2015

A Statistical Analysis and Modeling of Information Diffusion Across Online Social Networks

Michael Piserchia
University of Rhode Island, mvpiserchia@gmail.com

Follow this and additional works at: <https://digitalcommons.uri.edu/theses>

Recommended Citation

Piserchia, Michael, "A Statistical Analysis and Modeling of Information Diffusion Across Online Social Networks" (2015). *Open Access Master's Theses*. Paper 636.
<https://digitalcommons.uri.edu/theses/636>

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

A STATISTICAL ANALYSIS AND MODELING OF INFORMATION
DIFFUSION ACROSS ONLINE SOCIAL NETWORKS

BY

MICHAEL PISERCHIA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
STATISTICS

UNIVERSITY OF RHODE ISLAND

2015

MASTER OF SCIENCE THESIS
OF
MICHAEL PISERCHIA

APPROVED:

Thesis Committee:

Major Professor Natallia Katenka

Liliana Gonzalez

Corey Lang

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2015

ABSTRACT

In recent years, online social networks have become a very popular and effective forum for information exchange. These large, highly interconnected networks span the globe and have the ability to disseminate information in a fraction of the time it would take other communication networks. Given the myriad ways in which online social networks can be used, creating accurate, predictive models for the spread of information across them is very valuable. With that, modeling processes on large networks is a difficult task. It is computationally expensive, and usually prohibitive, to model a process on the entirety of a very large network. Given these complexities, creating smaller network graphs that are characteristically similar to the original networks graphs enable researchers to run models that are otherwise not feasible.

This project aims to create prototypic networks and model the spread of information across them using network-based epidemiological models to better understand how information spreads across an online social network. More specifically, the focus will be on the spread of the news of a scientific discovery, i.e., the Higgs boson particle, on Twitter.

ACKNOWLEDGMENTS

I would like to thank the Computer Science and Statistics Department for affording me the wonderful opportunity to pursue graduate studies as a statistics student. The support, both financial and academic, has been incredible and has made for a wonderful experience. I have learned so much over the last two years and it is truly a testament to the enormous amounts of time and energy the faculty invests in the success of the students. I would like to thank Professor Katenka, Professor Gonzalez, Professor Puggioni, Professor Robinson, Professor Lang, Professor Finizio, Beth Larimer, and Lorraine Berube for their integral part in my success. I would like to thank all of the other students in the department that have befriended and helped me over the last two years. I would also like to thank Professor Katenka for being my thesis advisor and helping me turn a few scatterbrained ideas and R code into a coherent thesis.

I would like to thank my mom for doing an amazing job of raising me. I mention this because all of my accomplishments are largely attributable to the unwavering love and support she has provided throughout my entire life. I would also like to thank my mother for showing me the beauty and value of science and mathematics. Lastly, I would also like to thank my entire family for their support.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 Introduction	1
1.1 Overview of Online Social Networks	2
1.2 Introduction to Information Diffusion and Epidemic Modeling	5
1.3 Case Study: News of the Higgs-Boson Discovery on Twitter	6
List of References	10
2 Review of Literature	11
2.1 Quantitative Analysis of Online Social Networks	11
2.2 Sampling of Network Graphs and Scalability	12
2.3 Network Processes: Information Diffusion and Epidemic Modeling	14
List of References	17
3 Methodology	20
3.1 Descriptive Analysis and Network Characterization	20
3.2 Social Network Sampling Algorithms	25
3.3 Information Diffusion and Epidemic Modeling	31

	Page
List of References	35
4 Results	36
4.1 Topology Characterization of Social Network	36
4.2 Social Network Sampling and Inference	42
4.3 Modeling the News of the Higgs-Boson Discovery on Twitter . .	58
List of References	64
5 Conclusion	65
5.1 Main Findings	65
5.2 Future Work	66
List of References	68
BIBLIOGRAPHY	69

LIST OF TABLES

Table		Page
1	Network Characteristic Summaries Including: Number of nodes, global clustering coefficients, average node degree, degree assortativity, and number of connected components for the observed network and the largest connected component (L.C.C.).	38
2	Induced Subgraph Characteristics. Estimates obtained from averaging the graph characteristics of 100 induced subgraphs sampled from the original social network.	45
3	Incident Subgraph Characteristics. Incident Subgraph Characteristics. Estimates obtained from averaging the graph characteristics of 100 incident subgraphs sampled from the original social network.	46
4	Snowball Sampling Characteristics. Snowball Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 snowball sampled from the original social network.	48
5	Power-Law Sampling Characteristics. Power-law Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 power-law graphs.	50
6	Barabasi-Albert Sampling Characteristics. Estimates obtained from averaging the graph characteristics of 100 Barabasi-Albert preferential attachment graphs.	51
7	Metropolis-Hastings Random Walk Sampling Characteristics. Metropolis Hastings random Walk Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 Metropolis Hastings random walk graphs.	52
8	Sensitivity C.I.s	53
9	Sensitivity Analysis	54

LIST OF FIGURES

Figure		Page
1	Twitter Wall with Retweet Option Selected. Courtesy of Twitter.com.	3
2	Undirected Friendship Graph with 7 elements and 8 links. . . .	5
3	Degree Distribution.	39
4	Hill Plot. of Power-law exponent, $\hat{\alpha}_k$, as a function of number of order statistics used, k . Stabilizes at $k \sim 40$ with a value of $\hat{\alpha}_k = 2.4$	39
5	Local Clustering Coefficient Histogram (Average: .147).	40
6	Local Density Histogram (Average: .221).	41
7	Vertex Degree Versus Average Neighbor Degree.	42
8	Induced Subgraph Sampling. Left Panel: Visualization of Induced Subgraph Sampling with $N_V = 2000$ and $N_E = 297$. Right Panel: Degree Distribution of Induced Subgraph.	44
9	Incident Subgraph Sampling. Left Panel: Visualization of Incident Subgraph Sampling with $N_V = 3689$ and $N_E = 2000$. Right Panel: Degree Distribution of Incident Subgraph.	46
10	Snowball Sampling. Left Panel: Visualization of Snowball Sampling with $N_V = 1412$ and $N_E = 14894$. Right Panel: Degree Distribution of Snowball Sampled Subgraph.	48
11	Power-Law Distribution Sampling, $P(k) = k^{-2.4}$. Left: Visualization of Power-Law Sampling with $N_V = 2000$ and $N_E = 2016$. Right: Degree Distribution of Power-law graph.	49
12	Barabasi-Albert Sampling Method. Left: Visualization of Barabasi-Albert preferential attachment graphs. Sampling with $N_V = 2000$ and $N_E = 1999$. Right: Degree Distribution of Barabasi-Albert preferential attachment graphs.	51

Figure		Page
13	Metropolis-Hastings Random Walk Sampling. Left: Visualization of Metropolis Hastings random walk graph. Sampling with $N_V = 1412$ and $N_E = 14894$. Right: Degree Distribution of Metropolis Hastings random walk graph.	52
14	Normalized mean square error (NMSE) of sample graph degree distributions as a function of degree index, k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a degree less than or equal to k and the proportion of nodes in the original graph with a degree less than or equal to k	56
15	Normalized mean square error (NMSE) of sample graph local clustering coefficient distributions as a function of all possible clustering coefficient values, i.e., zero to one, and denoted as k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a clustering coefficient less than or equal to k and the proportion of nodes in the original graph with a clustering coefficient less than or equal to k	57
16	Normalized mean square error (NMSE) of sample graph local density distributions as a function of all possible density values, i.e., zero to one, and denoted as k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a density less than or equal to k and the proportion of nodes in the original graph with a density less than or equal to k	57
17	Frequency of Twitter actions as a function of day. Three Twitter actions, Mentioning, Replying, and Retweeting are plotted as functions of time. RT stands for Retweeting (Top Panel), RE stands for replying(Middle Panel), and MT stands for mentioning(Bottom Panel).	59
18	CDF of Twitter actions as a function of day. Three Twitter actions, Mentioning, Replying, and Retweeting are plotted as functions of time. RT is retweeting, RE is replying, and MT is mentioning.	60
19	Proportion of Users Discovering The Higgs-Boson News Over Time as a Function of Time in Days.	61

Figure		Page
20	RMSE of SI models as a function of the transmission rate, β , for three graphs generated from sampling techniques. The minimum of each line represents the transmission rate at which a SI model on a particular graph best fits the actual proportion of Twitter users becoming aware of the Higgs boson news (Figure 19) The MHRW RMSE has the smallest minimum value of the three RMSE plots and has the best fit to the actual proportion of users becoming aware over time. MHRW Minimum at $\beta = .056$.	62
21	SI epidemic processes run across three of the sampled graph types. The black line is the actual information diffusion process across the entire Twitter network. This network is used to compute the mean square error for each process by comparing the proportion of informed users at each timestamp. The processes on the Barabasi-Albert and power-law models have a transmission rate of 1. The transmission rate for the process on the MHRW graph is .056.	62
22	Total number of iterations for epidemic processes run across the Barabasi-Albert and power-law graphs with a $\beta = .4$. The number of iterations visualized was extended from 7 to the total number required to completely reach the entire graph to illustrate that the graphs still conform to a shape similar to the MHRW and actual information diffusion curves (a sigmoid curve), but it takes more iterations to complete the process.	64

CHAPTER 1

Introduction

In recent years, online social networks have become a very popular and effective forum for information exchange. These large networks span the globe and have the ability to disseminate information in a fraction of the time it would take other communication networks. Given the myriad ways in which online social networks can be used, creating accurate, predictive models for the spread of information across them is very valuable. Modeling processes, however, on large networks is a difficult task. It is computationally expensive, and usually prohibitive, to model a process on the entirety of a very large network with most of the currently available software packages. This project aims to create prototypic networks and model the spread of information across them using network-based epidemiological models to better understand how information diffusion across an online social network. More specifically, the focus will be on the spread of the news of a scientific discovery, i.e., the Higgs boson, on Twitter.

The discovery of the Higgs boson was one of the greatest scientific achievements in modern science. The elusive nature of the particle and its hypothesized ability to unify the three major forces into what is known as the Grand Unified Theory made the discovery of this particle a very momentous achievement for the scientific community [1]. The particle and its importance had also reached the general public and had been dubbed the God Particle because of its potential ability to shed light on the framework of the universe [1]. News of such a momentous discovery was of a particular interest to the scientific community and the general public.

A discovery of this magnitude had not been made in the recent past, and had

certainly not been made since the explosion of popularity of online social networks. Given Twitter's role as a global, online news outlet [2], tracking the spread of such momentous news across Twitter offers a very interesting look at how significant events spread across social communities in the era of online social networks.

1.1 Overview of Online Social Networks

Online social networks began to form in the 1990s when the Internet started to become available to households and individual users [3]. These networks offered very limited functionality, e.g., forums, chatrooms, and instant messaging. However, social media websites grew much more popular at the turn of the 21st century when a website, named Friendster, was launched. Friendster attempted to mimic real world friendship communities by creating connections between users that had a common interest or considered to be friends. In the years to follow, sites such as LinkedIn, Myspace, Facebook, adopted the idea of organizing an online social networks that revolved around communities formed by shared interests and friendship.

The focus of this thesis is on another popular online social media website called Twitter. Twitter was created in 2006 and offered a slightly different social experience than the preceding social networks. Twitter allows users to post short messages, referred to as 'tweets', on their own profile wall. A user's wall is a space on their profile that is used to share various types of tweets. Users can share personal experiences, daily activities, momentous events, etc.

Twitter users form connections with other users in a way that is slightly different than other social networks such as Facebook or LinkedIn. The difference between the types of connections made on Twitter when compared to other social networks is that rather than being friends with another user, a Twitter user has the option of following another user. Following another Twitter user enables one

to view all of the postings made to their wall. This unidirectional relationship does not guarantee that the user being followed will see any of the content posted to the wall of the person doing the following. This relationship between users differs from other social networks because most other social networks establish bidirectional connections between users that are considered friends. The bidirectional connections require a mutual agreement of reciprocal information sharing, as opposed to the unidirectional structure that is used on Twitter. This structure of social relationships is atypical when one think about social networks because friendships is usually thought of as a reciprocal relationship. So, the idea of sharing information with ones friends must be altered when thinking about information exchange on Twitter.

Figure 1 displays a Twitter wall on which all of a user’s follower’s messages, or tweets, appear. Hashtags are frequently used to highlight key points of the tweet.

Figure 1: Twitter Wall with Retweet Option Selected. Courtesy of Twitter.com.

The Twitter wall of each user represents a "news feed", an aggregate of the information posted by all the of the users they are following on Twitter. The news feed is where each user interacts with the Twitter community. It is where each user gains access to all of the information being offered by the friends, news agencies, celebrities, etc., which they are following. Users also have the ability to perform a number of actions on the tweets of others. The three most popular actions

are retweeting, mentioning, and replying. Retweeting is the process of reposting another users message onto your own wall. This, in turn, offers that persons piece of information to all of the followers of the retweeter. Mentioning another user in one of their own tweets is an action that allows alerting users to the information contained in that particular post. Replying is a simple action that allows a user to reply to a given tweet. The replies to tweets are seen at the bottom of the tweet and resemble web forum discussions.

Until now, networks in this thesis have been described in a more relaxed manner with an emphasis on describing social networks in the real world. However, along with this applied notion, a more formal body of knowledge will also be required to fully understand social networks. The branch of mathematics known as graph theory provides the formal understanding necessary to analyze networks. Graph theory is an integral part of network analysis because networks are most commonly expressed and studied as mathematical graphs (see Chapter 3 for more details). Generally, networks represent a system of interacting elements, e.g., Twitter users follow each other. A network graph represents a set of predefined nodes (elements) and edges (links).

Figure 2 illustrates a hypothetical set of seven elements, for example, friends on a social network, and links connecting particular elements (friends). This grouping of people and the relationships between them are a very basic example of a graph.

Figure 2: Undirected Friendship Graph with 7 elements and 8 links.

1.2 Introduction to Information Diffusion and Epidemic Modeling

Information diffusion can be regarded as a dynamic process that takes place on a social network and can be thought of as the evolution of time-indexed vertex attributes on a given network [4]. For example, the spread of knowledge or a rumor through a population can be thought of as a dynamic, time indexed process in which the piece of information reaches a certain number of people at a given time period.

The information diffusion models are primarily derived from classical epidemiology. There are, however, differences between the spread of disease via physical contact and the spread of information online.

The spread of the disease via physical contact relies on geographic proximity of different individuals and usually the transmission of the disease is unintentional. Also, infected individuals may be unaware of their infection. In online social networks, information diffuses from one user to a follower or to individuals within his community. Also, information is transmitted intentionally. Physical contact and geographical location have much less bearing on the probability of whether not information will be transmitted from one user to another. Online social networks span the globe and, as a result, a Twitter user on one continent can potentially become infected with the information of the news story from a Twitter user that is on a different continent.

An important difference between Twitter and other social networks is unidirectionality, meaning that the potential for information diffusion can only move in one direction. For example, Person A may be following person B on twitter, and there will be a connection on the network, but it doesn't mean that Person A can spread the information to Person B. This notion of one-way information flows leads to interesting phenomena in epidemic spreads regarding the rates of infection [5].

In what follows in Chapter 3 and Chapter 4, network-based epidemic models similar to those discussed in [6] for computer viruses will be used to model the scientific discovery of the Higgs boson particle on Twitter. Epidemic modeling will be used to model the spread of information over time. Classical epidemiological models make the assumption that the population is homogenous and moving randomly throughout a given space. This assumption of modeling across a random network graph cannot be assumed when modeling the spread of information across Twitter (for reasons described earlier). Given that, network-based epidemiological models will be used instead of classical epidemiological models. With these network-based models, the paper aims to accurately predict the spread of information across a social network.

1.3 Case Study: News of the Higgs-Boson Discovery on Twitter

In what follows, the structure of the Twitter network graph as well as the data related to the news of the Higgs boson discovery on Twitter will be introduced. The nodes of the Twitter graphs represent individual users that possess accounts on Twitter. The connections, or edges, that comprise the network graph indicate that one account is following another account. Following allows one user to track the postings or comments of another user. If account A follows account B, the postings of B will appear on the newsfeed of A. Therefore, if B posts a particular news story or makes a reference to a particular event, it is likely that this information will

appear on the newsfeed of account A.

As mentioned earlier, there are three main actions that users can perform on Twitter including retweeting, mentioning, and replying. 'Retweeting is an action that can be performed by users who want to post someone else's postings on their own page. If account A enjoyed account B's comment or news story, he can retweet it and have it appear on his page. Mentioning and replying are two other features that are more self-explanatory than the previous features. If someone wants to mention another user in a post of theirs, they use the @ sign followed by the other account holder's name, and they will be notified of the posting. Replying is a chat-like feature that allows people to comment on any post they feel deserves a comment.

The connections, or edges, that comprise the network graph on Twitter create a unidirectional relationship that allows information to be transmitted in only one direction (e.g., from the followee to the follower). This type of connection implies that Twitter should be represented as a directed graph.

Conversely, online social networks like Facebook have bidirectional relationships between friends because information and communication can flow in both directions (from one friend to another, or vice versa).

Following on Twitter allows one user to track the postings or comments of another user. If account A follows account B, the postings of B will appear on the newsfeed of A. Therefore, if B posts a particular news story or makes a reference to a particular event, it is likely that this information will appear on the newsfeed of account A. Retweeting is an action that can be performed by users who want to post someone else's postings on their own page. If account A enjoyed account B's comment or news story, he can retweet it and have it appear on his page. Mentioning and replying are other features that are a little more self-explanatory

than the previous features. If someone wants to mention another user in a post of theirs, they use the @ sign followed by the other account holders name, and they will be notified of the posting. Replying is a chat-like feature that allows people to comment on any post they feel deserves a comment.

The social network of followers will serve as the underlying network graph on which the processes of information diffusion will be modeled. Although all three actions can contribute to the information diffusion on Twitter, only retweet and mentioning actions will be modeled as dynamic processes via network-based epidemic models.

The data used in this thesis was originally collected by De Domenico et al. [1] using the Twitter API service and can be found on the Stanford Network Analysis Project website [7]. The dataset contains four large network graphs that contain interactions related to the spread of the news about the Higgs boson. There exist social, retweet, mention, and reply networks. The social network is the largest one, containing roughly 450,000 nodes with 14,000,000 edges, and represents the following activity of all the users engaged in the spread of this news story. The retweet network, 425,000 nodes and 733,00 edges, represents all the retweets done during the spreading period. The mention and reply networks are significantly smaller and represent chatting done between different accounts holders.

Information diffusion processes depend on node connectivity and clustering that comprise the main characteristics of the underlying social network structure. In the beginning of Chapter 4, the following network characteristics will be analyzed: the number of connections each node has to the rest of the graph, the clustering of nodes, the cohesion of nodes, and the correlation between the number of links an element possess and the number of links its neighbors possess.

The social network of Twitter followers is similar to many other online social

networks in that its distribution of connections, or edges, closely follows an exponential distribution [8]. This means that the overwhelmingly majority of users have between zero and one hundred connections to other users, but there are a small percentage of users that have a very large number, upwards of 50,000, of connections to other users. This type of graph will have certain areas of very high clustering around these accounts that have 50,000 followers. The clustering coefficient is an important statistic. This measure of how closely nodes cluster is of importance because rates of information diffusion vary greatly depending how how connected a given node is to the surrounding nodes.

The network of Twitter users described previously is very large, both in number of nodes and edges. The size of the network makes it difficult, and often prohibitive, to model processes on it using available software packages. Sampling from this large network to obtain scaled down, but still representative, networks is an important first step taken before simulating information diffusion processes. Different sampling methods provide different networks and network structure characteristics. The network characteristics of the graphs created from each sampling technique are compared to the characteristics of the original network to determine which method proceeds the most similar scaled down, tractable network (see Chapters 3.2 and 4.2 for more information).

The remaining content of the thesis is organized as follows. Chapter 2 is dedicated to a review of the existing literature related to three main areas: quantitative analysis of online social networks, scalability of networks graphs, and information diffusion processes. Chapter 3 provides background information and mathematical definitions of network analysis characteristics, network sampling methods, and epidemic models for information diffusion. Main results and findings are summarized in Chapter 4, and followed by conclusions in Chapter 5.

List of References

- [1] M. D. Domenico, A. Lima, *et al.*, “Anatomy of a scientific rumor,” *Scientific Reports*, no. 4.
- [2] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 591–600.
- [3] “The history of social networking,” <http://www.digitaltrends.com/features/the-history-of-social-networking/>, accessed: 2015-06-24.
- [4] Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY, United States of America: Springer, 2009.
- [5] J. B. M Newman, S Forrest, “Email networks and the spread of computer viruses,” *Physical Review*, 66, 2002.
- [6] R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Physics Review Letters*, vol. 86, no. 14, april 2001.
- [7] “Stanford network analysis project,” <http://snap.stanford.edu/data/index.html>, accessed: 2014-09-12.
- [8] F. Fu, L. Liu, and L. Wang, “Empirical analysis of online social networks in the age of web 2.0,” *Physica A*, vol. 387, pp. 676–682, 2008.

CHAPTER 2

Review of Literature

2.1 Quantitative Analysis of Online Social Networks

Social scientists have studied the relationships and connections that people forge with one another in a network based framework since the 1930s [1]. These early social networks refer to a collection of entities that are not interacting on the Internet.

Social networks exists in many forms and include, but are not limited to, friendships among people, co-authorship on journal publications, business alliances between certain companies, ally networks between nations [1]. All of these social networks possess entities that are linked together and interaction between entities are dependent on links between the entities. Similarly, online social networks are complex systems in which interactions between entities (e.g., blog owners, Twitter users, etc.) take place on the Internet.

Social networks have been shown to exhibit properties of a nonrandom structure of links between the entities. The nonrandom structure of social networks is typically classified into one of two major categories. The first, small world networks, have been shown to be more typical of offline social networks [2]. Small world networks have have a structure that requires a small number of links to be traversed to reach any of the entities on the graph. This characteristic of a social network of friendships among people has been popularized by the idea of the small-world phenomenon, which was first studied by Milgram in [3]. This theory posits that if a one person is randomly selected, it would take a small number of friendship-link traversals, six is a popular, but controversial estimate, to reach any other person in the world. While this accuracy of this idea is controversial, it has been shown the average number of link traversals between two americans is six [4].

The second type, power law networks, are much more common among online social networks [2]. Power law networks are networks in which a large percentage of the entities have a small number of links to other entities and a small percentage of the entities have a very large number of links to other entities. Information networks, such as Twitter, have been shown to exhibit power law characteristics. It has also been shown that Internet networks such as blogging services often exhibit properties of power law networks [1]. The power-law networks also typically possess a very large connected component of entities that encompasses nearly the entire number of entities [2]. It will be illustrated in Chapter 4.1 that the Twitter network being analyzed also has a very large connected component that accounts for nearly all of the entities and links of the entire graph.

The increasing popularity of online social networks has led to a rapid increase in a network size. The popularity of online social networks has also increased the number of online social networks. This increase in size and frequency indicates that these networks are becoming increasingly more integrated into daily life, while also getting larger and more complex. Given this, the value of analyzing these networks is increasing, but the task of analyzing them is becoming increasingly difficult due to the very large size of the networks. Sampling methods are one way to address this problem.

2.2 Sampling of Network Graphs and Scalability

The impetus for sampling online social networks is motivated by the sudden explosion in popularity of these networks. This very rapid increase in popularity has dramatically increased the size of these networks and have made the networks intractable in their natural state. Thus, scaling down the massive online social networks using sampling methods is required before conducting any further inferential analyses [5].

Two popular sampling techniques used to create sampled networks are element-based and link-based. Element-based sampling techniques involve randomly selecting a subset of elements from the whole graph and constructing links between the subset of nodes based on a certain set of conditions [1]. Link-based sampling involves randomly sampling a set of edges and creating a network from the subset of edges [1]. The goal of sampling in this thesis, regardless of the technique employed, is to obtain a scaled down network that possesses link and element features that are similar to those of the original network.

Power law networks possess a small number of elements that have a very large number of connections to other elements in the graph. This characteristic is very different from random graphs, which possess a degree distribution which closely follows a normal distribution around a most common degree value. Given this, it is unsurprising that a large number of sampling techniques that apply to random graphs will fall short of preserving the characteristics of a very large scale free network. If nodes are randomly selected, it is very unlikely that the very high degree nodes will be selected. Furthermore, if the nodes are selected based on a likelihood that is proportional to their degree, it is even more unlikely that they will be selected. Given that these high-degree nodes are crucial when analyzing the topology of a power law network, it is easy to see how some of the more traditional sampling techniques mentioned below can lead to very misleading inference about the topology or processes of a given network.

It has been shown in [6] that subgraphs of a scale free network (a form of power-law network) that are obtained through random sampling do not produce a network that has an exponential degree distribution. Common sampling methods such as the induced subgraph sampling (a form of element-based sampling), incident subgraph sampling (a form of link-based sampling), snowball sampling, and many

others sampling techniques have been shown to be unable to preserve the degree distribution of a scale free network [7].

With that, it has been shown that random walk algorithms are better suited to sample representatively from a power law network [7]. Random Walk algorithms employ a Bayesian sampling method known as a Metropolis Hastings algorithm to crawl along the graph from vertex to another [8]. This Monte Carlo Markov Chain (MCMC) approach has been proven to be effective at preserving the characteristics of directed online social networks [9]. In fact, it has been shown that the Markov Chain of a Metropolis-Hastings Random Walk can be tailored to any vertex distribution [10].

The sampling designs discussed in this section will be considered as appropriate sampling techniques and described in detail in Chapter 3, and performed on the Twitter graph in Chapter 4.

2.3 Network Processes: Information Diffusion and Epidemic Modeling

The diffusion of information across online networks has been analyzed since the advent of the Internet. The idea of information cascading in a systematic manner across an online network was applied to computer viruses in the early nineties [11]. The result of this analysis showed that epidemiological models could effectively be applied to modeling information spread on a directed computer network [11]. It also showed that network topology greatly influenced the scope of computer virus propagation through a network of computers [11]. These early types of information diffusion model have been altered over the years to model the spread of information across online social networks.

News of important events and rumors have been described as infections of the mind, and this analogy to ideas and information acting as infectious pathogens has given rise to a large body of research devoted to modeling the spread of information

across social networks using epidemic modeling techniques [12].

This research led to numerous discoveries regarding the benefits and downsides of applying epidemic models to the spread of information [12]. It has been shown that network topology not only impacts the scope of the spread, but also the transmission rate required for an epidemic to spread [12]. Traditional epidemic models rely on an assumption that the people or entities being studied comprise a random network. This assumption of a random network means that all uninfected people are equally likely to be infected by a randomly selected infected person. As a network, this would mean that every person, or element, has a link to every other person in the network. As discussed earlier in this chapter, online social networks are rarely random and, as a result, the epidemic processes modeled on them exhibit characteristics that were previously thought to be very unlikely to occur.

Traditional epidemic models rely heavily on the value of the reproductive ratio, R_0 , which represents the average number of people a given person will infect over the course of the epidemic [13]. If the reproductive ratio is less than 1, then the epidemic will die out over time. Conversely, if the reproductive ratio is greater than 1, the epidemic will continue to spread until all elements have become infected [13]. This methodology of gauging the severity and scope of an epidemic based on R_0 only works for random networks. As one might expect, R_0 is not an effective gauge of epidemic spreading when dealing with nonrandom networks. In fact, it has been shown that epidemics can spread to exponentially in scale-free networks regardless of reproductive ratio and even transmission rate [14]. This result is due to the fact that a typical node is likely connected, or a short path away from, a highly connected node [15]. The existence of the highly connected nodes enable the epidemic to continue to spread even when the reproductive ratio and transmission

rates are still low [15]. Unsurprisingly, it has been shown that networks possessing exponential degree distributions are particularly susceptible to epidemics [16] .

Twitter is one of the few online social networks that has been proven to act more as a information dissemination service as opposed to to true social network with required reciprocal connections between users [17]. Furthermore, a topological analysis of the degree distributions shows low reciprocity of following [17]. This understanding of Twitter as an information diffusion medium has been used to construct various information diffusion models.

Current Twitter information diffusion models mostly fall into two broad categories. The first being a method of analyzing only the retweet, i.e. spreading, network and its topological characteristics over time [18]. The second method involves analyzing the retweet network in conjunction with underlying social network (i.e., A Twitter network of followers) [19].

The first method is primarily used to understand how cascading trees evolve over the course of a give time period. Cascading trees are directed networks that originate at a root node, an individual Twitter user in this case, and extend as far that individuals influence will reach [19]. This method of analyzing cascading sequences is often used when comparing the influence of one user with another [20]. This type of analysis is also used to study how far across the network the average users influence will go over a specified time period [18]. It is important to note that this type of analysis does not presume a specific social network structure prior to analyzing the retweet network.

The second method involves modeling the spread of information after characterizing the original social network. This type of analysis often includes epidemic modeling because one is aware of the network structure and the number of individuals present in a given population before creating the model. These analyses

have shown that epidemic models can accurately model the spread of news and rumors on a network of Twitter users [21]. The ability to model information diffusion using epidemic models is, in part, made possible by the potential for a rapid spreading of information on Twitter [22]. Epidemics are often intense and episodic, leading to a sharp spike in the number of infected individuals followed by a sharp decline [13]. This short-lived, intense activity on Twitter will be explored in more detail throughout Chapter 4.

List of References

- [1] Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY, United States of America: Springer, 2009.
- [2] F. Fu, L. Liu, and L. Wang, “Empirical analysis of online social networks in the age of web 2.0,” *Physica A*, vol. 387, pp. 676–682, 2008.
- [3] J. Travers, S. Milgram, J. Travers, and S. Milgram, “An experimental study of the small world problem,” *Sociometry*, vol. 32, pp. 425–443, 1969.
- [4] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC ’07. New York, NY, USA: ACM, 2007, pp. 29–42.
- [5] T. Wang, Y. Chen, Z. Zhang, P. Sun, B. Deng, and X. Li, “Unbiased sampling in directed social graph,” *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 401–402, Aug. 2010.
- [6] M. P. H. Stumpf, C. Wiuf, and R. M. May, “Subnets of scale-free networks are not scale-free: Sampling properties of networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4221–4224, 2005.
- [7] P. Ebbes, H. Thadakamalla, *et al.*, “Sampling large-scale networks: Insights from simulated networks.”
- [8] T. Wang, Y. Chen, and others”, “Understanding graph sampling algorithms for social network analysis,” *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference*, pp. 123–128, 2011.
- [9] B. F. Ribeiro, P. Wang, F. Murai, and D. Towsley, “Sampling directed graphs with random walks.” pp. 1692–1700, 2012.

- [10] P. Hu and W. C. Lau, “A survey and taxonomy of graph sampling,” *CoRR*, vol. abs/1308.5865, 2013.
- [11] J. O. Kephart and S. R. White, “Directed-graph epidemiological models of computer viruses.” in *IEEE Symposium on Security and Privacy*, 1991, pp. 343–361.
- [12] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, “Theory of rumour spreading in complex social networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457 – 470, 2007.
- [13] F. Bauer, P. van den Driessche, and J. Wu, *Mathematical Epidemiology*. New York, NY, United States of America: Springer, 2008.
- [14] R. Pastor-Satorras and A. Vespignani, “Epidemic spreading in scale-free networks,” *Physics Review Letters*, vol. 86, no. 14, april 2001.
- [15] S. Goel, A. Anderson, J. Hofman, and D. Watts, “The structural virality of online diffusion.”
- [16] Y. Moreno, R. Pastor-satorras, and R. Vespignani, “Epidemic outbreaks in complex heterogeneous networks,” *Eur. Phys. J. B*, pp. 521–529, 2002.
- [17] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 591–600.
- [18] I. Taxidou and P. M. Fischer, “Online analysis of information diffusion in twitter,” in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, ser. WWW Companion ’14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 1313–1318.
- [19] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, “Information diffusion in online social networks: A survey,” *SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, July 2013.
- [20] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: Quantifying influence on twitter,” in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM ’11. New York, NY, USA: ACM, 2011, pp. 65–74.
- [21] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, “Epidemiological modeling of news and rumors on twitter,” in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD ’13. New York, NY, USA: ACM, 2013, pp. 8:1–8:9.

- [22] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, “Understanding retweeting behaviors in social networks,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’10. New York, NY, USA: ACM, 2010, pp. 1633–1636.

CHAPTER 3

Methodology

3.1 Descriptive Analysis and Network Characterization

Network graphs are often used as mathematical representations of a complex systems of users interacting on the Internet. Descriptive analysis and network characterization is an important part of understanding and exploring the complexity of the Twitter network under consideration.

Formally, a network graph $G = (V, E)$ is a mathematical structure consisting a set of V elements (vertices) and a set of E links (edges), where the elements of E are unordered pairs u, v of distinct vertices $\{u, v\} \in V$ [1]. The number of vertices $N_V = |V|$ and the number of edges $N_E = |E|$ are often called the order and size of the the graph, respectively [1]. Edges can be thought of as the connections between vertices on a graph.

A graph $H = (V_H, E_H)$ is a subgraph of a larger graph $G = (V_G, E_G)$ if $V_H \subseteq V_G$ and $E_H \subseteq E_G$ [1]. Induced subgraphs will also be created and utilized when making inference about the whole network. Induced subgraphs are subgraphs that are constructed from a selected set of vertices, and then adding all edges associated with those vertices to the subgraph. Formally, an induced subgraph of G is a subgraph $G' = (V', E')$ where $V' \subseteq V$ is a predetermined subset of vertices and $E' \subseteq E$ is the subset of edges found on G that are associated with V' [1]. Later in the chapter, subgraphs and the sampling methods by which they are obtained, will be discussed and proposed as models for creating prototypic, scaled down networks.

The Twitter network, and its graph representation, will often be compared to a random graph to highlight the differences between topology and its role in epidemic processes. Given this, the concept of a random graph must be introduced.

Formally, a random graph, often referred to as an Erdos-Renyi Graph, is a collection $\mathcal{G}_{N_v, N_e}$ of all graphs G with $|V| = N_v$ and $|E| = N_e$, and assigns probabilities $\mathbb{P}(G) = \binom{N}{N_e}^{-1}$ to each $G \in \mathcal{G}_{N_v, N_e}$ where $N = \binom{N_v}{2}$ is the total number of unique vertices [1]. Random graphs rarely have dense cluster or high degree nodes.

Twitter users and the connections between users can be thought of as a mathematical graph with nodes representing users and the links between users as edges. Given a graph representation of a complex system like Twitter users, it is valuable to explore the characteristics and structural properties of this graph. With clearly defined edges on a particular graph G , it is important to understand what relationship these edges represents. Are they unidirectional or bidirectional processes? This distinction is essential for modeling information diffusion on social networks since the relationship between users dictates which way information can flow on the network. Directed graphs are comprised of unidirectional edges. A graph G for which edges E have a distinct ordering, i.e., $\{u, v\}$ is distinct from $\{v, u\}$, is a directed graph [1].

It is common to express vertex adjacencies, connections between the vertices of a graph, in the form of a matrix. Matrices prove to be useful for computation and data analysis. An adjacency matrix for a directed graph is a $N_V \times N_V$ binary matrix that represents all vertex connections on a graph with entries:

$$A_{i,j} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise,} \end{cases}$$

where $i, j \in V$.

A variation of the adjacency matrix is an adjacency list. The adjacency list is often used when the graph is particularly large and/or the adjacency matrix is very sparse, i.e. contains a large proportion of zero entries [1]. An adjacency list is a $N_e \times 2$ matrix in which each row represents a connection between the two nodes

in the row. Adjacency lists will be used throughout the analysis due to the very large size and sparsity of the Twitter graph.

An important characteristic of any given vertex, v , in the network, i.e., $v \in V$, is the number of edges connecting it to the rest of the vertices. This measure of connectedness is formally referred to as the degree of a given vertex d_v [1]. For a graph, G , let f_d be defined as the proportion of vertices $v \in V$ with a degree value $d_v = d$ [1]. Given that, $\{f_d\}_{d \geq 0}$ is referred to as the degree distribution of graph G [1]. Graphically, the degree distribution of G can be represented with a histogram.

Computing a degree distribution is a common starting point for characterizing a network since the distribution of connections between the vertices often indicates to which family of graphs a particular network belongs and helps in modeling a particular process over a graph [1]. It provides a very useful summary of the connectivity of the graph. This is crucial when determining how to predict a particular process over a graph [1]. As highlighted in Chapter 2, it is common for a social networks to possess a power law degree distribution. A power law degree distribution can be expressed as an exponential function in which the frequency of nodes with a given degree, f_d , sharply decreases as degree, d , increases:

$$f_d \propto d^{-\alpha}, \quad (1)$$

where α is an exponent to be estimated to determine the relationship between a degree value and the likelihood that a node will possess that degree value.

It is common to perform a log transformation on the degree distribution of power law networks to assess how closely the degree distribution follows an exponential curve. Recalling basic properties of exponential functions, a log-transformed degree distribution should approach a straight line. As seen in Equation 2, there

is a linear relationship between $\log(d)$ and $\log(f_d)$:

$$\log(f_d) \sim C - \alpha \log(d), \quad (2)$$

where C represents an arbitrary constant.

A number of regression-based approaches have been utilized to estimate the rate, α , at which frequency decreases as a function of degree, but these approaches have come under scrutiny for their ad hoc approach [1]. These methods are not advisable due to the disproportionate variability in the data at high degrees. A more mathematically rigorous estimator that is routinely used when dealing with power-law distributions is the Hill Estimator [1]. The rate of exponential degree decrease, $\hat{\alpha}_k$, estimated using the Hill Estimator, can be computed as follows:

$$\hat{\alpha}_k = 1 + \hat{\gamma}_k^{-1}, \text{ where } \hat{\gamma}_k^{-1} = \frac{1}{k} \sum_{i=0}^{k-1} \log \frac{d_{(N_v-1)}}{d_{(N_v-k)}}, \quad (3)$$

where $d_1 \leq \dots \leq d_{N_v}$ are the sorted vertex degrees [1]. This measure of α is iteratively computed for a large range of values of k with the hope that the function, γ_k^{-1} , stabilizes at a particular alpha level. A plot of all values of α with respect to k , known as a Hill Plot, can be used to find a stable value for α . An estimate of α may then be used in sampling from a power-law distributions when attempting to create prototypic networks for further analysis.

Another useful network characteristic is a graph density, which is defined as the ratio of the number of actual edges over the maximum number of possible edges [2]. Graph density can range from zero to one, with zero values indicating a very sparse graph and values close to one indicating a very dense graph. It is important to understand how sparse or dense a graph is when dealing with very large graph, and graph density is particularly useful in understanding graph topology. Numerically, the density of subgraph H can be expressed as:

$$den(H) = \frac{|E_H|}{|V_H|(|V_H| - 1)}, \quad (4)$$

where E_H being edges of the graph and V_H being vertices of the graph [1]. Note that defining the density of subgraph H allows for flexibility of computing density for a range of graphs from the entire graph, G , or one vertex, v [1].

The clustering coefficient of a graph is an important network characteristic which is related to density calculations. Two types of clustering coefficients, global and local, are typically used when describing network characteristics. Clustering coefficients can be thought of as the average of a set of densities that were calculated for every vertex of a graph.

Clustering coefficient can be defined using a more geometric thought process involving the number of triangles in the graph. In graph theory, a triangle is a complete subgraph of order three; and a connected triple is a subgraph with three vertices connected by two edges [1]. The local clustering coefficient, $cl(v)$, can be expressed as:

$$cl(v) = \tau_{\Delta}(v) / \tau_3(v), \quad (5)$$

where $\tau_{\Delta}(v)$ represents the number of triangles of G in which v falls and $\tau_3(v)$ representing the number of connected triples in which v has two edges [1]. The number of connected triples for a given vertex can also be thought of as a binomial coefficient, $\tau_3(v) = \binom{d_v}{2}$. Intuitively, $\binom{d_v}{2}$ represents the number of possible ways to choose two edges from all the possible edges of adjacent to v .

The global clustering coefficient can be computed as a weighted average of all local clustering coefficients:

$$cl_T(G) = \frac{\sum_{v \in V'} \tau_3(v) cl(v)}{\sum_{v \in V'} \tau_3(v)} = \frac{3\tau_{\Delta}(G)}{\tau_3(G)}. \quad (6)$$

The assortativity of a graph, r , measures how similar cohesive, clustered subsets of vertices are with one another. Formally, the assortativity of a graph is defined as:

$$r = \frac{\sum_{x,y} xy(f_{xy} - f_{x+}f_{+y})}{\sigma_x \sigma_y}, \quad (7)$$

where (x, y) are all the unique pairs of nodes, f_{x+} is the degree of x minus the number of edges connecting x and y , f_{+y} is the degree of y minus the number of edges connecting y and x , f_{xy} is the joint probability distribution of the rest of the edges, and σ_x and σ_y are the standard deviations of the frequency distributions of $\{f_{x+}\}$ and $\{f_{+y}\}$, respectively.

Intuitively, a similarity of clustered vertices makes sense because one would imagine these characteristic similarities are what drew them to cluster. Groups of friends tend to have similar interests, items frequently bought together are often used in conjunction with one another. In the context of the Twitter social network, the characteristics that will be analyzed are the degree of a given vertex and the degree of the vertex's neighbors. This relationship sheds light on the similarities or differences between Twitter users that are connected with one another.

3.2 Social Network Sampling Algorithms

The sheer size of the Twitter network makes analysis computationally expensive, and often prohibitive in available software packages. Given this, considerable effort is dedicated to creating network graphs that are characteristic of the social Twitter network, but much smaller. The smaller graphs generated are easier to analyze and visualize. These graphs, however, should possess similar network topology characteristics, e.g. degree distribution, average degree, graph density, clustering coefficient, and assortativity, in an attempt to accurately render a graph that is significantly scaled down, but still representative of the original network.

In order to determine if a sampling method has produced a reasonably similar and representative subgraph of the entire network, it is necessary to compute characteristics of the whole network that will serve as a benchmark by which to measure all sampled or generated subgraphs.

In what follows, six sampling techniques are introduced in an attempt to determine which method of sampling best captures the characteristics of the larger social network. The techniques are:

1. Induced Subgraph Sampling,
2. Incident Subgraph Sampling,
3. Snowball Sampling,
4. Power-Law Degree Sampling,
5. Barabasi-Albert Preferential Attachment Network Model, and
6. Metropolis-Hastings Random Walk Algorithm.

To evaluate how similar a subgraph's characteristic distribution is to the original graph, a normalized mean square error (NMSE) will be used [3]. This metric is utilized to compare the characteristic distributions, e.g., local density, local clustering coefficient, and node degree, of the sampled graph to the original graph. Let θ_k be the fraction of nodes in the graph that have less than or equal to a characteristic value of k , and $\hat{\theta}_k$ be the fraction of nodes in a subgraph that have less than or equal to a characteristic value of k . The normalized mean square error (NMSE) is defined as:

$$NMSE(k) = \frac{\sqrt{E[(\hat{\theta}_k - \theta_k)^2]}}{\theta_k}. \quad (8)$$

Next, each of the six outlined sampling algorithms will be described.

The first method, **induced subgraph sampling**, randomly samples vertices and creates corresponding induced subgraphs. An induced subgraph consists of all sampled vertices and all adjacent edges between those specific vertices in a graph G . Specifically, n vertices are randomly chosen from V yielding a subset H , where $H = \{v_1, \dots, v_n\}$. Then, all edges from vertex pairs $(v_i, v_j) \in E$ are added to the graph to connect the vertices [1]. This method of sampling is usually applied to social network research to determine a contact networks between randomly selected people in the network [1].

The second method, **incident subgraph sampling**, involves randomly selecting n edges from the edge set E from the graph G to create a subset of edges $E_H = \{e_1, e_2, \dots, e_n\}$, and then creating a subset of vertices, V_H , that comprise the ends of all edges within E_H . Another way to conceptualize the subset of vertices that is taken during indecent subgraph sampling is to identify all the unique vertices that comprise the edges of E_H . This sampling methods most noticeably differs from induced subgraph sampling since the probability of certain vertices being selected as part of the subgraph is not uniform across all vertices in the original graph. In fact, in incident subgraph sampling, the probability that a vertex is selected is proportional to its degree. Vertices with larger degrees will contain more edges and, thus, will be more likely selected. It is important to note that although this method is slightly better at identifying larger nodes, it only takes one edge from the large-degree vertex. This proves to be problematic for a network as large as the Twitter network because the likelihood of the subgraph becoming dense enough to be representative of the whole graph is extremely low, as illustrated in Chapter 4.

Unlike previous methods, **snowball sampling** is an iterative algorithm that

is used to sample the whole network. First, a randomly subset of vertices, V_H , is generated. Next, all the neighbors of all vertices in V_H are identified and added to the subset. All edges that have been walked along to arrive at the neighbors of the original vertices are the subset of edges, E_H . These steps can be repeated for any number of iterations to create a larger subgraph, H . The number of iterations performed is referred to as the stage of the snowball sampling method [1]. For example, in the context of a social network, if one randomly selected person, vertex v , is chosen to be the original subset, then a snowball sampling method of stage 1 would include all of v 's friends. A snowball sampling method of stage 2 would include v , all of v 's friends, and all of the friends of friends of v . It is clear that this method of sampling will capture local clustering since, by construction, it creates a cluster around every randomly selected vertex. However, when dealing with a very large network, the probability that two randomly chosen vertices will have common friends, thus linking the two clusters, is very low. This problem of disjoint clusters is illustrated in Section 4.2.

Sampling from a power law degree distribution involves constructing a new graph, H , using the unique degree values of nodes in the original network in conjunction with a power coefficient, α , obtained via the Hill estimator (see Section 3.1). A set of vertices is created from sampling from the newly created degree distribution that takes the form:

$$P(k) = k^{-\alpha}, \text{ where } k \text{ is degree.}$$

From this distribution, a sequence of degree values of size n is sampled and converted into a graph with $V_H = n$ vertices and the corresponding degree values taken from the random sample.

A **Barabasi-Albert preferential attachment** network has been chosen as a graph model suitable to replicate the true structure of the original social network

graph, G . This type of graph is formed similarly to the way in which a social community on Twitter is likely to form. Specifically, when each new node, v , is introduced and added to the subgraph H , a certain number of connections are formed with existing nodes. The new node connections made to existing nodes are dependent on how many connections the existing nodes already possess. More precisely, the new nodes are more likely to make connections with heavily connected nodes. As this process continues for all the nodes being introduced to the network, hub nodes begin to possess an extraordinarily large number of connections, while most nodes possess a relatively low number of connections. Hub nodes become the centers of preferential attachment graphs.

Intuitively, Barabasi-Albert preferential models properly mimic the process by which the Twitter network of followers is created. When new users join the social network, the users are offered a list of popular account/celebrities (highly followed twitter accounts) which might be of interest to them. And given that Twitter has also become a growing source of news and entertainment for many people, it is not unreasonable to presume that people opt to follow popular Twitter accounts of news outlets and celebrities, as opposed to randomly choosing Twitter accounts to follow. These type of graphs have a degree distribution that closely follows a power law distribution. While a large percentage of nodes have a much, much smaller number of connections, this means that there is a small number of nodes that have an incredibly large number of connections. These heavily connected nodes, sometimes referred to as super-spreaders in epidemic literature, play a crucial role in the severity of the epidemic spread. Due to the process by which new users prefer to connect with existing, popular users, a preferential attachment model might be more suited to replicating a Twitter network.

The last method used to sample from the social network is a Metropolis-

Hastings Random Walk (MHRW) algorithm that employs a Bayesian approach to randomly walk along the graph from one vertex to another if certain conditions are met. Formally, the algorithm can be described as follows: First, a randomly selected vertex, v , is sampled from the entire graph and set as the initial vertex. One of the edges of vertex v is randomly selected and is walked along to the corresponding neighbor vertex, w . Two proposals functions, $Q(v)$ and $Q(w)$, are created that correspond to the degree of nodes v and w , respectively. Next, $Q(v)$ is divided by $Q(w)$ and compared to a randomly generated number, p , from a continuous uniform distribution defined on the interval $[0, 1]$. If the ratio of the proposal functions $Q(v)$ and $Q(w)$ is greater than the randomly generated uniform number, p , the new vertex, w , is added to the subset of vertices and the process then starts over with vertex w as the starting point and its neighbors are analyzed. Otherwise, this process continues until a prespecified algorithm cost is met or until a specified vertex count is met. For this project, the algorithm is run until a specified vertex count is met that is similar in size to subgraphs obtained by the other sampling methods. Algorithm 1 illustrates pseudo-code of the MHRW algorithm that is used to sample the subgraph G^* from the entire social network G .

Data: Social Network Graph, G

Result: Subgraph, G^* , using Metropolis-Hastings Random Walk

Initialization;

```
while  $V(G) < 2000$  do
    Randomly Select a vertex,  $v$ , from  $G$ ;
     $Q(v) = d_v$ ;
    Random neighbor,  $w$ , of  $v$  is selected;
     $Q(w) = d_w$ ;
    Let  $p$  be a random sample from  $\text{runif}(0,1)$ ;
    if  $\frac{Q(v)}{Q(w)} > p$  then
        Crawl to  $w$ ;
         $w$  becomes initial node;
    else
        Randomly sample another neighbor from  $v$ ;
    end
end
```

Algorithm 1: Metropolis-Hastings Random Walk Algorithm.

3.3 Information Diffusion and Epidemic Modeling

The information diffusion process on a social network can be viewed as a Susceptible-Infected, or SI model; a type of epidemiological model that allows for each node in the network to become infected, and remain infected for the duration of the process. If an informed user can be considered infected, it is entirely possible for other Twitter users to retweet (become infected) another users story.

The important characteristic of the SI model is that once a user becomes infected, that user is considered infected for the entire duration of the epidemic process. The SI-type of model is chosen over other epidemic models because the

information diffusion model is concerned with how much of the network has become aware of the discovery of the news, not the frequency with which certain users retweet, mention, or reply to other users. Others epidemic models take into account the possibility to recover from the infection become susceptible after an initial infection. Although interesting, these models in their current form are not applicable to the information diffusion modeling at hand. Modified, network-based epidemiological models will be used to account the nonrandom topology of the Twitter network.

Traditional SI epidemic models, expressed by Equations 9 and 10 consist of two probabilities: probability of an infection, and probability that all individuals remain in the same state that they were in the previous time period (either infected or susceptible). $N_S(t)$ represents the number of susceptible individuals, $N_I(t)$ represents the number of infected individuals, and δt is the infinitesimal change in time. This model does not take network structure into account. In fact, it assumes that the population is homogenous and that each infected person is equally likely to infect any of the susceptible individuals. In terms of network structure, this would be referred to as a complete graph, or a graph in which every node has an edge connected to all other nodes.

$$\mathbb{P}(N_S(t + \delta t) = s - 1, N_I(t + \delta t) = i + 1 | N_S(t) = s, N_I(t) = i) \approx \beta s i \delta t \quad (9)$$

$$\mathbb{P}(N_S(t + \delta t) = s, N_I(t + \delta t) = i | N_S(t) = s, N_I(t) = i) \approx 1 - \beta s i \delta t \quad (10)$$

An important metric in epidemic modeling is the basic reproduction number, R_0 . The basic reproduction number is defined as the number of infections expected in the by a single infected person [1]. This number is very important in determining whether or not an epidemic will occur because it provides an estimate for the average number of people a given infected person will subsequently

infect. If $R_0 < 1$, then as time increases the infection will die out because each new person, on average, infects less than one new person. If $R_0 > 1$, an epidemic is likely to occur because each infected person is infecting more than one person. It is important to note that these epidemic predictions based on R_0 apply to homogenous populations. Later in this chapter, the basic reproduction number will be discussed in a network-based framework.

Given what is known about the topology of the social network, the traditional model represented in Equations 9 and 10 would not be suitable for a network-based information diffusion process. The model must be modified to restrict each node to have the potential to infect only its neighbors.

Equation 13 shows the network based epidemic process that will be modeled on the sampled subgraph (see [1] for more details). Let $\mathbf{X}(t)$ represents a stochastic process which forms a continuous-time Markov chain. This process is comprised of states vectors $\mathbf{x}$ that consist of either 1 or 0, with 1 being infected and 0 being susceptible. The state vectors and the numbers that comprise them represent the state of a given node over time. The coefficient β is the transmission rate that represents the probability that the neighbor of an infected vertex becomes infected, and $M_i(\mathbf{x})\delta t$ represents the number of neighbors of i infected at time t .

$$\mathbb{P}(\mathbf{X}(t + \delta t) = \mathbf{x}' | \mathbf{X}(t) = \mathbf{x}) \approx \begin{cases} \beta M_i(\mathbf{x})\delta t, & \text{if } x_i = 0 \text{ and } x'_i = 1 \\ 1 - \beta M_i(\mathbf{x})\delta t, & \text{if } x_i = x'_i \text{ and } x_i \in \{0, 1\} \end{cases} \quad (11)$$

where $\mathbf{x}'$ is a successive change of state from $\mathbf{x}$ involving only one element at a time.

The basic reproduction number for an SI epidemic process on a graph can be defined as:

$$R_0 = \frac{\mathbb{E}(d^2)}{\mathbb{E}(d)} - 1, \quad (12)$$

where $\{f_d\}$ is a degree distribution and $\mathbb{E}(d)$ & $\mathbb{E}(d^2)$ are the first and second moments of f_d , respectively [1]. While this reproduction number produces reasonable estimates for processes on random graphs, it is shown in [1] that the R_0 dramatically increases for graphs that have heterogeneous degree distributions. This results from the fact that $\mathbb{E}(d^2) \gg \mathbb{E}(d)$ and often the second moment is extremely large due to the long tail of a power-law degree distribution. This fact illustrates that epidemics are much more likely to occur on power-law networks because there are very high degree nodes that have the potential to infect a large number of other nodes.

The network-based epidemic process is modeled on the sampled subgraph by randomly selecting one vertex and applying the transmission rate to determine the number of neighbors that will become infected on the first iteration. Next, all newly infected vertices spread the infection to their respective neighbors according to the transmission rate.

A root mean square error (RMSE) metric is used to compare epidemic models of varied transmission rates over various types of graphs. The proportion of the Twitter network that is aware of the news is compared to the estimated proportion for each day. Transmission rates can vary from zero to one, and the minimum mean square error value identifies the transmission rate that best fits the actual spread of information on a particular graph type.

Basic graph theory, graphs characteristics, sampling techniques, and information diffusion processes introduced in this chapter will be applied to the Twitter social network in an attempt to explore how information diffuses across an online social network and summarized in Chapter 4.

More complex information diffusion models for the Twitter network will be discussed in Chapter 5 (Conclusions). These models incorporate the Twitter action of mentioning other users in a tweet.

List of References

- [1] Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY, United States of America: Springer, 2009.
- [2] P. Hu and W. C. Lau, “A survey and taxonomy of graph sampling,” *CoRR*, vol. abs/1308.5865, 2013.
- [3] T. Wang, Y. Chen, and others”, “Understanding graph sampling algorithms for social network analysis,” *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference*, pp. 123–128, 2011.

CHAPTER 4

Results

Formally, network graphs and the characteristics that are important when summarizing topological structure have been introduced in Chapter 3. In this chapter, the network characteristics will be presented in application to the Twitter network.

First, the node degree distribution analysis will be discussed in conjunction with a Hill estimator and its applications to subnetwork sampling from a power-law degree distribution. The corresponding local clustering coefficient and local density of the full network graph will also be examined. Next, six sampling techniques will be used and their corresponding subgraph visualizations and characteristic summaries including: clustering coefficients, average degree, assortativity, and density will be studied for each sampling method. Finally, a network-based information diffusion model run on the sampled graphs will be analyzed and compared to the observed spread of information on the social network, i.e., the spread of the news of the discovery of the Higgs particle on Twitter.

4.1 Topology Characterization of Social Network

Recall that the data used in this thesis was originally collected by De Domenico et al. [1]. The dataset consists of four large interaction networks related to the spread of the news of the scientific discovery in July 2012, namely social, retweet, mention, and reply networks. In this project, the main focus will be on the social network and the retweet network. The social network is the largest one, containing roughly 450,000 nodes with 14,000,000 edges, and represents the network of Twitter followers engaged in the dynamic spread of this news. The retweet network, with 425,000 nodes and 733,00 edges, represents all the retweets done during the

spreading period.

Table 1 summarizes network characteristics of the observed Twitter network. The average degree of this entire network is 65.07. It is important to note that while the average is roughly 65, the distribution of the numbers that comprise this average are extremely varied and skewed with degree values as high as 50,000. The L.C.C. column presents the same characteristics that were presented for the entire social network, but for the largest connected component (L.C.C.) of the Twitter network. As seen in Table 1, the size of the two graphs, the L.C.C. and the entire Twitter graph, only differ by 345 nodes. This indicates that there are few nodes that are disjoint from the graph. Values of assortativity and clogged clustering coefficients are also summarized in table 1. These values will be discussed later in the chapter.

Figure 3 shows that the probability of a given vertex within the social network possessing a certain degree value decreases very rapidly as degree increases. This indicates that the overwhelming majority of the Twitter users have a relatively small number of connections to other users and the number of people they are following or followed by is relatively modest when compared to the tail of the distribution. The tail of the distribution extends to roughly fifty thousand. These types of users are most likely celebrities within the particle physics world and can explain why one particular Twitter user possesses fifty thousand connections to other users.

Characteristic	Social Network	L.C.C.
Number of Nodes	456,626	456293
Number of Edges	14,855,842	14,855,497
Clustering Coefficient	.188	.199
Assortativity	-.1375	-.1376
Average Degree	65.07	65.11
No. of Components	157	1

Table 1: Network Characteristic Summaries Including: Number of nodes, global clustering coefficients, average node degree, degree assortativity, and number of connected components for the observed network and the largest connected component (L.C.C.).

Given the exponential nature of the degree distribution, it is very difficult to discern any knowledge of the tail because the first bin of the histogram is so large relative to the size of the other bins. Instead, the log transformed degree distribution is used to understand the extent to which the distribution is exponential. The log degree distribution function shows a relatively straight line, and suggests that the node degree distribution is exponential (see Figure 3). The smaller degree vertices tend to arch upwards and deviate slightly from a perfectly straight line, indicating that the smaller degree vertices comprise a portion of the distribution that may deviate slightly from an exponential shape than the larger degree vertices. However, once the degree values exceed one hundred, the distribution quickly follows a very straight line. This emergence of a linear trend in the log-transformed degree distribution indicates that the distribution is exponential.

Figure 3: Degree Distribution.

Recall that the exponential degree distribution, requires estimation of one parameter, α , that converges to the rate at which the function decreases. As mentioned in Chapter 3, the Hill estimator is used to estimate the exponential coefficient to better understand the nature of the exponential degree distribution.

Figure 4: Hill Plot. of Power-law exponent, $\hat{\alpha}_k$, as a function of number of order statistics used, k . Stabilizes at $k \sim 40$ with a value of $\hat{\alpha}_k = 2.4$.

Figure 4 shows that the exponential coefficient, α , quickly stabilizes to roughly

2.4 as k , the number of order statistics used by the hill estimator, increases. This value of 2.4 will be used later to construct a network subgraph with a power-law degree distribution similar to the degree distribution of the entire social network.

The distribution of the local clustering coefficients, illustrated in Figure 5, indicates that a large proportion of the vertices have a clustering coefficient below .2 and relatively similar to the global value of .188 (Table 1). A tail is definitely present with a noticeable number of vertices having the maximum number of possible triangles. The skewed distribution of local clustering coefficients indicates that the majority of twitter users do not form cliques with other users that are following the same accounts. Given that most people follow celebrity accounts that have a very large following (some celebrity accounts approaching 50,000 followers), it is unlikely that users would follow strangers that are also following the celebrity account. The small uptick at the end of the tail is likely attributable to small components of two or three users that are disjoint from the largest connected component, but connected to each other. These types of groups are likely users that were informed via another mode of media and decided to share the information with their close friends on Twitter.

Figure 5: Local Clustering Coefficient Histogram (Average: .147).

Figure 6: Local Density Histogram (Average: .221).

The histogram of the local densities of the individual vertices (Figure 6) peaks just below the average value of .221. However, the sudden increase in the vertices possessing a local density of .5 paired with the long right-hand tail certainly skew the distribution. The local density histogram is similar to the local clustering coefficient in that they both possess a noticeable spike in frequency at the value 1. The distribution of the local densities offers insight into the interactions of Twitter users. The majority of users have a local density below .3. These low density values offer an interoperation similar to that of the clustering coefficients. The relatively small number of edges of a subgraph compared the number of vertices indicates that there is not a lot of following occurring between the users that are all following the celebrity accounts.

Figure 7: Vertex Degree Versus Average Neighbor Degree.

The average degree of a vertex's neighbors indicates how connected a given vertex's neighbors are to the rest of the graph. Figure 7 illustrates that when a vertex's degree is relatively low, the variation of the average degree of its neighbors is very high. This suggests that low degree vertices are as likely to connect to low degree vertices as they are to very high degree vertices. However, as the degree values increase, the variation of average neighbor degree values converge to roughly 125 with less variance explained by a small number of higher degree nodes. Figure 7 visually supports a negative correlation between the degree of a vertex and the degree of its neighbors which, in turn, compliments the negative assortativity of -0.135 (Table 1) when examining the entire observed Twitter network.

4.2 Social Network Sampling and Inference

Sampling the large social network is done to obtain representative subgraphs that are small enough in size to be able used in a sequential modeling of information diffusion processes. As mentioned in Chapter 1 and Chapter 3, the network of Twitter followers under consideration is too large to be analyzed using currently available software packages, and, therefore, must be sampled for further modeling.

The following sampling methods will be used to sample the original social network and compared in terms of efficiency at producing subgraphs with a structure similar to the original social network:

1. Induced Subgraph Sampling,
2. Incident Subgraph Sampling,
3. Snowball Sampling,
4. Power-Law Degree Sampling,
5. Barabasi-Albert Preferential Attachment Model, and
6. Metropolis-Hastings Random Walk Graph Sampling Algorithm.

The induced subgraph sampling method produces graphs that are sparse and disconnected. A clustering of vertices takes shape in the center of the graph, but there are only a few clusters and they consist of a relatively small number of nodes compared to the entire subgraph (see Figure 8). This small amount of clustering is expected since there is a very large number of nodes to be sampled with low node degree. Intuitively, the likelihood that two randomly sampled nodes from roughly a half million nodes are connected or have common neighbors is low. Thus, it is not surprising that the degree distribution is heavily weighted towards a degree of zero because the likelihood that a given node from the sampled subset is connected to another node in the subset is also low. While the shape of the distribution is ostensibly similar to the original graphs degree distribution, the network graph visualization indicates that the very large proportion of zero degree nodes does not construct a subgraph that is representative of the original graph. All sampling methods are random in nature. In order to assess the performance of the proposed sampling methods, one hundred graphs were randomly generated

using each sampling method. The tables corresponding to each sampling method display characteristics that are averages and confidence intervals for all one hundred graphs for a given sampling method.

Table 2 shows that this sampling algorithm produces graphs with very low clustering coefficients, negative assortativity, and a density that is very close to zero, on average. Density and assortativity values for these graphs are reasonably close to the characteristics of the original Twitter network, but the clustering coefficient is significantly lower. Since clustering characteristics are very important when modelling information diffusion, a sampling method that produces graphs with such dissimilar clustering coefficients can be problematic. Thus, other sampling methods will need to be employed in an attempt to obtain a more representative social network.

Figure 8: Induced Subgraph Sampling. Left Panel: Visualization of Induced Subgraph Sampling with $N_V = 2000$ and $N_E = 297$. Right Panel: Degree Distribution of Induced Subgraph.

Characteristic	Average	5th Percentile	95th Percentile
Degree	0.29	0.20	0.42
Clustering Coeff.	0.03	0.00	0.08
Assortativity	-0.08	-0.23	0.06
Density	0.00007	0.00005	0.0001

Table 2: Induced Subgraph Characteristics. Estimates obtained from averaging the graph characteristics of 100 induced subgraphs sampled from the original social network.

The incident subgraph sampling method randomly selects edges from the roughly fourteen million edges present in the entire social graph. Once a subset of edges is selected, it collects all nodes that comprise the edges in the subset. Finally, it produces a graph of the edges and nodes from their respective subsets.

Figure 9 suggests that almost all of the edges in the graph are disjoint and connect only to a few vertices. A lack of clustering, highlighted by average the clustering coefficient summarized in Table 3, is even more apparent in this sampling technique since it is highly unlikely to sample two vertices have common edges. This is mostly because of the extremely large number of edges that are all equally likely of being selected. Also, one particular feature that is absent in the subgraph obtained using incident sampling that is present in all power-law networks is the existence of highly connected nodes, sometimes referred to as super spreaders in epidemiological literature. These vertices play a crucial role in the spread of information on a social media graph because of their ability to infect a very large number of other users. The absence of these types of vertices requires another method is better suited to adequately sample the social network.

Table 3 shows that the average connectivity of the nodes in incident subgraphs is very low, .29. This value is not a good estimate because it implies that a large number of the users in the social network have no connections to any other users. The global clustering coefficients for the sampled graphs are all very near to zero,

and the assortativity values are also very close to zero. These ranges of values for one hundred sampled subgraphs are significantly far removed from the true values of the original social network .

Figure 9: Incident Subgraph Sampling. Left Panel: Visualization of Incident Subgraph Sampling with $N_V = 3689$ and $N_E = 2000$. Right Panel: Degree Distribution of Incident Subgraph.

Characteristic	Average	5th Percentile	95th Percentile
Degree	1.27	1.25	1.28
Clustering Coeff.	0.0001	0.00	0.0023
Assortativity	-0.05	-0.07	-0.02
Density	0.0002	0.0018	0.00021

Table 3: Incident Subgraph Characteristics. Incident Subgraph Characteristics. Estimates obtained from averaging the graph characteristics of 100 incident subgraphs sampled from the original social network.

The snowball sampling technique, which randomly selects a subset of vertices from the original graph and includes all neighboring vertices of the selected vertices, creates a graph that is in some ways an improvement over previously discussed methods. Given the nature of including all the neighbors of the randomly selected subset of vertices, the clustering of this graph will be much higher than

previously discussed methods because every vertex that is selected with its cluster of neighbors. The disadvantage to snowball sampling a graph is that the likelihood that these neighborhood clusters have connections to other cluster that are randomly selected is very low. The chances of connections between clusters are higher than the chances of clustering in induced and incident subgraph sampling methods, but they are still very low compared to the original graph.

In terms of obtained graph characteristics, the average global clustering coefficient of the snowball sampled graphs is significantly greater than the clustering coefficient of the original graph (Table 4). Intuitively, this is expected because every single vertex in the snowball subgraph will contain all of its neighbors. The average of all of these local clustering coefficients will produce a global clustering coefficient that is higher than the original graphs coefficient. As shown in Figure 10, the clusters of vertices surrounding all the randomly selected seed vertices are quite dense. However, all of these clusters are disjoint, which is not representative of the original social.

Unlike network graphs obtained using the snowball sampling algorithm, the original social network is almost entirely connected and the largest connected component is only a few vertices less than the entire set of vertices in the whole graph (Table 1). The problem of disjoint clusters contributes to the higher than desired clustering coefficient (Table 4). This, in turn, would prove to be problematic when running information diffusion models on the subgraph because the information will not be able to disseminate past the cluster in which the infection began. This will contradict the dynamics of the actual information spread on the almost fully connected original social network. The redeeming aspect of a snowball sampling algorithm applied to a power-law network is that it is much more likely, as shown in Figure 10, to capture super spreader nodes from the original graph. The snowball

subgraph possesses a node with a degree of roughly 1,200. As mentioned earlier, these nodes are crucial when modeling information diffusion.

Figure 10: Snowball Sampling. Left Panel: Visualization of Snowball Sampling with $N_V = 1412$ and $N_E = 14894$. Right Panel: Degree Distribution of Snowball Sampled Subgraph.

Characteristic	Average	5th Percentile	95th Percentile
Degree	23.76	11.20	48.05
Clustering Coeff.	0.33	0.09	0.49
Assortativity	-0.03	-0.23	0.24
Density	0.01	0.0044	0.0217

Table 4: Snowball Sampling Characteristics. Snowball Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 snowball sampled from the original social network.

Sampling degree values from an exponential distribution that is similar in shape to the degree distribution of the entire social graph will require the degree sequence values and the exponent estimate, α , by which frequency will decrease as a function of degree. As shown by the Hill plot (Figure 4) the exponent estimate stabilizes at $\alpha = 2.4$, and will be used when creating an exponential distribution from which degree values will be sampled. The unique degree values from the

full social graph will be degree sequence that will be raised to the power of -2.4. This means that as degree increases, the probability of selecting that degree value will decrease. The subgraph that is sampled from this distribution produces level clustering (Figure 11) and the degree distribution is still exponentially decreasing as a function of degree. This graph modeling method has a fixed number of vertices, N_V , and a varying number of edges, N_E .

The average degree value, 1.99, for the power-law networks is significantly lower than that average degree value of the original social network, 65. The average of the global clustering coefficients is very near to zero, and significantly different than the global clustering coefficient of the social network. The assortativity confidence interval for the sampled graphs contains zero, indicating that there is no significant assortative mixing in these types of graphs.

Figure 11: Power-Law Distribution Sampling, $P(k) = k^{-2.4}$. Left: Visualization of Power-Law Sampling with $N_V = 2000$ and $N_E = 2016$. Right: Degree Distribution of Power-law graph.

Characteristic	Average	5th Percentile	95th Percentile
Degree	1.99	1.78	2.30
Clustering Coeff.	0.01	0.0043	0.0218
Assortativity	-0.001	-0.034	0.03
Density	0.0009	0.00089	0.0011

Table 5: Power-Law Sampling Characteristics. Power-law Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 power-law graphs.

The Barabasi-Albert preferential attachment network modeling produced a graph visualized in Figure 12. The left panel of Figure 12 shows a graph with large, dense clusters around a small number of nodes. Recall that this method creates graphs by iteratively adding one node to the graph until a specified vertex count is reached. The newly introduced nodes are more likely to connect to higher degree nodes. This preferential attachment model produces graphs with that possess a small number of highly connected nodes. While this is certainly a characteristic of the original network, the clustering in the rest of the graph is very low and the connections between the nodes that connect to the preferred nodes is almost nonexistent. These characteristics contribute to the graphs very low clustering coefficient that is near zero. Although the very high degree nodes were able to be constructed using this method, the very low clustering coefficient prompts further analysis into other sampling techniques to determine if another method is preferable.

Figure 12: Barabasi-Albert Sampling Method. Left: Visualization of Barabasi-Albert preferential attachment graphs. Sampling with $N_V = 2000$ and $N_E = 1999$. Right: Degree Distribution of Barabasi-Albert preferential attachment graphs.

Characteristic	Average	5th Percentile	95th Percentile
Degree	2.01	1.78	2.30
Clustering Coeff.	0.0102	0.0043	0.018
Assortativity	-0.00043	-0.035	0.034
Density	0.001	0.0088	0.0013

Table 6: Barabasi-Albert Sampling Characteristics. Estimates obtained from averaging the graph characteristics of 100 Barabasi-Albert preferential attachment graphs.

The Barabasi-Albert preferential attachment model requires a predetermined number of vertices, N_V , and the resulting number of edges, N_E , will be $|V_N| - 1$. This is true because at every iteration one new node is added and it makes one connection to an existing node, but the first node introduced does not have any possibility of connecting to an existing node. This results in a graph with an edge count that is one less than the vertex count.

Table 6 shows that the average global clustering coefficients of the Barabasi-Albert graphs are also very near to zero. The assortative mixing of these types

of graphs is also not significantly different than zero. The average values of the clustering coefficients and assortativity values of the Barabasi-Albert graphs are significantly different than original social network graph.

Figure 13: Metropolis-Hastings Random Walk Sampling. Left: Visualization of Metropolis Hastings random walk graph. Sampling with $N_V = 1412$ and $N_E = 14894$. Right: Degree Distribution of Metropolis Hastings random walk graph.

Characteristic	Average	5th Percentile	95th Percentile
Degree	20.41	20.21	20.69
Clustering Coeff.	0.1763	0.1683	0.1843
Assortativity	-0.1756	-0.1816	-0.169
Density	0.0044	0.00432	0.00468

Table 7: Metropolis-Hastings Random Walk Sampling Characteristics. Metropolis Hastings random Walk Graph Characteristics. Estimates obtained from averaging the graph characteristics of 100 Metropolis Hastings random walk graphs.

The Metropolis-Hastings Random Walk (MHRW) sampling method, detailed in Algorithm 1, produces graphs by analyzing the ratio of node degree value and the degree of one of its randomly selected neighbors. If this ratio is larger than a value randomly selected from a continuous uniform distribution on $[0,1]$, the algorithm includes the neighboring node and edge, and then restarts the selection

process from the newly selected node.

The graph network illustrated in Figure 13 is a product of a MHRW sampling method applied to the original social network. As can be seen in the left panel, there are distinct clusters that are connected by a relatively smaller number of edges. The average clustering coefficient of the one hundred graphs produced using this method is .178. This is the closest average of all six sampling techniques analyzed. This similarity between the clustering coefficient of the MHRW graphs and the original social network makes the MHRW method very appealing for producing prototypic graphs that can be used for information diffusion analyses.

Char./Method	Induced	Incident	Snowball	B.A.	Power-Law	MHRW
Avg. Degree	.369 (.009)	.292 (.001)	1.69 (1.81)	-.06 (~ 0)	-.056 (.04)	12.02 (.13)
C.C.	.002 (.002)	.00003 (.00007)	-.063 (.0001)	-.0092 (~ 0)	-.0084 (.0008)	0.0 (.001)
Assort.	.13 (.007)	-.03 (.004)	-.0457 (.0002)	.0031 (~ 0)	.00283 (.005)	.0486 (.0006)
Density	.00001 (.0004)	-.00035 (.00001)	-.0144 (~ 0)	-.0014 (~ 0)	-.0014 (.0007)	-.0058 (.00007)

Table 8: Sensitivity C.I.s

Induced Subgraph Sampling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	0.071	0.144	0.29	0.44
Clustering Coeff.	0.0003	0.021	0.03	0.0023
Assortativity	-0.005	-0.0004	-0.08	-0.01
Density	0.000074	0.00007	0.0007	.00006
Incident Subgraph Sampling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	1.078	1.45	1.27	1.37
Clustering Coeff.	0.00008	0.0001	0.0001	0.00038
Assortativity	-0.025	-0.037	-0.05	-0.055
Density	0.0005	0.0003	0.0002	0.00015
Snowball Sampling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	23.38	21.47	23.76	25.07
Clustering Coeff.	0.352	0.336	0.33	0.289
Assortativity	-0.046	-0.018	-0.03	-0.0003
Density	0.0178	0.0082	0.01	0.0034
Power-Law Graph Modeling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	2.046	2.02	1.99	1.99
Clustering Coeff.	0.017	0.013	0.01	0.0086
Assortativity	-0.000033	-0.0007	-0.001	0.0028
Density	0.002	0.0082	0.0009	0.0006
Barabasi-Albert Graph Modeling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	2.05	2.002	2.01	1.99
Clustering Coeff.	0.0178	0.013	0.0102	0.0086
Assortativity	-0.0003	-0.0028	-0.00043	0.0028
Density	0.002	0.0082	0.001	0.0006
MHRW Sampling				
Characteristics/Nodes	500	1000	2000	3000
Average Degree	10.68	23.33	20.41	22.7
Clustering Coeff.	0.159	0.178	0.1763	0.159
Assortativity	-0.225	-0.189	-0.1756	-0.1764
Density	0.0088	0.0065	0.0044	0.003

Table 9: Sensitivity Analysis

Along with global graph characteristics such as clustering coefficient, assortativity, and density, it is also important to compare the distributions of the local values of these characteristics to the distributions of the original network. A normalized mean squared error (NMSE) approach is used to compare the distributions of the sampled networks versus the original network. As discussed in Chapter 3.2, the normalized mean square error statistic is used to compare distributions using the difference in the proportion of nodes with a value less than or equal to a value for both distributions. As the characteristic value increases, the difference between the two proportions is expected to approach zero because both proportions will be approaching 1. Given this, the closer a given curve is to the x-axis, the better the fit. Another way to think about this is that if the original degree distribution was compared to itself the difference would be zero for every degree value and produce a line that runs along the x-axis.

As seen in Figure 14, the NMSE of the degree distributions for all the sampling methods decrease quickly, but there are noticeable differences in the values for each sampling algorithm. For instance, incident subgraph sampling has the least similar distribution because its NMSE is the largest for the smaller degree values. Induced subgraph, power-law, and snowball sampling all have similarly shaped MSE plots that exhibit similar degree properties. The MHRW and Barabasi-Albert sampling methods have MSE plots that are much closer to the x-axis, indicating a better fit. It is clear that the MHRW method and Barabasi-Albert produce degree distributions that are closest to the degree distribution of the original network.

Figure 14: Normalized mean square error (NMSE) of sample graph degree distributions as a function of degree index, k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a degree less than or equal to k and the proportion of nodes in the original graph with a degree less than or equal to k .

A NMSE analysis is also performed on the distributions of the local clustering coefficients to determine which sampling method produced a distributions of local clustering coefficients most similar to that of the original network. Figure 15 shows the MHRW and snowball sampling methods noticeably outperforming the other sampling methods. The NMSE plot for the snowball sampling method performs very well for small clustering coefficient values, but then rises above the rest of the plots when the clustering coefficient approaches .2. The MHRW sampling method produces a NMSE plot that performs slightly worse than the snowball method for local clustering coefficients between 0 and .1, but performs better once the clustering coefficient is beyond .1.

Figure 15: Normalized mean square error (NMSE) of sample graph local clustering coefficient distributions as a function of all possible clustering coefficient values, i.e., zero to one, and denoted as k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a clustering coefficient less than or equal to k and the proportion of nodes in the original graph with a clustering coefficient less than or equal to k .

Figure 16: Normalized mean square error (NMSE) of sample graph local density distributions as a function of all possible density values, i.e., zero to one, and denoted as k . NMSE is computed by computing the squared difference between the proportion of nodes in a sampled graph with a density less than or equal to k and the proportion of nodes in the original graph with a density less than or equal to k .

A NMSE analysis of the density distributions (Figure 16) was also performed. The distribution of the local density of a graphs nodes is compare the local density

distribution of the original social network. Induced subgraph sampling performs very poorly, relative to the other sampling methods, when attempting to preserve the density distribution of the original network. The reason for this poor fit is that a large percentage of the nodes have a degree of zero and are not connected to any part of the graph. Snowball sampling performs better than incident, but is still noticeably worse than the other four methods. The performance of the other four methods, namely Barabasi-Albert, Induced, Power-Law, and MHRW are all comparable.

Given the analyses performed in this section, the three graph models on which information diffusion models will be run on the MHRW graph, the Barabasi-Albert preferential attachment graph, and the power-law graph. Also, only these three models managed to provide enough similarity to the original Twitter network in terms of topological characteristics. Of these three, the MHRW graph types can be considered the best because they also possess clustering coefficients that are very similar to the original network, and have an assortativity that is very similar to the original network.

4.3 Modeling the News of the Higgs-Boson Discovery on Twitter

The discovery of the Higgs boson was a monumental event in scientific history and such news can expected to spread through a social network of people in the scientific community. The results below illustrate how the spread of this news across Twitter evolved over the course of a week. The three types of activity, namely retweeting, mentioning, and replying, are plotted separately in Figure 17. The results focus on the intensity of activity over time and how these actions on Twitter spread throughout the social network.

Figure 17: Frequency of Twitter actions as a function of day. Three Twitter actions, Mentioning, Replying, and Retweeting are plotted as functions of time. RT stands for Retweeting (Top Panel), RE stands for replying(Middle Panel), and MT stands for mentioning(Bottom Panel).

Overall, Figure 17 shows the frequency with which each action occurred over the course of the week. For all three types of activity, it is clear that there is a surge in activity in the middle of the week with periods of relatively low activity at the beginning and end of the week. It is important to note that while all three actions have a very similar distribution, the number of replies is much smaller than that of mentioning or retweeting.

Figure 18: CDF of Twitter actions as a function of day. Three Twitter actions, Mentioning, Replying, and Retweeting are plotted as functions of time. RT is retweeting, RE is replying, and MT is mentioning.

The cumulative distribution functions for all three actions (Figure 18) are very similar in shape and this is to be expected given the similar shapes of the intensity plots. The most notable characteristic of the CDFs is the sharp increase near the middle of the time period.

Until this point of the results section, the analyses of the network characteristics and the analyses of the temporal dynamics of the information spread have been separated. Moving forward, the two will be analyzed in conjunction to understand how the information diffuses over the social network. Figure 19 shows the proportion of users that have become aware, through one of the three actions, of the scientific discovery. Over the first few days, the news does not reach a large proportion of the network, but in the middle of the time period the proportion of people aware of the discovery sharply increases. This sigmoidal shape of the proportion of users to which the information has spread is similar in shape to the proportion of people who become infected in a SI epidemic model. This similarity further supports the use of an epidemic model when attempting to quantify the

scope of information diffusion on the social network.

Figure 19: Proportion of Users Discovering The Higgs-Boson News Over Time as a Function of Time in Days.

As discussed in Chapter 3, the transmission rate, β , is an important factor when modeling information diffusion as an epidemic process. Susceptible-Infected, network-based epidemic models were run on the MHRW, the Barabasi-Albert, and Power-law sampled graphs with transmission rate ranging from zero to one. The minimum of each MSE curve is the transmission rate that produces a model that best fits the actual spread of information across the social network. For example, Figure 20 shows the MSE curve produced for information diffusion models on a MHRW graph for transmission rates ranging from zero to one. A minimum value of β was reached at .056. The transmission rate of .056 indicates that at every iteration of the process, every neighbor of an informed node has a 5.6 percent chance of retweeting the information of their informed neighbor.

Figure 20: RMSE of SI models as a function of the transmission rate, β , for three graphs generated from sampling techniques. The minimum of each line represent s the transmission rate at which a SI model on a particular graph best fits the actual proportion of Twitter users becoming aware of the Higgs boson news (Figure 19) The MHRW RMSE has the smallest minimum value of the three RMSE plots and has the best fit to the actual proportion of users becoming aware over time. MHRW Minimum at $\beta = .056$.

Figure 21: SI epidemic processes run across three of the sampled graph types. The black line is the actual information diffusion process across the entire Twitter network. This network is used to compute the mean square error for each process by comparing the proportion of informed users at each timestamp. The processes on the Barabasi-Albert and power-law models have a transmission rate of 1. The transmission rate for the process on the MHRW graph is .056.

Next, a SI epidemic model process was run on each type of graph using the transmission rate for which each graph type had a minimum on the MSE plot, $\beta = .056$ for MHRW, and 1 for Barabasi-Albert and Power-law sampling. The diffusion models were depicted in figure 21 and superimposed on the original information spread (black curve). The SI model on the MHRW most closely fits the actual spread of information on the Twitter network. The processes on the Power-law and Barabasi-Albert graphs take longer than seven iterations to begin to spread to significant portions the graph. If the number of iterations were extended beyond seven, the processes on these two graph types would look similar in shape to the original spread on the Twitter network. For example, Figure 22 illustrates epidemic processes run on a Barabasi-Albert preferential treatment graph and a power law graph. The processes were illustrated to show that the increase in the proportion of informed users is similar to that of the process on the MHRW, but that a larger number of iterations is required. The sigmoid, or S-shaped, growth curve is common to processes on all three graphs, but the Barabasi-Albert and power law graph processes do not begin to increase until the seven day period of Twitter news spread has already passed.

Figure 22: Total number of iterations for epidemic processes run across the Barabasi-Albert and power-law graphs with a $\beta = .4$. The number of iterations visualized was extended from 7 to the total number required to completely reach the entire graph to illustrate that the graphs still conform to a shape similar to the MHRW and actual information diffusion curves (a sigmoid curve), but it takes more iterations to complete the process.

In summary, this chapter details the results of sampling algorithms and information diffusion processes. The six sampling algorithms were evaluated by comparing various sample graph characteristics to those of the original Twitter graph. The information diffusion processes were evaluated by comparing the actual proportion of informed users over to time to the proportions estimated in epidemic models. In the next chapter, results will be discussed and topics for future work and continued analysis will be discussed.

List of References

- [1] M. D. Domenico, A. Lima, *et al.*, “Anatomy of a scientific rumor,” *Scientific Reports*, no. 4.

CHAPTER 5

Conclusion

5.1 Main Findings

The discovery of the Higgs boson was a momentous scientific achievement. In the days surrounding this breakthrough, news of the discovery spread across online social networks. The focus of this thesis was to attempt to model the spread of this news using networks-based epidemic models. Three main analyses have been performed in order to understand this process. First, a characterization of the large social network has been performed in order to understand the topology of the network. These characterizations have been used as benchmarks to evaluate the performance of various sampling techniques. Finally, the information diffusion model have been run on a sampled subgraph to determine if an epidemic model could be used to model the spread of information across the Twitter network.

Characterization of the Twitter network, related to the spread of the Higgs boson news, has revealed information about the structure of an online social network. Specifically, it has been found that the degree distribution closely follows a power law function. This fact indicates that most users have relatively small number of connections to other users, but that there are a relatively small number of users that posses an extremely large number of connections to other users. These nodes are most likely celebrity accounts that are either prominent scientists or authority figures in the quantum physics world, or possibly official account of organizations involved the discovery, e.g., CERN. The negative assortativity value for the Twitter network also supports this idea since more popular users connect to a large number of nodes that are less connected.

The six sampling algorithms used to construct prototypic networks have offered a wide range of graph topologies. Specifically, incident and induced sampling

methods have failed to produce networks with enough connectivity between sampled nodes. Snowball sampling has provided graphs with clustering coefficients higher than the observed Twitter network. While an improvement from incident and subgraph sampling, the dense clusters of nodes produced by snowball sampling are disjoint from one another presenting obvious problems when attempting to model epidemic processes. Barabasi-Albert preferential attachment and power law graph models have created graphs that were more representative, based on NMSE characteristic analysis, than previous three methods. However, the MHRW algorithm has managed to produce the most representative prototypic network of the six sampling algorithms both in terms of connectivity and cohesion of the nodes.

The network-based SI epidemic process has been modeled over the MHRW graph, the Barabasi-Albert preferential attachment graph, and the power-law graph. The process that best fit the actual spread of information over time is the epidemic model on the MHRW graph with $\beta = .056$. Evidently, MHRW has been the only graph sampling process that exhibited a sharp increase in the proportion of nodes informed, in a way similar to the actual information diffusion over a seven-day period due to the characteristic similarities of the sample graph and the actual graph, namely the clustering coefficient, the assortativity, and the density.

5.2 Future Work

It is worth noting that the epidemic process modeled in this thesis only accounts for retweet behavior, but not the mentioning or replying behavior of users. Replies are more of a conversation between users and do not necessarily represent the spreading of information from one user to another. The reply network may, however, be used as to predict links between users to identify if users are more

likely to respond to the tweets of people with whom they have common interests or friends. Users have the ability to mention another Twitter user in a posting. This can also be thought of as a means of spreading information to other Twitter users. However, the method by which the information is spread is fundamentally different from retweeting because users do not decide to be mentioned in the same way that they decide to retweet another user's tweet. Modeling this behavior in conjunction with the retweeting behavior would involve creating a new stochastic process that incorporates two transmission rates and two neighborhoods.

$$\mathbb{P}(\mathbf{X}(t + \delta t) = \mathbf{x}' | \mathbf{X}(t) = \mathbf{x}) \approx \begin{cases} [\beta_1 M_i(\mathbf{x}) + \beta_2 M_i'(\mathbf{x})] \delta t, & \text{if } x_i = 0 \text{ and } x'_i = 1 \\ 1 - [\beta_1 M_i(\mathbf{x}) + \beta_2 M_i'(\mathbf{x})] \delta t, & \text{if } x_i = x'_i \text{ and } x_i \in \{0, 1\} \end{cases}, \quad (13)$$

where $M_i'(\mathbf{x})$ is another type of neighborhood that can represent the group of people likely to be mentioned by an informed person. This type of neighborhood could be determined using community detection methods that identify communities of nodes [1]. These communities could serve as a likely subset of nodes that could be mentioned by a particular user.

Varying transmission rates over time could also be incorporated into the epidemic model. The transmission rate, β , was a fixed parameter in this particular analysis, but varying the transmission rate by day could produce a better fitting model.

Information diffusion models on Twitter could also be applied to news that isn't scientific, e.g., political sotries, celebrity gossip, sports news, etc. [2]. Twitter offers a very wide array of news stories and these types of stories, and the people who discuss them, may differ greatly from the dynamics of action surrounding a scientific discovery.

List of References

- [1] C. M. A. Clauset, M. Newman, “Finding community structure in very large networks,” *Physical Review*, 70, p. 66111, 2004.
- [2] M. G. RODRIGUEZ, J. LESKOVEC, D. BALDUZZI, and B. SCHLKOPF, “Uncovering the structure and temporal dynamics of information propagation,” *Network Science*, vol. 2, pp. 26–65, 4 2014.

BIBLIOGRAPHY

- “The history of social networking,” <http://www.digitaltrends.com/features/the-history-of-social-networking/>, accessed: 2015-06-24.
- “Stanford network analysis project,” <http://snap.stanford.edu/data/index.html>, accessed: 2014-09-12.
- A. Clauset, M. Newman, C. M., “Finding community structure in very large networks,” *Physical Review*, 70, p. 66111, 2004.
- Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J., “Everyone’s an influencer: Quantifying influence on twitter,” in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM ’11. New York, NY, USA: ACM, 2011, pp. 65–74.
- Bauer, F., van den Driessche, P., and Wu, J., *Mathematical Epidemiology*. New York, NY, United States of America: Springer, 2008.
- Domenico, M. D., Lima, A., *et al.*, “Anatomy of a scientific rumor,” *Scientific Reports*, no. 4.
- Ebbes, P., Thadakamalla, H., *et al.*, “Sampling large-scale networks: Insights from simulated networks.”
- Fu, F., Liu, L., and Wang, L., “Empirical analysis of online social networks in the age of web 2.0,” *Physica A*, vol. 387, pp. 676–682, 2008.
- Goel, S., Anderson, A., Hofman, J., and Watts, D., “The structural virality of online diffusion.”
- Guille, A., Hacid, H., Favre, C., and Zighed, D. A., “Information diffusion in online social networks: A survey,” *SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, July 2013.
- Hu, P. and Lau, W. C., “A survey and taxonomy of graph sampling,” *CoRR*, vol. abs/1308.5865, 2013.
- Jin, F., Dougherty, E., Saraf, P., Cao, Y., and Ramakrishnan, N., “Epidemiological modeling of news and rumors on twitter,” in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD ’13. New York, NY, USA: ACM, 2013, pp. 8:1–8:9.
- Kephart, J. O. and White, S. R., “Directed-graph epidemiological models of computer viruses.” in *IEEE Symposium on Security and Privacy*, 1991, pp. 343–361.

- Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY, United States of America: Springer, 2009.
- Kwak, H., Lee, C., Park, H., and Moon, S., “What is twitter, a social network or a news media?” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 591–600.
- M Newman, S Forrest, J. B., “Email networks and the spread of computer viruses,” *Physical Review*, 66, 2002.
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B., “Measurement and analysis of online social networks,” in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC ’07. New York, NY, USA: ACM, 2007, pp. 29–42.
- Moreno, Y., Pastor-satorras, R., and Vespignani, R., “Epidemic outbreaks in complex heterogeneous networks,” *Eur. Phys. J. B*, pp. 521–529, 2002.
- Nekovee, M., Moreno, Y., Bianconi, G., and Marsili, M., “Theory of rumour spreading in complex social networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457 – 470, 2007.
- Pastor-Satorras, R. and Vespignani, A., “Epidemic spreading in scale-free networks,” *Physics Review Letters*, vol. 86, no. 14, april 2001.
- Ribeiro, B. F., Wang, P., Murai, F., and Towsley, D., “Sampling directed graphs with random walks.” pp. 1692–1700, 2012.
- RODRIGUEZ, M. G., LESKOVEC, J., BALDUZZI, D., and SCHLKOPF, B., “Uncovering the structure and temporal dynamics of information propagation,” *Network Science*, vol. 2, pp. 26–65, 4 2014.
- Stumpf, M. P. H., Wiuf, C., and May, R. M., “Subnets of scale-free networks are not scale-free: Sampling properties of networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4221–4224, 2005.
- Taxidou, I. and Fischer, P. M., “Online analysis of information diffusion in twitter,” in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, ser. WWW Companion ’14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 1313–1318.
- Travers, J., Milgram, S., Travers, J., and Milgram, S., “An experimental study of the small world problem,” *Sociometry*, vol. 32, pp. 425–443, 1969.

- Wang, T., Chen, Y., and others”, “Understanding graph sampling algorithms for social network analysis,” *Distributed Computing Systems Workshops (ICDCSW)*, 2011 31st International Conference, pp. 123–128, 2011.
- Wang, T., Chen, Y., Zhang, Z., Sun, P., Deng, B., and Li, X., “Unbiased sampling in directed social graph,” *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 401–402, Aug. 2010.
- Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., and Su, Z., “Understanding retweeting behaviors in social networks,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’10. New York, NY, USA: ACM, 2010, pp. 1633–1636.