially influence other user's views. Literature shows that modeling how users are connected, which is most commonly done through social relationships, and how information flows from one user to others is an important step in studying the information reach and user's influence (Peng, Zhu, Piao, Yan, & Zhang, 2011). Social graphs enable the calculation of a variety of factors that could be helpful in calculating users influence. Examples of such factors include:

- Reach: This is the number of people a user influences. These are the users exposed to user's published contents.
- Amplification: Amplification indicates the effect on audiences a user has. That is how much a user influences other users within his/her reach by having them act on his/her published contents.
- Impact: This is the influence of user's audience. It indicates the influence level of people who engage with user's contents. It's not just about how many users in one's reach; it's about getting one's contents to the right people, those who are capable of further spreading the contents. Having more connections will not effectively increase the impact of a user, but having influential connections will.

2.1 Twitter

One of the most important applications on Social media is Twitter (Twitter, 2016). It is a social micro blogging service in real time that allows users to post messages that contain 140 characters or less. Twitter allows users to establish friends and followers. Friends of a user are the users that the user follows, and followers of a user are the users that follow that user. The importance of Twitter comes as a consequence of the huge numbers of users that use the service on regular basis, and also the ability to use web services to interact with Twitter. Demand for Twitter interactions has fueled the development of a plethora of third party applications and mobile Apps for variety of device platforms such as smartphones, tablets, and personal computers.

When a user posts a message, which is called a tweet, the user can make this tweet public or private. User's followers get to see the user's tweets, and can reply to them, or repost them, called retweets. Twitter is different than any other social media service in the relationship of following; it is not necessary that the user has a friendship relationship with his/her followers. In other words, Twitter relationships are not bidirectional, so if user A follows user B, user B may not follow user A. Additionally, users A and B may not have any type of social relationship outside Twitter, yet, user A is interested in receiving the tweets of user B.

Researchers studying social networks tend to prefer using Twitter due to the great accessibility allowed by Twitter for retrieving data about users, posts, relationships,

metadata about users such as their profile information, and metadata about tweets such as geo location and timestamp. In this thesis, we developed an application to retrieve, manage, and process data from Twitter. This data is used for analyzing users influence within the Twitter. Although the focus is on Twitter, which represents a special case of social networks where relationships are unidirectional as opposed to bidirectional relationships, which are more common across other social networks, this should not constitute a limitation on the generality and universality of the approach, albeit tweaks may be necessary.

Most approaches aimed at studying social network influence use social graph representations, but in Twitter's case, we strictly use a "directed" graph due to the unidirectional relationships. In other words, Twitter's relationships are directed, so following a user doesn't necessarily mean that user follows you back. That is why the relationship between two users is best modeled by a directed link (Lee, Kwak, Park, & Moon, 2010).

2.2 Social Graphs

Many researchers modeled social network users and their relationships in the form of a directed graph, typically referred to as social graph (L. Tang & Liu, 2010). The social graph represents all, or some of, the social network users and their social relationships. This thesis studies the structure and characteristic of the social graph created by the data

retrieved from Twitter for the purpose of analyzing user's influence. Social network analysis usually involves the following approaches:

- Centrality Analysis: This analysis focuses on identifying the most important actors in a social network (L. Tang & Liu, 2010), thus very relevant to the thesis's topic of analyzing users influence. In any community, there can be many participants at a given period of time, some of them are more central than others, and we may refer to them as influencers or leaders. Influential users are more likely to acquire connections, and it is unlikely that someone could be influential on a social network without having followers, upon whom the user expresses his/her influence. Domingos, however, gave the example of advisors to celebrities, where the advisors themselves do not necessarily have a large number of followers, yet the celebrities whom they advise have a large number of followers, which makes the advisors influential, but through their advisees (Domingos, 2005). Different algorithms have been developed to identify influential users and also rank members of a social network based on their influence power such as NodeXL (NodeXL, 2016), Mathematica (Mathematica, 2016), and eigenvector centrality (L. Tang & Liu, 2010).
- Community Detection: Communities of interest may be created spontaneously, when users finds a topic or a discussion interesting. Some communities last for long periods and some others fade away in short periods. Community detection is an important task in social network analysis because communities represent real social groupings, maybe by users' interest or background. Studying communities can be

helpful in understanding user groupings and interests. Researchers have studied different methods for detecting community structures (L. Tang & Liu, 2010). When a group of users interact with one another they form a strong community effect. In order to measure the community effect, researchers have used the concept of transitivity, which simply means that friends of friend are likely to be friends as well. They used clustering coefficient to measure the probability of connections between the friends of a user (L. Tang & Liu, 2010).

Other factors that are considered in modeling social networks for influence analysis were identified in (Newman, 2003; J. Tang, Sun, Wang, & Yang, 2009) which include:

- Diameter: the length of the longest shortest path between any pair of nodes in the social graph.
- Node Degree: the number of edges incident to a node. However in directed graphs in-degree and out-degree are used instead, where in-degree is the number of head ends adjacent to a node, and out-degree is the number of tail ends adjacent to a node.
- Degree Distribution: the distribution of the number of nodes with each degree value. It was found that in large scale social networks, the degree distribution follows the power law, i.e. few nodes have high degrees (particularly in-degrees in directed graphs) and many others have few degrees.

- Clustering Coefficient: the degree to which nodes in a large social network tend to cluster together. Nodes in social networks tend to have a higher probability of forming clusters than of randomly created networks (D. J. Watts & Strogatz, 1998).
- Small World Effect: In large scale networks, any two nodes are not too far away, a phenomenon sometimes referred to as "six degrees of separation," where nodes are in most cases about 6 links, or fewer, away from one another (Leskovec & Horvitz, 2008).

Lie Thang et al. focused on measuring the strength of topic-level social influence quantitatively. They presented some questions such as: what are the representative nodes on a given topic? how to identify topic-level experts and their social influence to a particular node? and how to quickly connect to a particular node through strong social tie? (L. Tang & Liu, 2010). Some of the challenges identified by Lie Tang et al. while computing influence on social graphs include the following:

- Multi-aspects: The social influence is associated with different topics. One user can have high influence on another user on a particular topic, but may be the second user has a higher influence on the first user on another topic. Therefore, influence is topic (aspect) dependent.
- Node-Specificity: Social influence algorithms should not measure the global importance of nodes, but rather measure the importance of links between nodes, i.e. the directed influence between pairs of users.

- Scalability: Given the current size of social networks and their current of rate of growth, which is only expected to accelerate, it is essential for the social influence algorithms and techniques to scale well with such huge datasets.

2.3 Literature Review

The massive information generated on social networks everyday enticed researchers to study social influence because of the potential impacts on many businesses. Identifying users who can influence the opinions and decisions of other users, either positively or negatively, and quantifying user's influence have been the focus of many recent social network studies (Hill et al., 2011). This topic has been extensively studied for marketing, and to a lesser extent in other disciplines (Hill, Provost, & Volinsky, 2006). For example, in healthcare, studying how users may spread smoking behaviors through social networks is of interest to many entities (Christakis & Fowler, 2008).

Research efforts in social media used a wide variety of methods and algorithms to study user's influence and diffusion of information. But they can be categorized into three main approaches, namely, network topology, user characteristic, and topic sensitive. These approaches are not mutually exclusive, and in the reminder of this chapter will discuss the most notable efforts in each category.

2.3.1 Social Network Topology-based Approach

Research efforts that followed the network topology approach have primarily focused on leveraging the unidirectional follow relationship that exists between users. The findings of Bakshy et al. (Bakshy, Rosenn, Marlow, & Adamic, 2012) suggest that although weak ties can serve a critical bridging function, which was already discovered by (Granovetter, 1973) and (Onnela et al., 2007), the majority of influence results from exposure to individual weak ties, which indicates that most information propagation on social media is driven by simple contagion. This contrasts the conclusions of prior studies that suggested the densely connected users have higher influence (Aral & Walker, 2011; Backstrom, Huttenlocher, Kleinberg, & Lan, 2006; Centola & Macy, 2007; Centola, 2010). The study by Bakshy et al. focused on Facebook; hence their findings may not be applicable to all social networks, particularly those with unidirectional relationships, given Facebook's bidirectional relationships.

Bakshy et al. (Bakshy et al., 2012) examined the role of social networks in the propagation of information with a large-scale field experiment that randomized the exposure to information among 253 million Facebook users. Users who were exposed to information were significantly more likely to spread the information they receive, and do so sooner than those who were not exposed to such information. They also examined the relative role of strong and weak ties in information propagation, where they found that although stronger ties are individually more influential, the large number of weak ties is more responsible for the propagation of information. The authors claim that it is nearly impossible to determine

from observational data whether any particular interaction, mode of communication, or social environment is responsible for the propagation of information through social networks. They argue that weak ties have access to more diverse information because such users with weak ties are expected to have fewer mutual contacts.

Bakshy et al. (Bakshy, Hofman, Mason, & Watts, 2011) tracked the diffusion events that took place on the follower graph during a specific period. In order to do that, they combined two sources of data. The first data source was the public tweets broadcast that included bit.ly URLs. Second, they crawled the portion of the follower graph to get all users who had broadcast at least one URL over the same period of time. They also calculated the influence score for a given URL post and tracked the diffusion of the URL from its origin at a particular seed "node" through a group of reposts by the user's followers until the diffusion event terminated. They assumed that user A influenced user B, if user B post the same URL after user A did. If B has more than one friend who has previously posted the same URL, they determined three possible scenarios to assign the corresponding influence: first, by assigning full credit to the friend who posted first; second, by assigning full credit to the friend who posted it most recently; and third, by splitting the credit equally between all prior-posting friends.

Bakshy et al. (Bakshy et al., 2011) analyzed the attributes and influence of 1.6 million Twitter users and 74 million messages retweets over a two months period. Their findings suggest that the largest number of retweets tend to be generated by traditionally influential users who have a large number of followers. More interestingly, they considered marketing

strategies based on the relative cost of identifying potential influencers, versus compensating potential influencers. Their results contradicted the common belief of recognizing prominent users as most effective for information diffusion (Leavitt, Burchard, Fisher, & Gilbert, 2009) and showed that although in certain situations the most influential users are the most cost effective; it is more likely that the most cost-effective performance can be attained by leveraging average or even less-than-average influencers. Again, their focus was on cost effectiveness, and for it to have a utility an influence score has to be calculated.

Lee et al. proposed a method to find influential users by considering both the link structure and the temporal order of information adoption on Twitter (Lee et al., 2010). Their method emphasizes the importance of timeliness of information adoption; assuming that a user reads all tweets he/she receives in chronological order. Their method finds influential users based on the number of effective followers (readers) a user has, where a follower user can belong to one of two categories with respect to a new message: the user has already read the message or yet to read it. From their experiment, they found out that most of the influential users using their method were news media, which led them to claim that news media has significant influence in spreading information to effective users.

TunkRank (Tunkelang, 2009) is a measure of Twitter user's influence. It is developed on the foundations of Google's PageRank algorithm, but for Twitter. TunkRank uses two basic ideas for its influence metric. First, the amount of attention a user can give is spread out among all those followed by that user. The more users a user follows, the less attention

each followee gets. Second, user's influence depends on the amount of attention the user receives from his/her followers. The TunkRank score takes also into account the amount of attention generated by user's followers both directly and indirectly through their network of followers. TunkRank uses a 1–100 metric, where 100 is most influential and 1 is least influential. The algorithm behind the TunkRank is based on Daniel Tunkelang's Twitter influence algorithm (Tunkelang, 2009). The influence of a user, X, is calculated as follows:

$$Influence(X) = \sum_{Y \in Followers(X)} \big(1 + \mathrm{p} * \mathrm{Influence(Y)}\big)/||Following(Y)||$$

Equation 1. TunkRank Influence

Where*:*

- *Influence(X)* is the number of users who are expected to read a tweet that user *X* tweets, including all retweets of that tweet. For simplicity, it is assumed that if a person reads the same message twice, both readings count.
- If *X* is a member of *Followers(Y)*, then there is a *1/||Following(X)||* probability that *X* will read a tweet posted by *Y*, where *Following(X)* is the set of people that *X* follows.
- If *X* reads a tweet from *Y*, there's a constant probability, *p,* that *X* will retweet Y's tweet.

This model in particular, combined with the above assumptions, accounts for the inflation that occurs from people who follow in the hopes of reciprocity. There's less value in being

followed by someone who follows a lot of people, because that person is less likely to read their messages or retweet them.

The approach of developing influence models based on the social network topology is generally static and does not consider the activities or interests of individual users. It is important to note that the following relationship could indicate intimate friendship, common interests, and anything in between.

2.3.2 User Characteristics-based Model

Agarwal et al. (Agarwal, Liu, Tang, & Yu, 2008) asserted that since blogging has become a popular way for users to publish information on the Web, bloggers tend to share their sentiment, express their opinions about products and services, or provide recommendations and reviews. Bloggers tend to also communicate within groups. These groups are usually formed around particular interests where relevant information is disseminated. However not all bloggers are powerful in biasing the opinion of members of their groups, there are still influential bloggers, but these are not necessarily the most active ones. Agarwal et al. developed a model to identify influential bloggers by conducting experiments involving the whole history of blog posts of a blog site. Their experiments showed that their model is capable of identifying influential bloggers, who are not necessarily the most active bloggers. In other words, the authors proved that posting a large number of blog articles does not necessarily earn a blogger an influential status. The same may be true in the micro-blogging sphere as well where the number of posts by itself doesn't make a blogger

influential, but rather the combination of a number of factors. Otherwise, it will be a race of who can post more. Often, influence is established by bloggers' reach, i.e., number of direct and indirect connections, and the impact of their posts (in terms of number of impressions, likes, re-posts, etc.).

Ye et al. (Ye & Wu, 2010) focused on characterizing information propagation and social influence on social networks. Their experiment involved the collection of 58 million Twitter messages by 700,000 users and the study of message flows to understand how breaking news were spread. Their results proved that messages quickly propagate far away from the original author's immediate followers. Their study focused on examining the stability, assessment and correlation of social influence over time.

Kwak et al. studied the topological characteristics of Twitter and its power as a new medium of information sharing (Kwak, Lee, Park, & Moon, 2010). In their follower-following topology analysis using 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets, they revealed a deviation from the known characteristics of human social networks such as a non-power-law follower distribution, a short effective diameter, and low reciprocity (Newman, 2003). They also measured the user's influence on Twitter using the number of followers, PageRank, and the number of retweets. Whilst the first two methods were found to be similar, the number of retweets method indicated a gap in influence inferred from the number of followers and that from the popularity of one's tweets. One of the interesting findings of Kwak et al. is that retweeted messages reached an average of 1,000 users regardless to the number of

followers of the original author. Upon the first retweet, a message is retweeted by second, third, and fourth level followers indicating a fast diffusion of information after the first retweet.

They also made a comparative analysis of three different measures of influence: namely followers, retweets, and user mentions which were used to evaluate the social influence. Their results show that the number of followers may not be the best measure of the influence. The findings of Kwak et al. with respect to the importance, or lack thereof, of the number of followers as a single measure for user's influence was also confirmed by the study conducting by Cha et al. (Cha et al., 2010).

Cha et al. (Cha et al., 2010) assert that social media connections could represent intimate friendships, common interests, passion for news, celebrity gossip, or a number of other reasons. The directed links resulting from these relationships indicate the flow of information and the influence of a user. In their study, Cha et al. used Twitter data to study the dynamics of user influence across topics and time. They also compared indegree (followers), retweets, and mentions as individual measures of influence. In addition to confirming the findings of Kwak et al. (Kwak et al., 2010) that indegree is not the most effective measure of user's influence, they found that influential users can hold significant influence over a variety of topics. However, their study also revealed user's influence is not gained spontaneously. Gaining and maintaining influence requires a great deal of deliberate activities such as increased level of personal involvement and limiting tweets to

few topics. This finding suggests that influential users are more predictable than suggested by theoretical models such as those developed by Watts (D. Watts, 2007).

Huang et al. (Huang, Liu, Chen, & Cheng, 2013) developed a dynamic algorithm based on the concepts of social diversity of the influenced users and dynamic influence propagation to identify the influencer users on Twitter using aggregate information. Their study suggests that due to the nature of rapid changes on social networks and to reflect the flow of influence spread, the patterns of influence propagation should be updated dynamically. The temporality of the influence relationship seems to be an intriguing aspect. In my analysis, data will be collected during a specific period of time and the influence will be studied over that period. There exists the potential for relationships to change over time, however this will be out of the scope of my study.

Huang et al.'s algorithm calculated social diversity using user interactions, which is changing dynamically, as compared to previous methods (Huang, Liu, Lin, & Cheng, 2013) that predominantly used community structure or static influence propagation. The biggest disadvantage of this user characteristics based approach for calculating user's influence is that it does not take into consideration the interaction between users and their interests, where both factors may be important.

2.3.3 Topic Sensitive Model

Weng et al. (Weng, Lim, Jiang, & He, 2010) analyzed a sample Twitter dataset and found that the majority of Twitter users follow their followers, indicating strong presence of reciprocity. They claim that this finding is explained by the phenomenon of *homophily*, which is the tendency of individuals to bond with individuals who share similar interests (McPherson, Smith-Lovin, & Cook, 2001). In Twitter, this means users who are interested in certain topics tend to follow other users who are interested in the same topics. Their study also included proposing a new influence measure called TwitterRank, measures the influence of Twitter users taking into account the topical similarity between users, the link structure, and the number of tweets by each user. The authors admit that TwitterRank can be skewed by if users deliberately publish a large number of tweets, and that it could be improved by incorporating other interactions between Twitter users. TwitterRank is considered an extension of Google's PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998). The experimental results showed that some Twitter users follow other users not necessarily because of their topical similarity (homophily phenomenon), which contradicts the findings of Cha et al. (Cha et al., 2010). The biggest drawback of this approach is its lack of consideration of user's activity. In other words, users who do share interests but do not actively engage in discussions, forward, or reply to other user's messages will likely be less influential, but this model does not factor their interaction into influence score calculations.

2.3.4 Summary

Given that all the surveyed efforts seem to focus on one approach or the other, and given the inherent drawbacks in the individual approaches, I believe a composite score that encompasses all three approaches (user characteristics, network topology, and topic sensitivity) will be more accurate in calculating user's influence. The proposed approach is explained in the following chapter.

Due to the lack of a standard dataset that is used by all the surveyed methods, this research collects three datasets from Twitter and implements the general theme of each of the three approaches (topology, user characteristics, and topic sensitive models). The topology based model focuses primarily on the number of followers and the number of friends followed by the user, while the user characteristics model focuses on the user's total number of tweets and the number of friends. It is obvious that both the topology and user characteristics models are not sensitive to a particular topic. Conversely, the topic sensitive model focuses primarily on the number of topic specific tweets, the number of replies, and the number of favorites of those tweets. It is important to note that my implementation of these approaches does not resemble any particular implementation of those surveyed, but rather it focuses on the main theme of the different approaches.

Chapter 3

COMPOSITE INFLUENCE SCORE

3.1 Social Influence Modeling Methodology

Prior research has focused on modeling social networks using topic-based models, topology-based models, or user characteristics-based models for identifying influencer users, as discussed in chapter 2. While the topology-based model introduced in (Aral & Walker, 2011; Backstrom et al., 2006; Bakshy et al., 2012; Centola & Macy, 2007; Centola, 2010; Granovetter, 1973; Leavitt et al., 2009; Lee et al., 2010; Onnela et al., 2007; Tunkelang, 2009) are criticized for being static, the user characteristics-based model introduced in (Agarwal et al., 2008; Cha et al., 2010; Huang et al., 2013; Huang et al., 2013; Kwak et al., 2010; Newman, 2003; D. Watts, 2007; Ye & Wu, 2010) do not account for interactions between users or user's interests. Additionally, the topic-based models introduced in (McPherson et al., 2001; Page et al., 1998; Weng et al., 2010), while sensitive to user's interests, they typically suffer from ignoring the inclusion of inter-user interactions. The limitations of the individual approaches, primarily due to their static nature, have led to several inaccuracies. First, a user with a large number of connections may not be an authentic source of information on every topic, therefore the influence of messages posted by that user will not always have the same influence on followers. Second, the followers of a user may not all be interested in the topic of every message posted by that user. A good example of this is the connections (follow relationship on Twitter) established between family members who may have completely different interests.

Due to the aforementioned limitations of the static approaches for identifying influencers, there is an obvious need for more dynamic approaches to identifying social network influencers. For example, Apple may be interested in identifying influential Twitter users who write about the company's new iWatch. Apple could send an iWatch for influential users to closely review the product, or invite them to special events where they can take a closer look at the watch and gain more knowledge of the product and its features. Ideally, these efforts would motivate those users to write more positively about the iWatch. Given their influence power, which is why they were selected in the first place, they could spread positive sentiment and hence enhance the product's reputation.

This thesis research introduces a dynamic topical approach for identifying social media influencers using a composite measure that includes the following factors:

- Number of re-tweets: This factor indicates the authoritativeness of the user
- Number of followers: This factor indicates the reachability of the user

In this approach a topical sub-network is constructed from all the users who post messages pertinent to a particular topic, *p*, within a particular time frame, *t*. The assumption is that if *t* is long enough, users interested in *p* would have posted at least one post. This approach will produce a graph representing the community of users interested in topic *p*.

It is not uncommon for social network users to echo the posts of other users. Every time a message is echoed by another user this is considered a vote for the trustworthiness of the message's original author. In Twitter, echoing a message is called re-tweeting, and

fortunately Twitter uses metadata to tag re-tweets and relate them to the original message. In this research, the cumulative number of all user's message retweets by the network's other users is used as a factor of that user's authoritativeness. For example, if user *u* posted two messages, *m1* and *m2*, on the same topic, *p*, where *m1* was retweeted *r1* times and *m2* was retweeted *r2* times by other users in that topical network, the influence of user *u*, *i*, will be in part determined by the total amount of retweets of his/her messages on that topic within the analysis time frame, *t*, which is quantified by *r1+r2*.

The second factor in the influence composite score is the total number of users who have received the original message or the re-postings of it. This is same as the total number of followers of the original message's author and the followers of each user who re-tweeted the original message. This factor is usually referred to as reachability (Hanneman & Riddle, 2005). We assume that not every user in social networks participate actively in posting messages even on topics they are interested in. In other words, some users prefer to be more passive and consumers of the information and content generated by other users who they may follow. For example, if user u who has 500 followers, posts a tweet that gets retweets by some other user who has 1000 followers, then u's message has reached a total of 1500 users, which constitutes more reachability, and hence influence (partially indirect though), than if it was retweeted by another user who has only 10 followers (limited reachability). Of course, one of the challenges that we will have to address is users overlap. If a particular user is following two other users and both retweet the same message, ideally that user should be counted once for the sake of quantifying the reachability. Although it will be very difficult to always get a complete list of followers and identify the intersection

of followers of different users for, primarily, the amount of processing required, there is a more subtle reason for not doing that. Since our primary goal is measuring user's influence, one can argue that if a user receives the same message more than once that reinforces the message and contributes to the original author's influence, hence, the reachability measure will count the number of message deliveries, every instance of it.

If n users exist in a particular topical network, we will calculate a composite influence score for each user in that network. Of course the higher the influence score the more influential the user is and vice versa. The following section provides a detailed description of how the composite score will be calculated.

3.2 Social Influence Modeling

Unlike traditional media types, users of social networks play an active role in ranking and re-broadcasting the information they receive. For example, newspapers influence could be measured by the number of subscriptions and the number of ads. Likewise, the influence of a television show could be determined by the number of viewers, or the number of listeners in case of radio shows. But on Twitter the situation is different where participants can "favorite" and/or "reply" a post, giving them a more active role. Therefore, user interactions with respect to other users' posts will be taken into consideration to compute the composite influence score. Additionally, the score is dynamic, which means user tweets, as well as the actions (and reactions) of other users, impact the user's influence score.

The Reachability of message *m* by user *u* will be measured in this model as the number of followers of the user. If user *u* posts *M* messages, then *u*'s reachability will be the sum over *u*'s messages.

$$Reach_m = Followers_{u,m}$$

Equation 2. User Reachability Factor

Where:

- $Followers_{u,m}$ is the number of followers of user *u* at the time of posting message *m*

The Impact generated by message *m* of user *u* will be measured as the sum of the number of "favorite", number of "retweet", and number of "reply" *m* receives. The relative contributions will be adjusted by weight factors *w1*, *w2* and *w3*.

$$Impact_m = w1 * Favorite_m + w2 * Retweet_{m +} w3 * Reply_m$$

Equation 3. Message Impact Factor

The Influence of user *u* is defined as the sum of the user's relevant messages Reachability and Impact.

$$Inf_u = Impact_m + w4 * \sum_{m \epsilon M} Reach_m$$

Equation 4. User Influence

Where:

- M is the set of user u's relevant messages
- $Reach_m$ is the reachability of message m
- $Impact_m$ is the impact of message m
- $w4$ is a weight factor representing the relative contribution of user's reachability

Twitter users have additional set of characteristics that may be useful in identifying the influential user. While performing the experiments, as will be discussed in Chapter 5, it was determined that user's influence might benefit from the inclusion of those characteristics. Those characteristics are identified as the total number of tweets made by a user (not necessarily in the topic of interest). This is an indication of user's overall activity. The second parameter is the number of user's friends, which is a topological factor. Last, although the number of messages posted by user u in a particular topic, M, is used for aggregating the reachability of user's individual messages, as shown in Equation 4, it was added as an independent parameter in computing user's influence. The hypothesis is that those additional parameters will be included in the influence model and experiments will demonstrate whether they have an impact. The additional parameters will update the influence of user u to the updated to the one below, Equation 5.

$$Inf_u = w1 * Favorite_m + w2 * Retweet_{m+} w3 * Reply_m + w4 * \sum_{m \epsilon M} Followers_m + w5 * Statuses_u + w6 * Friends_u + w7 * Tweets_u$$

Equation 5. Updated User Influence

3.3 Social Influence Implementation

3.3.1 Twitter API

Twitter uses OAuth 1.0A to provide authorized access to its Application Program Interface (API) via the v1.1 Authentication Model. The model provides two modes: application-user authentication and application-only authentication. This thesis uses the application-only authentication because the function of the application does not depend on the user of the application. The application-only authentication is a form of authentication where the application makes API requests on its own behalf, without a user context.

Twitter uses REST APIs to provide programming read and write access to its data. Twitter communicates data using to applications using JSON objects. But since the application monitors and process Tweets data in real-time, Twitter's streaming API is used within this thesis. The streaming API continuously delivers new responses to REST API queries over a long-lived HTTP connection and provides a low latency access to the global stream of Tweet data.

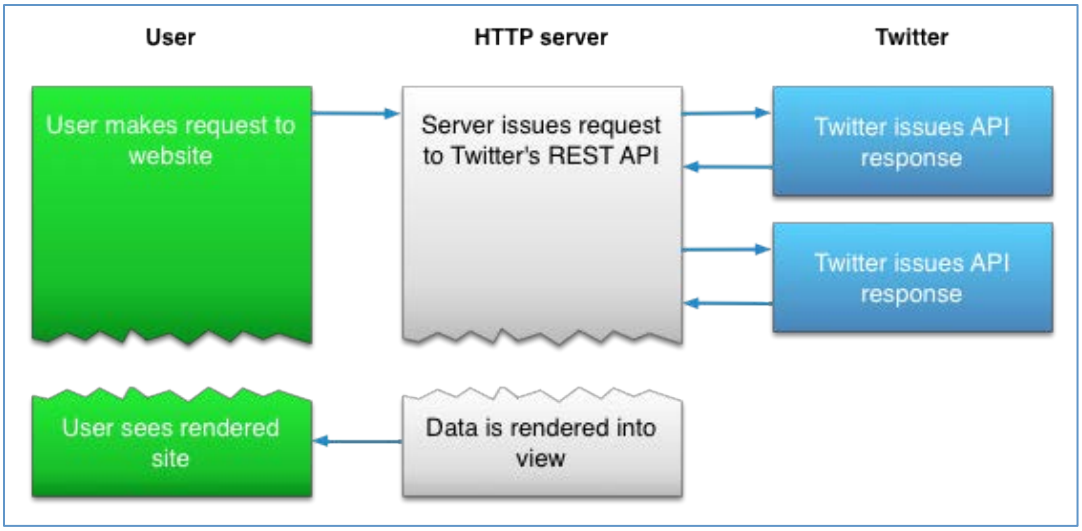

Figure 2. Twitter Connection Model

Twitter API provides four main objects: Tweets, Users, Entities, and Places. Tweets are the basic atomic building block of all things on Twitter. See below for a brief explanation of the main objects:

- Tweets: also known more generically as “status updates,” and they can be embedded, replied to, liked, unliked and deleted.
- Users: can be anyone or anything. Users can tweet, follow, create lists, have a timeline, be mentioned, and be looked up.
- Entities: provide metadata and additional contextual information about content posted on Twitter. Entities are never divorced from the content they describe, and are returned wherever Tweets are found in the API. Entities are essential for resolving URLs.
- Places: are specific, named locations with corresponding geo coordinates. They can be attached to Tweets by specifying a place_id when tweeting.

3.4 Evaluation Plan

Given that the objective of this research is to identify topic influencers in social networks, it makes sense to compare the effectiveness of the composite influence score to the previous methods using information diffusion. Information diffusion, or information spread, has been used in many research studies (Gomez-Rodriguez, Song, Du, Zha, & Scholkopf, 2016; Herzig, Mass, & Roitman, 2014; Kempe, Kleinberg, & Tardos, 2003). Information diffusion is a measure of the spread of contagions (tweets in this research) through actions such as sharing and forwarding (favorite, like and retweet in this research) enabled by social networks.

Finding the smallest set of users (source nodes) in Twitter that maximizes the spread of information in a limited amount of time depends dramatically on the dynamics of the underlying social graph. This has been proven to be an NP-hard problem by (Gomez-Rodriguez et al., 2016). This thesis rather than trying to find such an optimal set of influencers, it postulates that the composite score approach will produce a more accurate set than those produced by the surveyed approaches. The evaluation experiments of this research will use the following independent parameters:

Social Graph: This is the network of users who tweet and respond to tweets pertinent to a certain topic. In our experiments, we use three different topics from different domains. The topics are: the Jaguars NFL team, Hurricane Matthew, and Climate Change. Data collection lasted for several days for each of the datasets from Twitter live feeds, which

provides us with a large enough dataset. Of course the size of the data is a function of the topic and user's interest in the topic.

Influencers Set Size: This is the number of top influencers selected to measure information diffusion. However this could be a percentage of the size of the social graph or an absolute number, it was decided that the three experiments would use percentages.

Influence Score Model: The composite influence model developed by this thesis will be compared to the topic, user-characteristic, and topology based models.

For each combination of the 36 independent parameter values (3 topics x 4 influence scores x 3 influencers set sizes) we will follow a similar methodology to that presented in (Herzig et al., 2014) to compute information diffusion. Specifically, we will use the information diffusion model presented by Kempe et al. (Kempe et al., 2003), which provides a way to quantify the amount of information spread within a network.

We hypothesize the composite influence score will identify a larger number of the actual influencers, using information diffusion values, over the other methods for each of the three datasets.

3.4.1 Information Diffusion

The information diffusion was used in many research projects as a robust metric for assessing the spread of information. Since an influencer user is one who can reach a large number of users, such a user will have a large information diffusion value. In this thesis adopted the information diffusion measure as the baseline for comparing the different methods. In Chapter 4, the information diffusion measure is represented by the CummFollowers value. As discussed later, CummFollowers is the sum of the number of followers of the user and the followers of all users who retweeted that user's tweet.

In our experiments we collected data over limited periods of time (two weeks for each experiment topic groups), which allowed us to collect every tweet posted on a specific topic. Consequently, as shown in Figure 3, this provides for the opportunity to parse the social graph starting from the user's original tweet and going through every retweets and identifying the number of user followers. This parsing process has to be repeated for every tweet by the user in the specific topic. Although this method enables the computation of CummFollowers, which represents the information diffusion, it is impractical for organizations interested in studying user's influence to collect such large amounts of data and parse the resulting social graphs for each user and each relevant tweet on continuous basis. We postulate that the composite score will provide a simpler, accurate alternative. In other words, if the CummFollowers value can be predicted using the composite score presented here. Similar to typical machine learning projects, during the training phase a model is developed, and once proven accurate it can be used in production to classify or

predict some value with much less effort. Of course, such models may need to be reconstructed to correspond to emerging trends. In the context of identifying influencer users, the construction of the model, as proposed here, requires calculating CummFollowers, but once the model is created, it can be used to identify influencers. If the accuracy of the model drops, the model will need to be reconstructed.

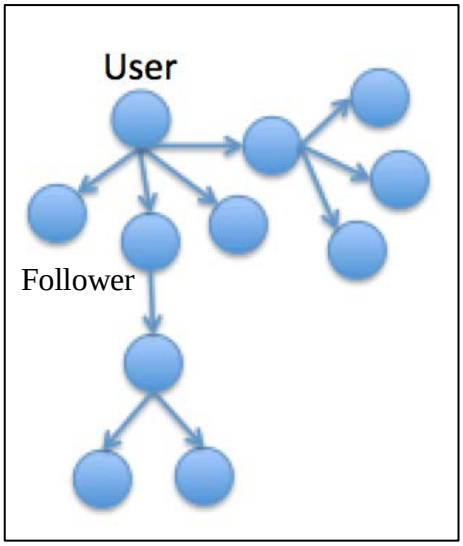

Figure 3. Social Graph Parsing

To put things in perspective, the Hurricane Matthew experiment, discussed in Chapter 5, has collected data from 1,026,769 unique users who posted 2,164,142 tweets that were retweeted 602,461 times, as noted in Table 1. The calculation of CummFollowers requires parsing such a very large social graph, where as if the influence score can be computed using the independent variables and still produce accurate identification of influencer users, that would be a much simpler process.

3.4.2 Predictive Models

The collected data in those experiments will be used to develop regression equations for each of the surveyed models well as the composite score models. The regression equations will identify the best value for the weight factors (*w1, w2, w3, w4, w5, w6* and *w7*) presented in Equation 5 above. The model will calculate a composite influence score that is closest to the value of user's CummFollowers. In other words, the regression model will try to calculate user's influence score that is as close as possible to the CummFollowers. This approach is typically referred to as curve fitting. Once the regression model is developed, and the weight factors are identified, the model can be used in real-time settings to identify set of influencer users by predicting their composite influence score, without having to compute CummFollowers.

Additionally, other types of supervised predictive models were examined to determine how accurate they could compute influence scores. The computed scores by those models will be compared to the CummFollowers to determine their accuracy. Again, if those predictive models produce high accuracy results that will suggest that they could be used in real-time sittings to predict user's composite influence score without having to compute CummFollowers.

Chapter 4

RESEARCH METHODOLOGY

Design science is a problem-solving paradigm. The earlier focus of design science in Information Systems (IS) was primarily on the impacts of IT artifacts on organizations, teams and individuals. More recently the focus has shifted to the development of IT artifacts in the context of solving real-world problems in a particular application domain (A. R. Hevner, March, Park, & Ram, 2004). The creation of such artifacts usually involves activities such as analysis, design, implementation, and use of information systems. Design science research in IS addresses most challenging problems. Those are typically characterized by: unstable requirements, complex interactions among the different subsystems and subcomponents, flexibility to changes in design processes and artifacts, high dependence on human input, and high dependence on human soft skill. Hevner has identified three design science research cycles that are typical in any design research project as shown in Figure 4.

The focus on design science research is on three processes. The Relevance Cycle relates the problem and application domain with the design science activities. Connecting the design science activities with the literature, scientific foundations, and expertise is the primary focus of the Rigor Cycle. Central to the design science process is the Design Cycle. It iterates between research processes and the development of artifacts and processes. A design science research project must have these three cycles.

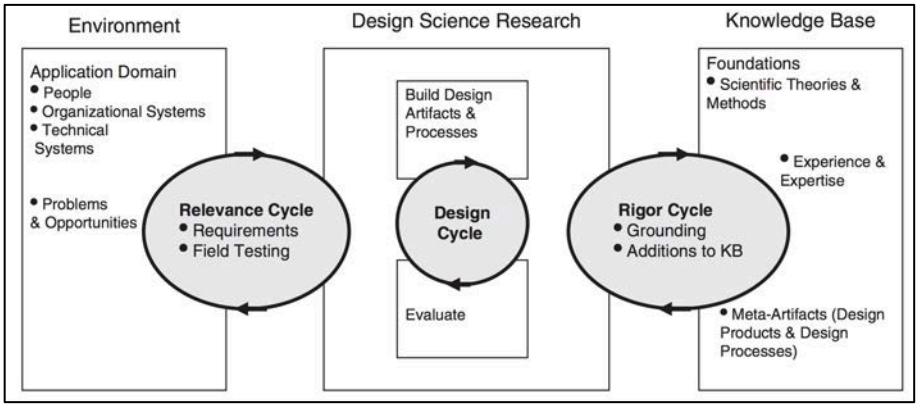

Figure 4. Design Science Research Cycles (A. Hevner, 2007; A. R. Hevner et al., 2004)

Henver et al. provide the following set of principles for conducting and evaluating good design science research in IS (A. R. Hevner et al., 2004).

Principal 1: Design as an Artifact

The first principal states that design science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation. The primary goal of this thesis is to produce a viable model for computing a composite score for identifying the influencers in social networks, as described in chapter 3. Therefore, this thesis satisfied the design as an artifact principal.

Principal 2: Problem Relevance

The second principal states that the objective of design science research is to develop technology-based solutions to important and relevant business problems. The problem of identifying influencers in social networks has been widely investigated by many researchers. Efficient solutions to this problem could impact many domains, e.g. marketing,

as outlined in chapters 1 and 2; therefore the problem relevance principal of design science research is met.

Principal 3: Design Evaluation

The third principal states that the utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods. The composite influence score was evaluated through a well-researched method, i.e. information diffusion that has been introduced and widely applied in the literature, as explained in chapter 3. Accordingly, this this satisfied the evaluation requirements of the design evaluation principal.

Principal 4: Research Contributions

The fourth principal states that effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies. This thesis produces a composite influence-scoring model that is founded on previously researched models. We argue that the composite score model proves to be superior to other models based on the information diffusion criteria. Therefore, the research contribution of this thesis meets the principal of the research contributions of design science research in IS.

Principal 5: Research Rigor

The fifth principal states that design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact. The composite

influence model presented in chapter 3 uses some of the well-researched principals, described in chapter 2, and was evaluated using rigorously vetted methodologies as explained in section 3.4. This thesis meets the research rigor principal required by the design science research guidelines.

Principal 6: Design as a Search Process

The sixth principal states that the search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment. The development of the composite influence score involves the exploration of various methods for combining the Impact and Reachability factors, as explained in chapter 3. This guarantees this thesis meets the design as a research process principal of design science research in IS.

Principal 7: Communication of Research

The seventh, and last, principal of design science states that the research must be presented effectively to both technology-oriented and management-oriented audiences. This principal is met through a written thesis and an oral public presentation. It may also result in publications in IS journals and conferences. Therefore, this research satisfies the required communication of research principal of design science research in IS.

Chapter 5

EXPERIMENTS

Three experiments were developed to evaluate and validate the new composite influence score. Three Twitter datasets were collected for the experiments. To do that, two C# programs were developed, Data Collection and Screen Scraping. Both programs are explained in the following two sections.

5.1 Data Collection Program

The Data Collection program was written in C# to interact with Twitter and retrieve tweets that contain any number of keywords (and phrases) provided by the user through Twitter's 1.1 REST API. The API allows receiving live tweets as JSON objects. For a program to interact with Twitter, it has to be registered and issued an access token, access token secret, customer key, and customer secret. Below is a snippet of how these values are encoded in the App.config configuration file (see Figure 5). After the program establishes connection with the Twitter service, it can send a query with any number of keywords. Twitter only considers OR operation, so if a tweet is posted and it has any of the keywords that tweet will be sent, as JSON object, to the Data Collection program. Keywords can be multi-word phrases. Figure 6 shows the preparations required of connecting to Twitter streaming API.

```
<?xml version="1.0" encoding="utf-8"?>
<configuration>
  <appSettings>
    <add key="loglevel" value="ALL"/>
    <add key="use_queue" value="false"/>
    <add key="multithread" value="false"/>
    <add key="customer_key" value="9hZk0YgdgivV8mVyG8PwQ"/>
    <add key="customer_secret" value="3kd2GOGi3LW7oaI8zbgP7CEQyWCS1fA62QAe61RmtI"/>
    <add key="access_token" value="287299076-MBBNAqox2hsVKupP7eRQww22Bms4VKn3cmoTeUa8"/>
    <add key="access_token_secret" value="unnxhxXJQL1Mc0kWrHvUwuysp7Yc7Zx02HyJGDvDmMiPF"/>
    <add key="stream_url" value="https://stream.twitter.com/1.1/statuses/filter.json"/>
```

Figure 5. Twitter Authentication in App.config

```
string postparameters = "&track=iwatch";
string streamUrl = ConfigurationManager.AppSettings"stream_url";
string oauthToken = ConfigurationManager.AppSettings"oauth_token";
string oauthTokenSecret = ConfigurationManager.AppSettings"oauth_token_secret";
string oauthConsumerKey = ConfigurationManager.AppSettings"oauth_consumer_key";
string oauthConsumerSecret =
ConfigurationManager.AppSettings"oauth_consumer_secret";
string oauthVersion = ConfigurationManager.AppSettings"oauth_version";
string oauthSignatureMethod =
ConfigurationManager.AppSettings"oauth_signature_method";


string baseFormat =
    "oauth_consumer_key={0}&oauth_nonce={1}&oauth_signature_method={2}" +
   "&oauth_timestamp={3}&oauth_token={4}&oauth_version={5}" + postparameters;

var baseString = string.Format(baseFormat, oauthConsumerKey, oauth_nonce,
                               oauthSignatureMethod, oauth_timestamp, oauthToken,
                               oauthVersion);

baseString = string.Concat("POST&", Uri.EscapeDataString(streamUrl), "&",
                               Uri.EscapeDataString(baseString));

var compositeKey = string.Concat(Uri.EscapeDataString(oauthConsumerSecret),
                                 "&", Uri.EscapeDataString(oauthTokenSecret));
```

Figure 6. Twitter Connection Preparation

Figure 7 shows a code snippet that demonstrates how a Twitter query is submitted and the JSON response is received and prepared for processing.

```
webResponse = (HttpWebResponse)webRequest.GetResponse();
responseStream = new StreamReader(webResponse.GetResponseStream(), encode);

while (noEndofStream){

         jsonText = responseStream.ReadLine();
         dynamic obj = JsonUtils.JsonObject.GetDynamicJsonObject(jsonText);
}
                                                              }
```

Figure 7. Twitter receiving JSON objects

The program then parses those objects, identifies a pre-defined set of important elements, and stores them in a SQL Server database table. When the program receives a JSON object, it parses the received object to extract the values of the following keys: id; text; created_at; in_reply_to_screen_name; in_reply_to_user_id; in_reply_to_status_id; retweet_count; user.id; user.screen_name; user.followers_count; user.friends_count; user.favorites_count; user.statuses_count. Figure 8 shows a diagram of the main components of a sample Twitter JSON object. An actual Twitter JSON object is provided in Appendix A.

The extracted values for these keys are then inserted into a relational database table. As an example, the below SQL statement was used to create the Jaguars table (see Figure 9). The pseudo code of the Data Collection program is shown in Figure 10, and the complete listing of the program's source code is provided in Appendix II.

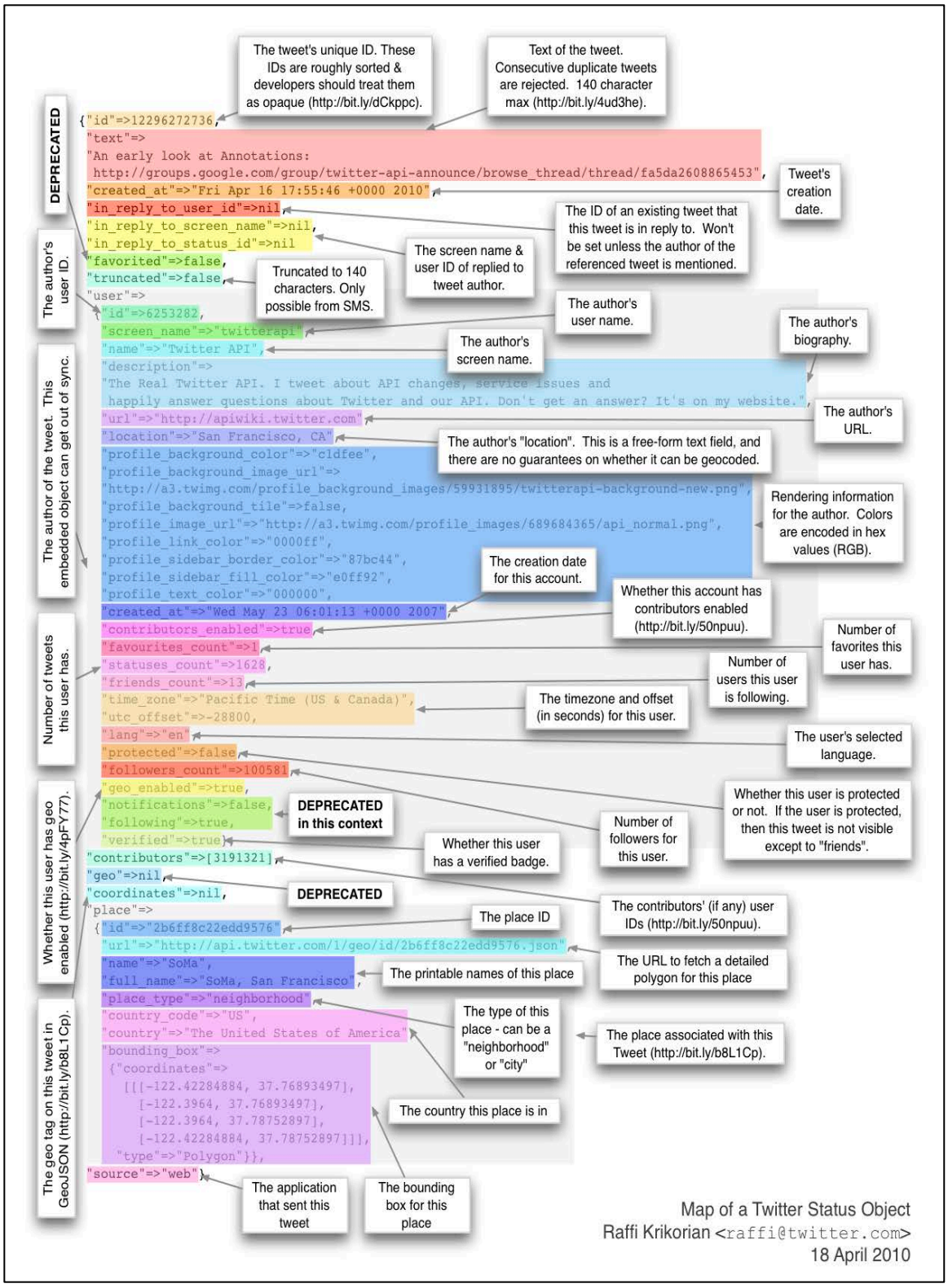

Figure 8. Sample Twitter JSON Object

```
USE [Twitter]
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [Jaguars](
      [tweetId] [bigint] NOT NULL,
      [tweetText] [nvarchar](255) NULL,
      [tweetCreatedAt] [datetime] NULL,
      [tweetInReplyToScreenName] [nvarchar](50) NULL,
      [tweetInReplyToUserId] [bigint] NULL,
      [tweetInReplyToStatusId] [bigint] NULL,
      [tweetFavoriteCount] [int] NULL,
      [userId] [bigaint] NULL,
      [userScreenName] [nvarchar](50) NULL,
      [userFollowersCount] [int] NULL,
      [userFriendsCount] [int] NULL,
      [userFavoritesCount] [int] NULL,
      [userStatusesCount] [int] NULL,
      [IfRetweetedTweetText] [nvarchar](300) NULL,
      [NumberOfRetweets] [int] NULL,
      [NumberOfRetweetFollowers] [int] NULL,
      [NumberOfReplies] [int] NULL,
 CONSTRAINT [PK_Tweets] PRIMARY KEY CLUSTERED
(
      [tweetId] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY = OFF,
ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
```

Figure 9. SQL Statement to Create Jaguars Table

```
Load authentication information from App.config
Compose Twitter search query using Keywords
Create a message queue if one doesn't exist
While (true){
      Send query with authentication information to Twitter API service
      using POST method
      Get a Twitter JSON objects stream
      While (stream has more objects){
            Parse the current JSON object
            Extract interesting parameters from the current JSON object
            Compose a SQL query with the extracted parameter values
            Execute a database INSERT query to insert a new record
            Advance to the next JSON object in the stream
      }
}
```

Figure 10. Pseudo Code for the Data Collection Program

5.2 Screen-Scraping Program

The second program is for screen scraping. A week after a tweet is posted, the screen scraping program reads that individual tweet record from the database, composes the URL of that tweet, downloads that tweet's webpage, parses the HTML content, and extracts the number of favorites. The screen-scraping program performs these operations for each tweet for each dataset. Figure 11 provides the pseudo code for the program. The complete listing of the Screen Scraping program is included in Appendix C. Additionally, Figure 12 provides a depiction of the data collection and screen-scraping processes. The following sections describe the datasets collected for the three experiments.

```
While (true){
      Select tweet from table if createdAt is < one week ago and not
      processed
      Extract interesting values
      Construct URL in the format https://twitter.com/" + screenName +
      "/status/" + tweetID);
      Download HTML of URL using WebClient()
      Parse the HTML page
      Extract the number of likes (favorites)
      Update the tweet's record with the extracted value of likes
}
```

Figure 11. Pseudo Code for the Screen-scraping Program

5.3 Collection of Datasets

Three datasets were collected for the three experiments. The first dataset focused on the Jaguars NFL team between September 25 and September 30, 2016, during the NFL season. The Jacksonville, Florida, is the Jaguars hometown. On September 25 the Jacksonville Jaguars lost to the Baltimore Ravens. That topic collected 58,531 tweets.

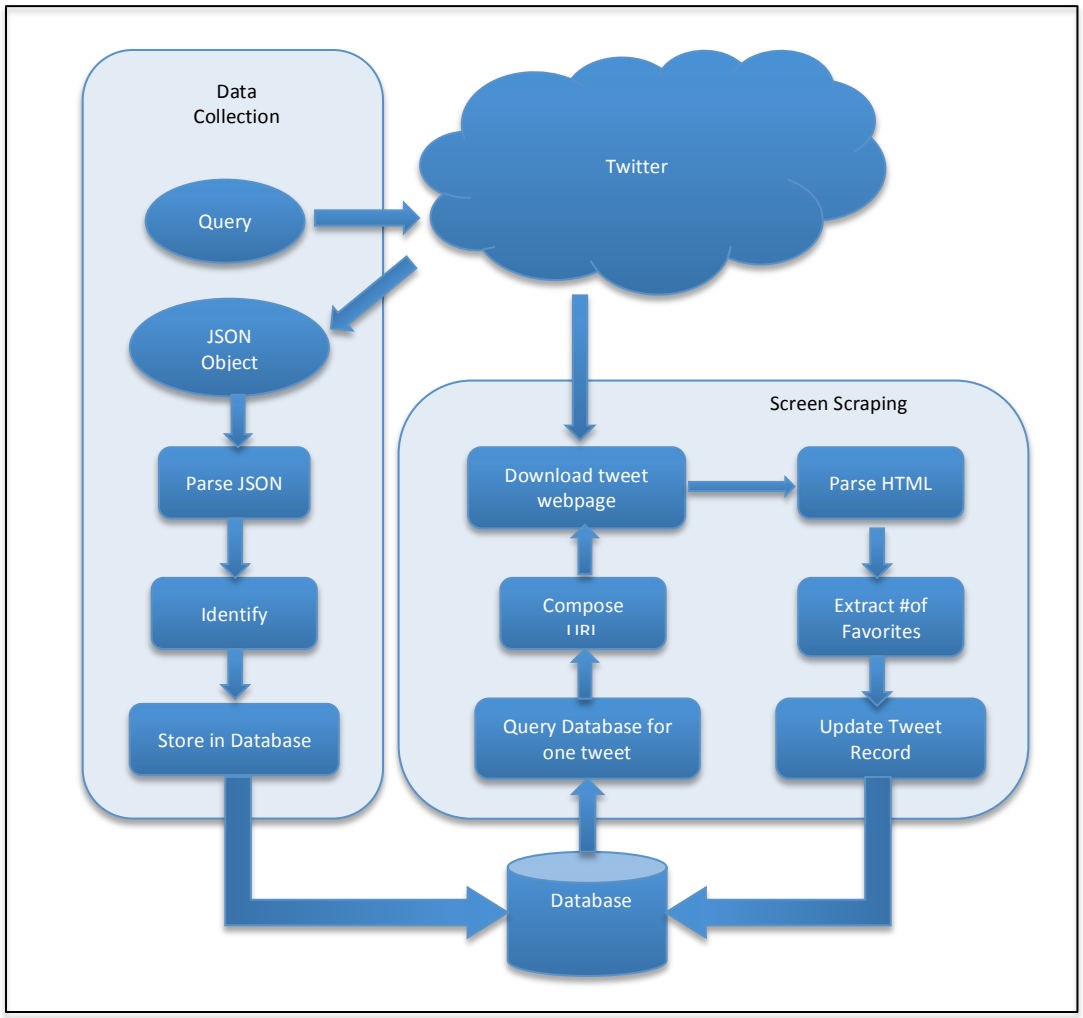

Figure 12. Data Collection and Screen-scraping Processes

The second experiment focused on a global issue, climate change. Data were collected from October 26 through November 2, 2016. 294,758 tweets were collected during this experiment.

The last experiment was related to Hurricane Matthew that hit the southeast United States in the fall of 2016. Hurricane Matthew was an extremely destructive and long-lived hurricane that reached category five, which is the highest hurricane category, at sometimes and was at category four for most of the time. It claimed the lives of 49 individuals in the United States. The data collection started on the early morning of October 7, when the hurricane was near Florida shores, and ended on October 14. That experiment collected more than two million tweets (exactly 2,164,142). Although the hurricane had a serious regional impact on the southeast United States, it attracted significant attention both nationally and internationally. The following table, Table 1, summarizes the three datasets.

	Jaguars (NFL)	**Climate Change**	**Hurricane Matthew**
Start Time (EST)	9/25/2016 18:08:59	10/26/2016 11:53:20	10/7/2016 04:54:08
End Time (EST)	9/30/2016 18:08:59	11/02/2016 11:53:20	10/14/2016 04:54:08
Number of Tweets	58,531	294,758	2,164,142
Number of Retweets	13,826	52,065	602,461
Unique Users	31,127	174,324	1,026,769
Average Tweets Per User	1.88	1.69	2.11

Table 1. Datasets Overview

Although each dataset was stored in a different database table, the three tables have the same set of attributes. Each row represents an individual tweet. The attributes of the tables are described in Table 2 below.

Column Name	Data Type	Description	Remarks
tweetId	bigint	Tweet ID	Automatically set by Twitter
tweetText	nvarchar (255)	Tweet Text	As tweeted by user
tweetCreatedAt	datetime	Tweet Creation Date and Time	Automatically set by Twitter
tweetInReplyToScreenName	nvarchar (50)	Screen Name to whom this tweet is a reply to	Automatically set by Twitter
tweetInReplyToUserId	bigint	User ID to whom this tweet is a reply to	Automatically set by Twitter
tweetInReplyToStatusId	bigint	Tweet ID to which this tweet is a reply to	Automatically set by Twitter
tweetFavoriteCount	int	Number of times this tweet is favorite by other users	Value extracted and set by the screen scraping program
userId	bigint	ID of user made the tweet	Automatically set by Twitter
userScreenName	nvarchar (50)	Screen name of user made the tweet	Automatically set by Twitter
userFollowersCount	int	Number of followers of user made the tweet	Automatically set by Twitter
userFriendsCount	int	Number of users followed by the user	Automatically set by Twitter
userFavoritesCount	int	Total number of favorites received by the user	Automatically set by Twitter
userStatusesCount	int	Total number of tweets made by the user	Automatically set by Twitter
IfRetweetedTweetText	nvarchar (300)	The tweet text should the tweet be retweeted	Set by a SQL query
NumberOfRetweets	int	Number of times this tweet is retweeted	Value extracted and set by the screen scraping program
NumberOfRetweetFollowers	int	Number of followers of all users who retweeted this tweet	Set by a SQL script
NumberOfReplies	int	Number of times this tweet was replied to by other users	Set by a SQL script

Table 2. Tweets Table

Chapter 6

ANALYSIS OF RESULTS

Summarized datasets were collected from each of the three database tables, Jaguars, Matthew and Climate Change, such that each row in the resulting summary datasets represents individual users. The query used to produces the summary from the Jaguars table is shown in Figure 13. The queries used to collect data from the Climate Change and Hurricane Matthew tables are very similar, except for the table name.

```
SELECT          userid,
                MAX(userscreenname) ScreenName,
                COUNT(*) Tweets,
                MAX(userfollowerscount) Followers,
                MAX(userfriendscount) Friends,
                MAX(userstatusescount) Statuses,
                SUM(tweetfavoritecount) Favorites,
                SUM(NumberOfReplies) Replies,
                SUM(numberofretweets) Retweets,
                SUM(numberofretweetfollowers) CummFollowers
FROM            jaguars
GROUP BY        userid order by CummFollowers desc
```

Figure 13. Collecting Data From Jaguars Table

The summary datasets had 31,127 users from the Jaguars, 174,324 users from Climate Change, and 1,026,769 users from Hurricane Matthew. The attributes in each summary dataset are described in Table 3.

Attribute Name	Attribute Description	Remarks
UserID	Twitter user ID as assigned by Twitter.	Copy from Tweets table, Table 2.
ScreenName	Twitter user screen name.	Copy from Tweets table, Table 2.
Tweets	Total number of tweets made by the user for the topic during the data collection period.	Calculated for every user using SQL's Count function and Group By statement on userID attribute from Table 2 above.
Followers	Total number of followers for the user.	Calculated for every user using SQL's Max function on userFollowersCount from Tweets table and Group By statement on userID attribute.
Friends	Total number of users followed by the user.	Calculated for every user using SQL's Max function on userFriendsCount from Tweets table and Group By statement on userID attribute.
Statuses	Total number of tweets made by the user at all times.	Calculated for every user using SQL's Max function on userStatusesCount from Tweets table and Group By statement on userID attribute.
Favorites	Total number of favorites for all tweets posted by the user for the topic during the data collection period.	Calculated for every user using SQL's Sum function on tweetFavoriteCount from Tweets table and Group By statement on userID attribute.
Replies	Total number of replies received for all tweets posted by the user for the topic during the data collection period.	Calculated for every user using SQL's Sum function on NumberOfReplies from Tweets table and Group By statement on userID attribute. NumberOfReplies is calculated using a SQL Script.
Retweets	Total number of retweets received for all tweets posted by the user during the data collection period.	Calculated for every user using SQL's Sum function on NumberOfRetweets from Tweets table and Group By statement on userID attribute.
CummFollowers	The total number of users who were exposed the tweets posted by the user during the data collection period. This includes the followers of users who retweeted the user's tweets on the topic.	Calculated for every user using SQL's Sum function on NumberOfRetweetFollowers from Tweets table and Group By statement on userID attribute. NumberOfRetweetFollowers is calculated using a SQL Script. The script creates an intermediate column and populates it with how a tweet will look if it is retweeted. Retweets, are tweets that have the same text of the original tweet with a "RT @" prefix. The script then scans the Tweets table for each tweet to find if there are matches to its retweet form. For a table of n tweets that is n table scans, which is very time consuming even after creating an index on the intermediate column.

Table 3. List of Attributes in Summary Datasets

As indicated in above, our one of the objectives of this thesis is to identify topic influencers in Twitter. To achieve this we need to compare the accuracy of the composite influence score to the previous methods using the measure of information diffusion

(represented here with CummFollowers). As stated earlier, information diffusion, or information spread, is a measure of the spread of contagions (tweets in this research) through actions such as sharing and forwarding (favorite, reply and retweet in this research) enabled by social networks. Information diffusion has been used in many research studies (Gomez-Rodriguez et al., 2016; Herzig et al., 2014; Kempe et al., 2003). The CummFollowers attribute represents the information diffusion over the respective experiment duration.

6.1 Analysis of the Jaguars Dataset

As stated above, the summary dataset has 31,127 unique users. The user with the largest CummFollowers, posted 81 tweets on the Jaguars topic during the data collection period (from 18:08:59 September 25, 2016 to 18:08:59 September 30, 2016). Those 81 tweets were retweeted 1170 times, marked as favorite 3,329 times, replied to 506 times, and most importantly reached 62,329,092 users (CummFollowers). It should not be surprising that the most influential user has the screen name of “Jaguars” and is the team’s official account.

6.1.1 Regression Analysis of Jaguars Dataset

Regression analyses were performed on the Jaguars dataset (excluding the UserID and ScreenName because they represent IDs and do not convey a user characteristic or activity). The first analysis focused on determining the regression equation of the

CummFollowers and the independent variables of the composite score as indicated in Equation 5 above, i.e. *Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets.* The purpose of this analysis is to determine the values of weight factors *w1, w2, w3, w4, w5, w6 and w7* that would calculate a composite score that is closest to CummFollowers. WEKA 3.8 (Frank, Hall, & Witten, 2016) was used for that analysis. Weka produced a regression equation represented here by Equation 6.

CummFollowers = -74.5647 * *Tweets* - 0.0387 * *Followers* +
0.1292 * *Friends* + 0.0116 * *Statuses* - 11.8512 * *Favorites* +
119219.7589 * *Replies* + 704.7357 * *Retweets* - 2689.4623

Equation 6. Jaguars Regression: Composite Score

Similar analyses were done using the independent variables associated with the topology, user characteristics, and topic sensitivity models, as discussed in Chapter 2. Equations 7, 8 and 9 represent the regression equation of each model independent variables with the CummFollowers, respectively.

CummFollowers = 0 * *Friends* + 0.06 * *Followers* + 2802.2773

Equation 7. Jaguars Regression: Topology

CummFollowers = 0 * *Statuses* + 980.5443 * *Tweets* + 0 * *Friends* + 1281.6571

Equation 8. Jaguars Regression: User Characteristics

CummFollowers = -52.153 * *Tweets* - 12.9031 * *Favorites* +
120841.8945 * *Replies* - 2124.0443

Equation 9. Jaguars Regression: Topic Sensitivity

A [0, 10] discrete influence score value was computed for each method using the above equations, such that the percentile, rounded to the nearest ten, for each user with non-zero CummFollowers value. Score value was set to zero for all users with zero CummFollowers. A user with influence score=10 is one with 95 percentile CummFollowers or higher. Similarly, influence core=9 represents a user with CummFollowers percentile in the range [85,94], and influence score=8 represents a user with CummFollowers percentile in the range [75,84], etc. This mapping was performed for the composite score model, and the surveyed models (topology, user characteristics and topic sensitivity).

Three analyses were performed on the Jaguars dataset, one for the 95 or higher percentile influencers, one for the 85 or higher, and one for 75 or higher, to evaluate the accuracy of the composite score model, as compared to the surveyed models.

Table 4 summarizes the number of users identified to be influential, based on 95 percentile (influence score=10), 85 percentile (influence score=10) and 75 percentile (influence score=10), using CummFollowers (representative of information diffusion) as the baseline. The data in Table 4 show that the CummFollowers identified 88 users with influence score 10 (i.e. ≥95 percentile), of those the composite score model identified 30 matching users of the same influence score of 10, which represents 23%. That means the composite score model is 23% accurate. That is a slightly larger accuracy when compared to the accuracy of the topology, user-characteristics or topic-sensitive models, which have accuracies of 18%,

14% and 22%, respectively. If we compare the accuracy of those methods in identifying influential users of score 8 or higher (≥75 percentile), we find that the composite score model is 23% accurate, while topology, user-characteristics and topic-sensitive methods have accuracies of 20%, 20% and 18%, respectively. Those results show that the composite score model is more accurate in identifying influential users compared to the models surveyed in the literature.

Percentile	**Cumm Followers**	**Topology Matches**		**User-Char. Matches**		**Topic-Sens. Matches**		**Composite Score Matches**	
	Number of Users	Number of Users	%	Number of Users	%	Number of Users	%	Number of Users	%
≥95	88	16	18%	12	14%	19	22%	20	23%
≥85	262	49	19%	49	19%	51	19%	81	31%
≥75	436	87	20%	89	20%	79	18%	99	23%

Table 4. Jaguars Dataset Comparison of Influencer Identification Methods

6.1.2 Attribute-Ranking for Jaguars Dataset

Attribute-ranking analysis using Information Gain (Berrar & Dubitzky, 2013) and Chi-Square (Chernoff & Lehmann, 2012) algorithms, both available in WEKA 3.8, was performed on the composite score model independent variables. Table 5 shows the ranking of each attribute with respect to identifying the ranking score. Both algorithms resulted in the same ranking of the independent variables. While *Friends* and *Statuses* were ranked at the bottom, *Retweets* and *Tweets* were ranked on top.

Attribute	Info Gain		Chi-Squared	
	Rank	**Value**	**Rank**	**Value**
Retweets	1	0.30906	1	56663.068
Tweets	2	0.03986	2	3398.812
Followers	3	0.02788	3	2584.104
Replies	4	0.01324	4	1891.0655
Favorites	5	0.00828	5	447.1757
Friends	6	0.00456	6	164.3188
Statuses	7	0.00339	7	142.523

Table 5. Jaguars Attribute Ranking

6.1.3 Predictive Analysis of Jaguars Dataset

Regression analysis was used to determine the weight factors relating the independent variables to CummFollowers. Regression analysis showed that the composite score model is more accurate in identifying the influential users than the surveyed methods using the Jaguars dataset. Besides using regression analysis, WEKA was used to explore the predictability of Twitter user's influence using machine-learning algorithms. In other words, additional methods were used to predict value of CummFollowers, so one does not have to calculate it, and consequently avoid social graph parsing. The seven independent variables of the composite score model (*Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets*) were used along with CummFollowers as the dependent variable with the Voting Feature Interval (VFI) (Demiroz & Guvenir, 1997) and the J48 decision tree algorithms (Quinlan, 1993). In both algorithms 66% of the dataset was used to train the model, and the remaining 34% was used for testing. Both sets were randomly identified by WEKA. The Confusion matrices for the test dataset as produced by both algorithms are presented in Table 6 and Table 7.

Actual Score	Predicted Score										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	9491	201	70	30	12	24	2	47	28	71	63
1	3	6	17	0	22	2	3	1	1	2	0
2	0	6	12	0	27	0	3	7	1	1	0
3	1	5	16	1	18	0	13	3	1	1	0
4	0	2	9	3	14	0	9	6	1	1	3
5	2	3	14	0	18	1	8	7	2	1	3
6	0	1	13	2	14	0	6	15	3	0	1
7	0	2	10	0	5	1	8	18	7	1	3
8	0	2	10	0	5	1	4	9	14	9	3
9	0	1	3	2	6	0	3	10	5	15	18
10	2	1	1	0	1	0	2	0	2	9	16

Table 6. Jaguars VFI Confusion Matrix

The dark shaded cells in the below table show the number of users whose influence scores were correctly predicted, i.e. match the information diffusion score. Out of the 10,583 users in the test portion, the influence score of 9,594 users were correctly predicted, which is about 91% accuracy. However, if only most influential users were considered (identified here as those of score 8 or higher), the influence scores of only 45 users (out of 154) were accurately predicted with a low accuracy rate of 29%. If a small range of tolerance is allowed, we may consider the prediction of a user's score of 8, 9 or 10 as a valid prediction, without being particular about the exact value, the algorithm's accuracy in predicting influential users goes up to 59% (91 out of 154 users). It is worth noting that using VFI with the recommended tolerance produces higher predictive accuracy than regression.

The Confusion matrix resulting from the J48 algorithm is shown in Table 7 below. The overall predictive accuracy for the test dataset is about 96%; however, the accuracy of predicting influential users is only 27%. If tolerance is applied the accuracy goes up to

56%. Although J48 is better than regression in predicting the influence score of influential users, it is slightly less accurate than the VFI algorithm. WEKA has numerous algorithms and most of them were tried but VFI and J48 were among few algorithms that produced better results than regression and for that reason are included in this thesis.

Actual Score	Predicted Score										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	10012	11	3	3	1	5	3	1	0	0	0
1	7	10	9	7	7	9	4	2	0	2	0
2	3	15	4	7	7	6	7	3	4	1	0
3	6	7	8	10	6	8	4	5	5	0	0
4	4	4	4	9	10	10	4	0	3	0	0
5	5	9	6	9	9	1	10	6	1	3	0
6	1	5	9	6	4	7	11	5	7	0	0
7	4	1	1	3	6	6	6	13	11	3	1
8	3	2	4	5	2	4	3	8	9	11	6
9	2	6	4	0	4	6	1	5	11	21	3
10	1	2	0	1	2	2	0	0	4	11	11

Table 7. Jaguars J48 Confusion Matrix

Although the composite score model proved to be more accurate in identifying influential users when compared to the topology, user-characteristics, and topic-sensitive models as discussed in section 6.1.1. The additional experiments discussed in this section demonstrate that other predictive models could also produce even better results. Clearly, VFI and J48 are superior in accurately identifying influential users to regression analysis using the Jaguars dataset. In the following sections, the same methodology will be used to explore whether those results will hold for other datasets, or are only pertinent to the Jaguars dataset.

6.2 Analysis of the Climate Change Dataset

The Climate Change summary dataset has 174,324 unique users. The most influential user posted only two tweets on the Climate Change topic during the data collection period (from October 26 11:53:20 to November 2 11:53:20, 2016). Those two tweets were retweeted 7796 times, marked as favorite zero times, replied to 36 times, and reached 23,044,692 users. The most influential user has the screen name of "SenSanders", that is the official Twitter account of United States Senator Bernie Sanders (D).

6.2.1 Regression Analysis of Climate Change Dataset

Regression analyses were performed on the Climate Change dataset. The first analysis focused on determining the regression equation of the CummFollowers and the independent variables of the composite score as indicated in Equation 5 above (*Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets).* The purpose of this analysis is to determine the values of weight factors *w1, w2, w3, w4, w5, w6 and w7* that would calculate a composite score as close as possible to CummFollowers for each user. WEKA 3.8 (Frank et al., 2016) was used for that analysis. Weka produced a regression equation represented here by Equation 10.

CummFollowers = 23.9952 * *Tweets* - 0.0006 * *Followers* - 0.5615 * *Favorites* + 1034.1691 * *Replies* + 2735.0472 * *Retweets* + 130.4577

Equation 10. Climate Change Regression: Composite Score

Similar analyses were performed using the independent variables associated with the topology, user characteristics, and topic sensitivity models, as discussed in Chapter 2. Equations 11, 12 and 13 represent the regression equation of each model independent variables with the CummFollowers, respectively.

$$CummFollowers = 0.0199 * Followers + 0.079 * Friends + 751.6481$$

Equation 11. Climate Change Regression: Topology

$$CummFollowers = 221.9099 * Tweets + 0.1062 * Friends + 453.9627$$

Equation 12. Climate Change Regression: User Characteristics

$$CummFollowers = 135.527 * Tweets + 33581.4648 * Replies + 0 * Favorites + 330.3731$$

Equation 13. Climate Change Regression: Topic Sensitivity

A [0, 10] discrete influence score value was computed for each method using the above equations, such that the percentile, rounded to the nearest ten, for each user with non-zero CummFollowers value. Score value was set to zero for all users with zero CummFollowers. A user with influence score=10 is one with 95 percentile CummFollowers or higher. Similarly, influence core=9 represents a user with CummFollowers percentile in the range [85,94], and influence score=8 represents a user with CummFollowers percentile in the range [75,84], etc. This mapping was performed for the composite score model, and the surveyed models (topology, user characteristics and topic sensitivity).

Three analyses were performed on the Climate Change dataset, similar to those used in the analysis of the Jaguars dataset, one for the 95 or higher percentile influencers, one for the 85 or higher, and one for 75 or higher, to evaluate the accuracy of the composite score model, as compared to the surveyed models.

Table 8 summarizes the number of users identified to be influential, based on 95 percentile (influence score=10), 85 percentile (influence score=10) and 75 percentile (influence score=10), using CummFollowers as the baseline. The data in Table 8 show that the CummFollowers identified 343 users with influence score 10, of those the composite score model identified 193 matching users of the same influence score of 10, which represents 56%. That means the composite score model is 56% accurate. That is a much higher accuracy than the accuracies of the topology, user-characteristics or topic-sensitive models. Those models have accuracies of 13%, 10% and 16%, respectively. If we compare the accuracies of those methods in identifying influential users of score 8 or higher ≥75 percentile), we find that the composite score model is 67% accurate, while topology, user-characteristics and topic-sensitive methods have accuracies of 15%, 14% and 17%, respectively. Those results show that the composite score model is more accurate in identifying influential users compared to the models surveyed in the literature.

Percentile	**Cumm Followers**	**Topology Matches**		**User-Char. Matches**		**Topic-Sens. Matches**		**Composite Score Matches**	
	Number of Users	Number of Users	%	Number of Users	%	Number of Users	%	Number of Users	%
≥95	343	43	13%	33	10%	56	16%	193	56%
≥ 85	1027	136	13%	116	11%	162	16%	639	62%
≥75	1712	250	15%	237	14%	296	17%	1143	67%

Table 8. Climate Change Dataset Comparison of Influencer Identification Methods

Attribute	Info Gain		Chi-Squared	
	Rank	Value	Rank	Value
Retweets	1	0.23841	1	314852.416
Tweets	3	0.01828	3	9083.498
Followers	2	0.02492	2	12794.629
Replies	6	0.00569	4	5629.816
Favorites	4	0.01033	6	3081.471
Friends	5	0.00914	5	3196.606
Statuses	7	0.00526	7	1321.333

Table 9. Climate Change Attribute Ranking

6.2.2 Attribute-Ranking for Climate Change Dataset

Attribute-ranking analysis was performed on the Climate Change dataset using Information Gain (Berrar & Dubitzky, 2013) and Chi-Square (Chernoff & Lehmann, 2012) algorithms. Table 9 shows the ranking of each attribute with respect to identifying the ranking score. The ranking of the attributes using the two methods were similar, but not exactly the same. *Statuses* was ranked last and *Retweets* and *Followers* were ranked first and second, respectively, in both methods.

6.2.3 Predictive Analysis of Climate Change Dataset

Regression analysis showed that the composite score model is more accurate in identifying the influential users than the surveyed methods using the Climate Change dataset. WEKA was used to explore the predictability of Twitter user's influence using machine-learning algorithms. Voting Feature Interval (VFI) (Demiroz & Guvenir, 1997) and the J48 decision tree algorithms (Quinlan, 1993) were applied on the seven independent variables

of the composite score model (*Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets*) and the CummFollowers dependent variable using WEKA to explore the predictability of Twitter user's influence using the Climate Change dataset. VFI and J48 algorithms were used such that 66% of the dataset was used to train the model, and the remaining 34% was used for testing. Both sets were randomly identified by WEKA. The Confusion matrices for the test dataset as produced by both algorithms are presented in Table 10 and Table 11.

Actual Score	**Predicted Score**										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	56654	77	0	37	3	1	11	18	18	39	250
1	14	132	0	91	10	0	1	1	0	0	4
2	5	98	0	93	17	0	0	7	0	1	6
3	14	95	0	75	14	1	0	14	6	0	3
4	14	65	0	102	17	1	0	22	9	1	3
5	9	43	0	80	11	0	0	22	30	2	8
6	7	52	0	74	9	0	1	29	33	5	5
7	6	27	0	77	12	0	0	32	54	15	11
8	2	19	0	49	8	1	2	29	75	19	22
9	6	13	0	23	4	0	0	22	78	37	42
10	1	0	0	3	0	0	0	7	16	20	74

Table 10. Climate Change VFI Confusion Matrix

The dark shaded cells in the above table show the number of users whose influence scores were correctly predicted. Out of the 59,270 users in the test portion, the influence score of 57,097 users were correctly predicted, which is about 96% accuracy. However, if only influential users were considered (those of score 8 or higher), the influence scores of only 186 users (out of 572) were accurately predicted with a low accuracy rate of 33%. If a small range of tolerance is allowed we may consider the prediction of a user's score of 8, 9 or 10 as a valid prediction, without being particular about the exact value, the

algorithm's accuracy in predicting influential users goes up to 67% (383 out of 572 users). VFI with the recommended tolerance produces similar predictive accuracy to regression analysis.

The Confusion matrix resulting from the J48 algorithm is shown in Table 11 below. The overall predictive accuracy for the test dataset is about 97%; however, the accuracy of predicting influential users is only 28%. If tolerance is applied the accuracy goes up to 56%. Results suggest that J48 is less accurate in predicting the influence score of influential users than regression analysis and the VFI algorithm. Although other predictive algorithms were applied using WEKA, VFI and J48 were most accurate using the Climate Change dataset.

Actual Score	Predicted Score										
	0	1	2	3	4	5	6	7	8	9	10
0	57013	23	24	10	17	8	9	2	0	2	0
1	35	59	42	19	37	23	18	9	3	8	0
2	21	46	36	31	29	21	23	10	3	6	1
3	18	39	29	36	30	29	11	18	3	9	0
4	17	32	29	38	23	38	20	19	12	6	0
5	18	27	21	26	21	23	29	24	7	7	2
6	14	23	19	10	20	40	36	22	18	13	0
7	3	14	19	21	16	32	36	31	32	25	5
8	7	17	6	10	20	25	23	28	45	37	8
9	7	4	6	11	12	17	18	23	46	56	25
10	1	0	1	0	0	2	8	7	14	29	59

Table 11. Climate Change J48 Confusion Matrix

In addition to the Jaguars dataset, in the Climate Change dataset also proved that the composite score method is more accurate in identifying influential users when compare to topology, user-characteristics, and topic-sensitive methods as discussed in section 6.2.1.

The additional experiments discussed in this section demonstrate that other predictive models could also to produce good results. In the following sections, the same methodology is used to explore whether those results will continue to hold for the Hurricane Matthew dataset.

6.3 Analysis of the Hurricane Matthew Dataset

The Hurricane Matthew summary dataset has 1,026,767 unique users. The most influential user posted 107 tweets on Hurricane Matthew during the data collection period (from October 7 04:54:08 to October 21 7:55:12, 2016). Those tweets were retweeted 14,422 times, marked as favorite 27,785 times, replied to 102 times, and reached 57,649,357 users. The most influential user has the screen name of "CNN", that is the official Twitter account of the Cable News Network (CNN).

6.3.1 Regression Analysis of Hurricane Matthew Dataset

Regression analyses were performed on the Hurricane Matthew dataset. The first analysis focused on determining the regression equation of the CummFollowers and the independent variables of the composite score as indicated in Equation 5 above (*Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets).* The purpose of this analysis is to determine the values of weight factors *w1, w2, w3, w4, w5, w6 and w7* that would calculate a composite score as close as possible to CummFollowers for each user. WEKA 3.8 (Frank

et al., 2016) was used for that analysis. Weka produced a regression equation represented here by Equation 14.

$$\text{CummFollowers} = 232.6515 * \textit{Tweets} + 0.1058 * \textit{Followers} + 0.1595 * \textit{Friends} - 0.0116 * \textit{Statuses} + 5.5896 * \textit{Favorites} + 217621.6325 * \textit{Replies} + 1194.1698 * \textit{Retweets} - 110.8545$$

Equation 14. Hurricane Matthew Regression: Composite Score

Similar analyses were performed using the independent variables associated with the topology, user characteristics, and topic sensitivity models, as discussed in Chapter 2. Equations 15, 16 and 17 represent the regression equation of each model independent variables with the CummFollowers, respectively.

$$\text{CummFollowers} = 0.2837 * \textit{Followers} + 0.1759 * \textit{Friends} + 1818.3059$$

Equation 15. Hurricane Matthew Regression: Topology

$$\text{CummFollowers} = 1019.2117 * \textit{Tweets} + 0.5974 * \textit{Friends} + 0 * \textit{Statuses} + 239.4336$$

Equation 16. Hurricane Matthew Regression: User Characteristics

$$\text{CummFollowers} = 296.96 * \textit{Tweets} + 7.425 * \textit{Favorites} + 327686.9977 * \textit{Replies} + 69.811$$

Equation 17. Hurricane Matthew Regression: Topic Sensitivity

A [0, 10] discrete influence score value was computed for each method using the above equations, such that the percentile, rounded to the nearest ten, for each user with non-zero

CummFollowers value. Score value was set to zero for all users with zero CummFollowers. A user with influence score=10 is one with 95 percentile CummFollowers or higher. Similarly, influence core=9 represents a user with CummFollowers percentile in the range [85,94], and influence score=8 represents a user with CummFollowers percentile in the range [75,84], etc. This mapping was performed for the composite score model, and the surveyed models (topology, user characteristics and topic sensitivity).

Three analyses were performed on the Hurricane Matthew dataset, similar to those used in the analysis of the Jaguars and Climate Change datasets, one for the 95 or higher percentile influencers, one for the 85 or higher, and one for 75 or higher, to evaluate the accuracy of the composite score model, as compared to the surveyed models.

Table 12 summarizes the number of users identified to be influential, based on 95 percentile (influence score=10), 85 percentile (influence score=10) and 75 percentile (influence score=10), using CummFollowers as the baseline. The data in Table 12 show that the CummFollowers identified 2,493 users with influence score 10, of those the composite score model identified 756 matching users of the same influence score of 10, which represents 30%. That means the composite score model is 30% accurate. That is a much higher accuracy than the accuracies of the topology, user-characteristics or topic-sensitive models. Those models have accuracies of 16%, 12% and 22%, respectively. If we compare the accuracies of those methods in identifying influential users of score 8 or higher (≥75 percentile), we find that the composite score model is 42% accurate, while topology, user-characteristics and topic-sensitive methods have accuracies of 20%, 19%

and 15%, respectively. Those results show that the composite score model is more accurate in identifying influential users compared to the models surveyed in the literature.

Percentile	Info. Diff.	Topology Matches		User-Char. Matches		Topic-Sens. Matches		Composite Score Matches	
	Number of Users	Number of Users	%	Number of Users	%	Number of Users	%	Number of Users	%
≥ 95	2493	401	16%	293	12%	544	22%	756	30%
≥85	7479	1397	19%	1276	17%	1242	17%	3160	42%
≥75	12465	2431	20%	2377	19%	1877	15%	5256	42%

Table 12. Hurricane Matthew Dataset Comparison of Influencer Identification Methods

6.3.2 Attribute-Ranking for Hurricane Matthew Dataset

Similar to the other datasets, attribute-ranking analysis was performed on the Hurricane Matthew dataset using Information Gain (Berrar & Dubitzky, 2013) and Chi-Square (Chernoff & Lehmann, 2012) algorithms. Table 13 shows the ranking of each attribute with respect to identifying the ranking score. The ranking of the attributes using the two methods were less similar than the ranking in the other two datasets. *Replies* was ranked last using Information Gain and *Statuses* was ranked last using Chi-Square, while *Retweets* was ranked first by both methods.

Attribute	Info Gain		Chi-Squared	
	Rank	Value	Rank	Value
Retweets	1	0.27941	1	1863962.651
Tweets	3	0.02584	3	79390.664
Followers	2	0.02843	2	104686.355
Replies	7	0.00522	4	74654.338
Favorites	4	0.01399	5	25903.606
Friends	5	0.00904	6	21284.982
Statuses	6	0.00605	7	10379.291

Table 13. Hurricane Matthew Attribute Ranking

6.3.3 Predictive Analysis of Hurricane Dataset

Consistent with the results of the other two datasets, regression analysis showed that the composite score model is more accurate in identifying the influential users than the surveyed methods using the Hurricane Matthew dataset. Using the same dataset, Hurricane Matthew, WEKA was used to explore the predictability of Twitter user's influence using machine-learning algorithms. Voting Feature Interval (VFI) (Demiroz & Guvenir, 1997) and the J48 decision tree algorithms (Quinlan, 1993) were applied on the seven independent variables of the composite score model (*Favorite, Retweet, Reply, Followers, Statuses, Friends and Tweets*) and the CummFollowers dependent variable to explore the predictability of Twitter user's influence. VFI and J48 algorithms were used such that 66% of the dataset was used to train the model, and the remaining 34% was used for testing. Both sets were randomly identified by WEKA. The Confusion matrices for the test dataset as produced by both algorithms are presented in Table 14 and Table 15.

The dark shaded cells in the above table show the number of users whose influence scores were correctly predicted. Out of the 349,101 users in the test portion, the influence score of 334,187 users were correctly predicted, which is about 96% accuracy. However, if only influential users were considered (those of score 8 or higher), the influence scores of only 1920 users (out of 4211) were accurately predicted with a low accuracy rate of 46%. If a small range of tolerance is allowed we may consider the prediction of a user's score of 8, 9 or 10 as a valid prediction, without being particular about the exact value, the algorithm's accuracy in predicting influential users goes up to 95% (4021 out of 4211

users). With the recommended tolerance VFI produces higher predictive accuracy than regression.

Actual Score	Predicted Score										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	331850	9	16	329	0	2	1	5	623	38	190
1	2	1	0	532	1	0	0	1	1168	0	4
2	0	1	0	515	0	0	1	0	1172	2	6
3	1	0	0	415	0	0	1	0	1267	0	4
4	1	1	0	362	0	1	5	0	1311	0	4
5	3	1	0	268	0	0	0	0	1380	1	8
6	0	0	0	229	0	0	1	0	1462	2	7
7	0	0	0	140	0	0	0	0	1529	2	15
8	2	0	0	111	0	0	0	0	1484	5	40
9	2	0	0	66	0	0	0	1	1409	4	284
10	1	0	0	7	0	0	0	0	360	3	432

Table 14. Hurricane Matthew VFI Confusion Matrix

The Confusion matrix resulting from the J48 algorithm is shown in Table 15 below. The overall predictive accuracy for the test dataset is about 96%; however, the accuracy of predicting influential users is only 31%. If tolerance is applied the accuracy goes up to 57%. Although J48 is better than regression in predicting the influence score of influential users, it is less accurate than the VFI algorithm.

The Hurricane Matthew dataset also proved that the composite score model is more accurate in identifying influential users when compare to topology, user-characteristics, and topic-sensitive models as discussed in section 6.3.1. VFI and J48 demonstrated that they could produce more accurate prediction of influential users than regression, consistent with our finding in the other two datasets.

Actual Score	Predicted Score										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	332317	192	147	118	99	84	41	26	26	12	1
1	223	359	293	240	206	153	86	72	55	21	1
2	161	307	315	298	230	141	93	67	58	22	5
3	134	261	280	282	214	202	147	81	59	23	5
4	144	250	256	240	225	198	142	115	85	27	3
5	106	172	194	202	200	210	214	178	122	56	7
6	91	155	158	184	179	208	244	212	159	98	13
7	79	124	121	131	157	162	233	280	249	121	29
8	45	90	108	93	112	108	185	252	328	261	60
9	41	42	56	55	79	92	124	179	338	518	242
10	7	9	13	10	10	18	25	42	75	147	447

Table 15. Hurricane Matthew J48 Confusion Matrix

6.4 Summary

The above results show some consistent patterns in the three datasets. The first of which is the ranking of top relevant attributes to the influence score. In all experiments, using Information Gain and Chi-square algorithms, the Retweets parameter was always on top. Also, Tweets and Followers were usually ranked high. On the other hand, Statuses nearly always ranked last. Also, Friends usually ranks very low.

The second observation is about the regression analysis performed on the three datasets. Looking at the accuracies of the influence score models, it is clear that the composite score model, using regression formulas (Equations 6, 10 and 14), was consistently more accurate than the topology, user-characteristics, and topic based models (see Tables 4, 8 and 12).

Additionally, VFI and J48 were more accurate in predicting the influence score of users than regression, almost in every experiment, as shown in Table 16 below. Given that the vast majority of users are not influential, the overall accuracy may be misleading. The accuracy of identifying the influential users is much lower than the overall accuracy, but using a more flexible counting scheme, where distinguishing 75th percentile from 85th percentile or 95th percentile is not particularly important, can significantly improve that accuracy. Last, from table 16, it is clear that VFI produces more accurate predictions than J48 and regression.

	Jaguars	**Climate Change**	**Hurricane Matthew**
Composite Score Accuracy of Influential Users	23%	56%	30%
VFI Overall Accuracy	91%	96%	96%
VFI Accuracy of Influential Users	29%	33%	46%
VFI Accuracy of Influential Users with Tolerance	59%	67%	95%
J48 Overall Accuracy	96%	97%	96%
J48 Accuracy of Influential Users	27%	28%	31%
J48 Accuracy of Influential Users with Tolerance	56%	56%	57%

Table 16. Summary of Various Algorithm Accuracies

Chapter 7

CONCLUSION

Social media sites and services have become the target for the marketing efforts of many organizations. With the steady growth of the number of business transactions that are performed on the Internet, the extent of such marketing efforts is expected to only grow. Literature shows that the accuracy of those efforts highly depends on identifying influential users who can evangelize and spread reviews for services and products offered by these organizations.

The three leading approaches for modeling user's influence on social media have been criticized for being static and not responsive to the dynamic nature of social media networks, or not practical in terms of covering important aspects such as user's interests. Therefore, there is a need for developing a user influence score that addresses the concerns present in existing models.

This thesis introduced a dynamic influence score model that takes into consideration multiple factors such as user's interest, reachability and impact. Three experiments were developed to evaluate the model by collecting twitter data and extracting values necessary for computing the composite influence score. The model evaluation compared the accuracy of the composite score using regression, VFI and J48 algorithms to topology, user-characteristics and topic based models. Results show that the composite score model is

more accurate than the topology, user characteristics and topic sensitivity models, in addition to being dynamic and comprehensive.

7.1 Future Directions

One possibility for extending this research would be through the creation of an automated end-to-end process that allows users to specify a topic (through keywords and key phrases) and a period of time and produce a sorted list of influential users. Such a system would allow other researchers to continue to develop newer influencer identification models and compare their results to the one presented here.

Another possibility would be to investigate the potential of adjusting the composite score approach to be adaptable to changes over an extended period of time such that more recent interactions have more impact on the score than older interactions.

REFERENCES

Print Publications:

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008). Identifying the influential bloggers in a community. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Palo Alto, California, USA. 207-218. doi:10.1145/1341531.1341559

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Manage.Sci., 57*(9), 1623-1639. doi:10.1287/mnsc.1110.1421

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in Bakshy large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA. 44-54. doi:10.1145/1150402.1150412

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Hong Kong, China. 65-74. doi:10.1145/1935826.1935845

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, Lyon, France. 519-528. doi:10.1145/2187836.2187907

Berrar, D., & Dubitzky, W. (2013). Information gain. In W. Dubitzky, O. Wolkenhauer, K. Cho & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 1022-1023). New York, NY: Springer New York. doi:10.1007/978-1-4419-9863-7_719"

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194-1197. doi:10.1126/science.1185231

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology, 113*(3), 702-734. doi:10.1086/521848

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. *4th International AAAI Conference on Weblogs and Social Media (ICWSM),*

Chernoff, H., & Lehmann, E. L. (2012). The use of maximum likelihood estimates in X^2 tests for goodness of fit. In J. Rojo (Ed.), *Selected works of E. L. lehmann* (pp. 541-549). Boston, MA: Springer US. doi:10.1007/978-1-4614-1412-4_47"

Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *N Engl J Med, 358*(21), 2249-2258. doi:10.1056/nejmsa0706154

Demiroz, G., & Guvenir, H. A. (1997). Classification by voting feature intervals. In M. van Someren, & G. Widmer (Eds.), *Machine learning: ECML-97: 9th european conference on machine learning prague, czech republic, april 23--25, 1997 proceedings* (pp. 85-92). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/3-540-62858-4_74"

Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems, 20*(1), 80-82.

Frank, E., Hall, M. A., & Witten, I. H. (2016). Weka workbench online appendix for data mining: Practical machine learning tools and techniques morgan kaufmann, fourth edition, 2016.

Gomez-Rodriguez, M., Song, L., Du, N., Zha, H., & Scholkopf, B. (2016). Influence estimation and maximization in continuous-time diffusion networks. *ACM Trans.Inf.Syst., 34*(2), 9:1-9:33. doi:10.1145/2824253

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology, 78*(6), 1360-1380.

Hanneman, R., & Riddle, M. (2005). *Introduction to social network methods*. University of California.

Herzig, J., Mass, Y., & Roitman, H. (2014). An author-reader influence model for detecting topic-based influencers in social media. *Proceedings of the 25th ACM Conference on Hypertext and Social Media,* Santiago, Chile. 46-55. doi:10.1145/2631775.2631804

Hevner, A. (2007). A three-cycle view of design science research. *Scandinavian Journal of Information Systems, 19*(2), 87-92.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Mis Q., 28*(1), 75-105.

Hill, S., Benton, A., Ungar, L., Macskassy, S., Chung, A., & Holmes, J. (2011). A cluster-based method for isolating influence on twitter. *21st Workshop on Information Technologies and Systems, WITS,* Shanghai, China.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science, 22*(2), 256-275.

Huang, P., Liu, H., Chen, C., & Cheng, P. (2013). The impact of social diversity and dynamic influence propagation for identifying influencers in social networks. *Web Intelligence,* 410-416.

Huang, P., Liu, H., Lin, C., & Cheng, P. (2013). A diversity-dependent measure for discovering influencers in social networks. *Information Retrieval Technology - 9th Asia Information Retrieval Societies Conference, {AIRS} 2013, Singapore, December 9-11, 2013. Proceedings,* 368-379. doi:10.1007/978-3-642-45068-6_32

Kempe, D., Kleinberg, J., & Tardos, \. (2003). Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Washington, D.C. 137-146. doi:10.1145/956750.956769

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web,* Raleigh, North Carolina, USA. 591-600. doi:10.1145/1772690.1772751

Leavitt, A., Burchard, E., Fisher, D., & Gilbert, S. (2009). The influentials: New approaches for analyzing influence on twitter. *Webecology Project,*

Lee, C., Kwak, H., Park, H., & Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. *Proceedings of the 19th International Conference on World Wide Web,* Raleigh, North Carolina, USA. 1137-1138. doi:10.1145/1772690.1772842

Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th International Conference on World Wide Web,* Beijing, China. 915-924. doi:10.1145/1367497.1367620

McPherson McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415-444. doi:10.1146/annurev.soc.27.1.415

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167-256. doi:10.1137/S003614450342480

Onnela, J. P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., . . . Barabasi, A. L. (2007). Structure and tie strengths in mobile communication networks. *Proc.Natl.Acad.Sci.USA, 104*(18), 7332-7336.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web.* Stanford University.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Found.Trends Inf.Retr., 2*(1-2), 1-135. doi:10.1561/1500000011

Peng, H., Zhu, J., Piao, D., Yan, R., & Zhang, Y. (2011). Retweet modeling using conditional random fields. *ICDM Workshops,* 336-343.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Paris, France. 807-816. doi:10.1145/1557019.1557108

Tang, L., & Liu, H. (2010). Graph mining applications to social network analysis. In C. C. Aggarwal, & H. Wang (Eds.), *Managing and mining graph data* (pp. 487-513) Springer.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of'small-world'networks. *Nature, 393*(6684), 409-410.

Watts, D. (2007). Challenging the influentials hypothesis. *WOMMA Measuring Word of Mouth, 3*(4), 201-211.

Weng, J., Lim, E., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining,* New York, New York, USA. 261-270. doi:10.1145/1718487.1718520

Ye, S., & Wu, S. F. (2010). Measuring message propagation and social influence on twitter.com. *Proceedings of the Second International Conference on Social Informatics,* Laxenburg, Austria. 216-231.

Electronic Sources:

comScore. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. Retrieved from https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior

Gantz, J., & Reinsel, D. (2010). The digital universe decade – are you ready? Retrieved from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf

Grove, J. (2012). Microsoft buys influence for bing with klout investment, partnership. Retrieved from http://venturebeat.com/2012/09/27/microsoft-bing-klout

Internet Live Stats. (2016). Internet users. Retrieved from http://www.internetlivestats.com/internet-users/

Klout. (2015a). Klout for business: American airlines. Retrieved from http://kcdn3.klout.com/static/images/docs/casestudies/American-Airlines-case-study.pdf

Klout. (2015b). Klout: Motorola case study. Retrieved from https://www.lithium.com/pdfs/casestudies/Lithium-Klout-Motorola-Case-Study.pdf

Klout. (2016). Retrieved from https://klout.com

Mathematica. (2016). Retrieved from https://www.wolfram.com/mathematica/

NodeXL. (2016). Retrieved from https://nodexl.codeplex.com/

Parr, B. (2010). Klout now measures your influence on facebook. Retrieved from http://mashable.com/2010/10/14/facebook-klout/#pZWqZOVQB5qo

Tunkelang, D. (2009). A twitter analog to PageRank. Retrieved from http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

Twitter. (2016). Retrieved from http://www.twitter.com

Wong, K. (2014). The explosive growth of influencer marketing and what it means for you. Retrieved from http://www.forbes.com/sites/kylewong/2014/09/10/the-explosive-growth-of-influencer-marketing-and-what-it-means-for-you

Appendix A

SAMPLE JSON OBJECT


```
{{
  "created_at": "Sun Mar 20 20:39:08 +0000 2016",
  "id": 711653320922898432,
  "id_str": "711653320922898432",
  "text": "I earned a Fitbit Adjustment of 12 calories. #LoseIt",
  "source":  "<a  href=\"http://www.loseit.com\"  rel=\"nofollow\">Lose
It!</a>",
  "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": 254282009,
    "id_str": "254282009",
    "name": "Eric N. Lott",
    "screen_name": "TheRealEricL",
    "location": "BUFFALO, NEW YORK",
    "url": null,
    "description": "I'm a poker chips & marvel comic book collecting,
movie watching, music listening, Las Vegas vacationing, positive people
loving, foodie fool",
    "protected": false,
    "verified": false,
    "followers_count": 393,
    "friends_count": 1226,
    "listed_count": 16,
    "favourites_count": 1763,
    "statuses_count": 13128,
    "created_at": "Sat Feb 19 00:25:15 +0000 2011",
    "utc_offset": -14400,
    "time_zone": "Eastern Time (US & Canada)",
    "geo_enabled": true,
    "lang": "en",
    "contributors_enabled": false,
    "is_translator": false,
    "profile_background_color": "131516",
    "profile_background_image_url":
"http://abs.twimg.com/images/themes/theme14/bg.gif",
    "profile_background_image_url_https":
"https://abs.twimg.com/images/themes/theme14/bg.gif",
    "profile_background_tile": true,
    "profile_link_color": "009999",
    "profile_sidebar_border_color": "EEEEEE",
    "profile_sidebar_fill_color": "EFEFEF",
    "profile_text_color": "333333",
```

```
    "profile_use_background_image": true,

    "profile_image_url":
"http://pbs.twimg.com/profile_images/699200679500455936/8_koJnTY_normal.j
pg",
    "profile_image_url_https":
"https://pbs.twimg.com/profile_images/699200679500455936/8_koJnTY_normal.
jpg",
    "profile_banner_url":
"https://pbs.twimg.com/profile_banners/254282009/1455537408",
    "default_profile": false,
    "default_profile_image": false,
    "following": null,
    "follow_request_sent": null,
    "notifications": null
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 0,
  "favorite_count": 0,
  "entities": {
    "hashtags":
      {
        "text": "LoseIt",
        "indices":
          45,
          52

      }
    ,
    "urls": ,
    "user_mentions": ,
    "symbols":
  },
  "favorited": false,
  "retweeted": false,
  "filter_level": "low",
  "lang": "en",
  "timestamp_ms": "1458506348244"
}}
```

Appendix B

DATA COLLECTION PROGRAM

```
using System;
using System.Configuration;
using System.IO;
using System.Messaging;
using System.Net;
using System.Runtime.Serialization.Json;
using System.Text;
using System.Threading;
using System.Web;
using System.Data.SqlClient;

 namespace TwitterStreamClient
 {
     public class TwitterStream : OAuthBase
     {
         private readonly string access_token =
ConfigurationManager.AppSettings"access_token";
         private readonly string access_token_secret =
ConfigurationManager.AppSettings"access_token_secret";
         private readonly string customer_key =
ConfigurationManager.AppSettings"customer_key";
         private readonly string customer_secret =
ConfigurationManager.AppSettings"customer_secret";

         public void Stream2Queue()
         {
             SqlConnection con = new SqlConnection("Server= localhost;
Database= twitter; Integrated Security=SSPI;");
             con.Open();
       //      string cmdString="INSERT INTO Tweets(tweetID, userID,
msgText, createdAt) VALUES (@val1, @val2, @val3, @val4)";
             string Const_TwitterDateTemplate = "ddd MMM dd HH:mm:ss
+ffff yyyy";
             string cmdString = "INSERT INTO Climate(tweetId, tweetText,
tweetCreatedAt, tweetInReplyToScreenName, tweetInReplyToUserId,
tweetInReplyToStatusId, tweetFavoriteCount, tweetRetweetCount, userId,
userScreenName, userFollowersCount, userFriendsCount, userFavoritesCount,
userStatusesCount) VALUES (@val1, @val2, @val3, @val4, @val5, @val6,
@val7, @val8, @val9, @val10, @val11, @val12, @val13, @val14)";


             //Twitter Streaming API
             string stream_url =
ConfigurationManager.AppSettings"stream_url";

             HttpWebRequest webRequest = null;
             HttpWebResponse webResponse = null;
             StreamReader responseStream = null;
```

```
            MessageQueue q = null;

            string useQueue =
ConfigurationManager.AppSettings"use_queue";
            string postparameters =
(ConfigurationManager.AppSettings"track_keywords".Length == 0 ?
string.Empty : "&track=" +
ConfigurationManager.AppSettings"track_keywords") +

//(ConfigurationManager.AppSettings"follow_userid".Length == 0 ?
string.Empty : "&follow=" +
ConfigurationManager.AppSettings"follow_userid") +

(ConfigurationManager.AppSettings"location_coord".Length == 0 ?
string.Empty : "&locations=" +
ConfigurationManager.AppSettings"location_coord");
            // postparameters =
"track=hurricanematthew&track=HurricaneMatthew&track=hurricane%20matthew&
track=Hurricane%20Matthew&locations=-122.75,36.8,-121.75,37.8";
 //            postparameters =
"track=hurricanematthew&track=HurricaneMatthew&track=hurricane%20matthew&
track=Hurricane%20Matthew&locations=-85.010359%2C25.004785%2C-
80.242292%2C30.674588";
            postparameters =
"track=climatechange%2CClimateChange%2Cclimate%20change%2CClimate%20Chang
e";
         //    postparameters = "locations=-85.010359%2C25.004785%2C-
80.242292%2C30.674588";


            // 25.004785, -85.010359
             //   30.674588, -80.242292
            if (!string.IsNullOrEmpty(postparameters))
            {
                if (postparameters.IndexOf('&') == 0)
                    postparameters = postparameters.Remove(0,
1).Replace("#", "%23");
            }

            int wait = 250;
            string jsonText = "";

            Logger logger = new Logger();


            try
            {
                //Message Queue
                if (useQueue == "true")
                {
                    if (MessageQueue.Exists(@".\private$\Twitter"))
                        q = new MessageQueue(@".\private$\Twitter");
                    else
                        q = MessageQueue.Create(@".\private$\Twitter");
                }

                while (true)
                {
```

```
                    try
                    {
                        //Connect
                        webRequest = (HttpWebRequest)
WebRequest.Create(stream_url);
                        webRequest.Timeout = -1;
                        webRequest.Headers.Add("Authorization",
GetAuthHeader(stream_url + "?" + postparameters));

                        Encoding encode = Encoding.GetEncoding("utf-8");
                        if (postparameters.Length > 0)
                        {
                            webRequest.Method = "POST";
                            webRequest.ContentType = "application/x-www-
form-urlencoded";

                            byte _twitterTrack =
encode.GetBytes(postparameters);

                            webRequest.ContentLength =
_twitterTrack.Length;
                            Stream _twitterPost =
webRequest.GetRequestStream();
                            _twitterPost.Write(_twitterTrack, 0,
_twitterTrack.Length);
                            _twitterPost.Close();
                        }

                        webResponse = (HttpWebResponse)
webRequest.GetResponse();
                        responseStream = new
StreamReader(webResponse.GetResponseStream(), encode);

                        //Read the stream.
                        while (true)
                        {
                            jsonText = responseStream.ReadLine();

                            //Success
                            wait = 250;

                            dynamic obj =
JsonUtils.JsonObject.GetDynamicJsonObject(jsonText);
                            string txt = (string)obj"text".ToLower();

                                using (SqlCommand comm = new
SqlCommand())
                                {

                                    comm.Connection = con;
                                    comm.CommandText = cmdString;

                                    DateTime createdAt =
DateTime.ParseExact((string)obj"created_at", Const_TwitterDateTemplate,
new System.Globalization.CultureInfo("en-US"));


comm.Parameters.AddWithValue("@val1", obj"id");
```

```
comm.Parameters.AddWithValue("@val2", obj"text");

comm.Parameters.AddWithValue("@val3", createdAt);

comm.Parameters.AddWithValue("@val4", obj"in_reply_to_screen_name");

comm.Parameters.AddWithValue("@val5", obj"in_reply_to_user_id");

comm.Parameters.AddWithValue("@val6", obj"in_reply_to_status_id");

comm.Parameters.AddWithValue("@val7", obj"favorite_count");

comm.Parameters.AddWithValue("@val8", obj"retweet_count");

comm.Parameters.AddWithValue("@val9", obj"user""id");

comm.Parameters.AddWithValue("@val10", obj"user""screen_name");

comm.Parameters.AddWithValue("@val11", obj"user""followers_count");

comm.Parameters.AddWithValue("@val12", obj"user""friends_count");

comm.Parameters.AddWithValue("@val13", obj"user""favorites_count");

comm.Parameters.AddWithValue("@val14", obj"user""statuses_count");

                                    comm.ExecuteNonQuery();
                                }
                            }
                            //Abort is needed or responseStream.Close() will
hang.
                            webRequest.Abort();
                            responseStream.Close();
                            responseStream = null;
                            webResponse.Close();
                            webResponse = null;
                        }
                        catch (WebException ex)
                        {
                            Console.WriteLine(ex.Message);
                            logger.append(ex.Message,
Logger.LogLevel.ERROR);
                            if (ex.Status ==
WebExceptionStatus.ProtocolError)
                            {
                                //-- From Twitter Docs --
                                //When a HTTP error (> 200) is returned,
back off exponentially.
                                //Perhaps start with a 10 second wait,
double on each subsequent failure,
                                //and finally cap the wait at 240 seconds.
                                //Exponential Backoff
                                if (wait < 10000)
                                    wait = 10000;
                                else
                                {
                                    if (wait < 240000)
                                        wait = wait*2;
```

```
                            }
                        }
                        else
                        {
                            //-- From Twitter Docs --
                            //When a network error (TCP/IP level) is
encountered, back off linearly.
                            //Perhaps start at 250 milliseconds and cap
at 16 seconds.
                            //Linear Backoff
                            if (wait < 16000)
                                wait += 250;
                        }
                    }
                    catch (Exception ex)
                    {
                        Console.WriteLine(ex.Message);
                        logger.append(ex.Message,
Logger.LogLevel.ERROR);
                    }
                    finally
                    {
                        if (webRequest != null)
                            webRequest.Abort();
                        if (responseStream != null)
                        {
                            responseStream.Close();
                            responseStream = null;
                        }

                        if (webResponse != null)
                        {
                            webResponse.Close();
                            webResponse = null;
                        }
                        Console.WriteLine("Waiting: " + wait);
                        Thread.Sleep(wait);
                    }
                }
            }
            catch (Exception ex)
            {
                Console.WriteLine(ex.Message);
                logger.append(ex.Message, Logger.LogLevel.ERROR);
                Console.WriteLine("Waiting: " + wait);
                Thread.Sleep(wait);
            }
        }


        public void QueueRead()
        {
            MessageQueue q;
            string multiThread =
ConfigurationManager.AppSettings"multithread";
            Logger logger = new Logger();

            try
```

```
            {
                if (MessageQueue.Exists(@".\private$\Twitter"))
                    q = new MessageQueue(@".\private$\Twitter");
                else
                {
                    Console.WriteLine("Queue does not exists.");
                    return;
                }

                while (true)
                {
                    Message message;
                    try
                    {
                        message = q.Receive();
                        message.Formatter =
                            new XmlMessageFormatter(new
{"System.String"});
                        if (multiThread == "true")
                            ThreadPool.QueueUserWorkItem(MessageProcess,
message);
                        else
                            MessageProcess(message);
                    }
                    catch
                    {
                    }
                }
            }
            catch (Exception ex)
            {
                Console.WriteLine(ex.Message);
                logger.append(ex.Message, Logger.LogLevel.ERROR);
            }
        }


        public void MessageProcess(object objMessage)
        {
            status status = new status();
            Logger logger = new Logger();
            DataContractJsonSerializer json = new
DataContractJsonSerializer(status.GetType());

            try
            {
                Message message = objMessage as Message;

                byte byteArray =
Encoding.UTF8.GetBytes(message.Body.ToString());
                MemoryStream stream = new MemoryStream(byteArray);

                //TODO:  Check for multiple objects.
                status = json.ReadObject(stream) as status;

                Console.WriteLine(message.Body.ToString());

                //TODO: Store the status object
```

```
                DataStore.Add(status);
            }
            catch (Exception ex)
            {
                Console.WriteLine(ex.Message);
                logger.append(ex.Message, Logger.LogLevel.ERROR);
            }
        }

        private string GetAuthHeader(string url)
        {
            string normalizedString;
            string normalizeUrl;
            string timeStamp = GenerateTimeStamp();
            string nonce = GenerateNonce();


            string oauthSignature = GenerateSignature(new Uri(url),
customer_key, customer_secret, access_token, access_token_secret, "POST",
timeStamp, nonce, out normalizeUrl, out normalizedString);


            // create the request header
            const string headerFormat = "OAuth oauth_nonce=\"{0}\",
oauth_signature_method=\"{1}\", " +
                                        "oauth_timestamp=\"{2}\",
oauth_consumer_key=\"{3}\", " +
                                        "oauth_token=\"{4}\",
oauth_signature=\"{5}\", " +
                                        "oauth_version=\"{6}\"";

            return string.Format(headerFormat,
                Uri.EscapeDataString(nonce),
                Uri.EscapeDataString(Hmacsha1SignatureType),
                Uri.EscapeDataString(timeStamp),
                Uri.EscapeDataString(customer_key),
                Uri.EscapeDataString(access_token),
                Uri.EscapeDataString(oauthSignature),
                Uri.EscapeDataString(OAuthVersion));
        }
    }
}
```

Appendix C

SCREEN SCRAPING PROGRAM

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Data.SqlClient;
using System.Data;
using System.Threading;
 namespace ScreenScrape
 {
     class Program
     {
         static void Main(string args)
         {
             int progress = 0;
             int retweetCount;
             int likesCount;
             String tweetID ="";
             String screenName ="";
             SqlConnection connection = new SqlConnection("Server=
localhost; Database= twitter; Integrated Security=SSPI;");
             SqlConnection connection2 = new SqlConnection("Server=
localhost; Database= twitter; Integrated Security=SSPI;");

             SqlCommand sqlStatement = new SqlCommand("select tweetid,
userscreenname from matthew where tweetcreatedat >= '2016-10-12
16:27:37.000' and tweetcreatedat < '2016-10-14 04:54:08.000' order by
tweetcreatedat asc;", connection);
             SqlCommand sqlUpdate = new SqlCommand("update matthew set
tweetFavoriteCount=@val1, tweetretweetcount=@val2 where tweetid=@val3;",
connection2);
             connection.Open();
             SqlDataReader dr = sqlStatement.ExecuteReader();

             if (dr.HasRows)
             {
                 connection2.Open();
                 while (dr.Read())
                 {
                     tweetID = dr0.ToString();
                     screenName = dr1.ToString();

                     Console.WriteLine(progress++);

                     try
                     {

       string source = ScreenScrape("https://twitter.com/" + screenName +
"/status/" + tweetID);
```

```
                        retweetCount = getRetweetCount(source);
                        likesCount = getLikesCount(source);

                        if (retweetCount != 0 || likesCount != 0)
                        {
                            sqlUpdate.Parameters.AddWithValue("@val1",
likesCount);
                            sqlUpdate.Parameters.AddWithValue("@val2",
retweetCount);
                            sqlUpdate.Parameters.AddWithValue("@val3",
tweetID);

                            sqlUpdate.ExecuteNonQuery();
                            sqlUpdate.Parameters.Clear();
                        }
                    }
                    catch (Exception ex)
                    {
                    }
                }
                connection2.Close();
            }
                connection.Close();
        }
         public static string ScreenScrape(string url)
         {
                return new System.Net.WebClient().DownloadString(url);

         }
         public static int getRetweetCount(string source)
         {
             int retweetCount = 0;
             int retweetStart = source.IndexOf("Retweets <strong>") + 17;
             int retweetEnd = source.IndexOf("</strong>", retweetStart);

             string retweetSub = source.Substring(retweetStart,
retweetEnd - retweetStart);
             if (retweetSub != null && retweetSub != "" &&
int.TryParse(retweetSub, out retweetCount)) ;

             return retweetCount;
         }a

         public static int getLikesCount(string source)
         {
             int likesCount = 0;
             int likesStart = source.IndexOf("Likes <strong>") + 14;
             int likesEnd = source.IndexOf("</strong>", likesStart);
             string likesSub = source.Substring(likesStart, likesEnd -
likesStart);
             if (likesSub != null && likesSub != "" &&
int.TryParse(likesSub, out likesCount)) ;

             return likesCount;
                              }     }  }
```

VITA

Doaa H. Gamal has a Bachelor of Science degree from Cairo University in Mass Communication, 1992 and expects to receive a Master of Science in Computer and Information Sciences from the University of North Florida, April 2017. Dr. Kathikeyan Umapathy of the University of North Florida is serving as Doaa's thesis advisor. Dr. Lakshmi Goel and Dr. Sandeep Reddivaari are members of Doaa's thesis committee.

Doaa is currently employed as Data Warehouse Analyst at JEA and has been with the company for 4 years. Prior to that, Doaa worked ABB in Cairo, Egypt. Doaa has on-going interests in Big Data, Data Analytics and Database Systems and has extensive experience with Oracle's Business Analytics suite.

Doaa is married for 20 years with three children, Honya, Farrah and Omar. Doaa and her family have lived in Jacksonville, Florida since 2000.