MODELING THE SPREAD OF INFORMATION ON TWITTER

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Mathematics

By

Darian Ignatius

2018

SIGNATURE PAGE

THESIS:	MODELING THE SPREAD OF INFORMATION ON TWITTER
AUTHOR:	Darian Ignatius
DATE SUBMITTED:	Summer 2018

Department of Mathematics and Statistics

Dr. Hubertus von Bremen
Thesis Committee Chair
Mathematics & Statistics

Dr. Jillian Cannons
Mathematics & Statistics

Dr. Jennifer Switkes
Mathematics & Statistics

ACKNOWLEDGMENTS

I would like to thank Dr. Von Bremen for his support and guidance, not only on this thesis, but as an advisor both in and out of the classroom. I would also like to thank the entire faculty of the Department of Mathematics, as they have all been absolutely amazing and have helped me throughout my time here. Additionally, I would like to thank Dr. U.J. Fan and Lita Patel of the Mechanical Engineering Department as they have been great mentors from the start of my journey at CPP. Finally, thank you to all of my friends and family, as their continued support pushes me everyday and motivates me to produce my best work.

ABSTRACT

Compartmental models have been used in epidemiology for many years to study the spread of infectious diseases throughout the world. In this thesis, we are recreating and extending the work done by others in [1] to apply one of these models, the *SEIZ* model, to the spread of news and rumors on Twitter. After deriving the model and discussing its background, we obtained data regarding 6 events, 3 real news stories and 3 rumors. We showed that the method used to minimize the error between the model and the actual data was quite accurate, and that the model was able to work very early on in a story or with limited information. We also attempted to find several combinations of parameters which could distinguish the stories between news and rumors, but no consistent results were found. Finally, we restricted the amount of data fed into the model, and took a look at its ability to estimate the number of tweets in the future.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the increasing use of social media by groups of all ages, platforms such as Twitter, Facebook, Instagram, and Snapchat have become primary sources of news for many people. The initial exposure some people have to a news story can come from many sources like major news networks, articles on the internet, or original posts on one of these platforms. From there, these stories spread among the population quite rapidly through social media. It is estimated that over 80% of Americans have accounts on social media [2]. This allows a large portion of the population to be constantly updated on current events from anywhere, and this efficient spread of information can be quite beneficial. Emergencies such as amber alerts or active shootings can reach millions of people within minutes, with many of them being in the affected location. People can quickly take appropriate action without having to wait to be informed by official news reports. However, this rapid spread of news can also be harmful when false information breaks out in a similar manner. By analyzing the spread of information on social media and trying to understand its patterns, we can look at ways to get news out to the world most efficiently while also figuring out methods to mitigate or even stop the propagation of rumors.

Compared with other social media platforms, Twitter is the ideal network for obtaining and analyzing data for information propagation. From about 18 million active users in 2009 to over 330 million active users in 2018, the network Twitter reaches has grown substantially [3]. Although Twitter may not be as big as some of its competitors, its posts are fairly simple and the interactions between users contribute heavily to the spread of information. Instead of just having posts, likes, and comments, Twitter also has "retweets" and replies, which can extend the life of a post much further than the original users network. In Table 1.1, we define some of the terms and keywords used with Twitter, which will be referred to for the remainder of this thesis.

Table 1.1: Terminology Used in Twitter Data

Term	Meaning
Tweet	Any post on Twitter, containing text, pictures, videos, or website links
Retweet	A user reposting another tweet while giving credit to the original post's user
Hashtag	Keyword(s) used to associate a tweet with a specific topic
Likes	A user can save a tweet to their own page, which allows them and others to view it later

Twitter began in 2006, and reached one billion posted tweets by May 2009. Today, an average of one billion tweets are posted just every two days [4]! Though information does spread with other social media platforms, it is much easier to track the network reached with Twitter and follow the timeline for which this all occurs. The usage of social media was drastically different between 2009 and now, but data from 2009 will be used for the analysis in this thesis. The second half of 2009 displayed by far the largest rate of growth of Twitter, with an increase in tweets per day of nearly 1000% displayed over this period [4]. Restrictions on data acquisition by Twitter did not allow more recent tweets to be obtained, but ideally the models would work just as well, if not better, with more current data.

We can apply compartmental models from epidemiology to the information on Twitter to get a better understanding of its propagation. These models break the population down into compartments, and then each compartment has a probability of transitioning from one to another. The "population" is anyone who can be affected by the disease being studied, or for our, any tweet related to the story. The simplest of these models is the *SI* model, with the compartments being Susceptible, S, and Infected, I. All members of the population belong to one of these compartments. This very simple version is not practical, however, as typically diseases are spread with much more complexity. A more realistic model is *SIS*, where we have the same compartments, but members transition from Susceptible to Infected and back to Susceptible. This is attractive for modeling things such as the flu or the cold. There are a variety of other compartments we can use to create different models. The *SEIZ* model is one that has already been analyzed in [1] by Jin et al., and it has been shown to work quite well with analyzing the spread of news on Twitter. In this model, our compartments are Susceptible, S, Exposed, E, Infected, I, and Skeptic, Z.

The *SEIZ* model has already been applied in [1] to several Twitter data sets and shown it is quite effective in modeling the spread of news on Twitter. It has also been shown that within the model's parameters, there is potential to show a distinction between real news stories and rumors. The purpose of this thesis is to look at more data sets to verify the model's ability, look at the distinction between real and fake news in these parameters, and attempt to use this to detect rumors early on in their propagation. We apply the *SEIZ* model to look at the propagation of tweet numbers for a story, rather than the users themselves. We also take a look at the ability of the model to predict the number of tweets over a longer duration than the model has seen. In the coming chapters, we will discuss the stories used in our Twitter data sets and how Python was used to search and filter larger data sets for the desired tweets. Then, we will go into detail on the Matlab scripts used to optimize the *SEIZ* model's parameters against the data. The results will be analyzed for each story and then the applications of our findings will be talked about. Before discussing the implementation of Matlab and the results of the *SEIZ* model, we will go through the background of derivation of the model in the next chapter.

Chapter 2

Background

Compartmental models are a mathematical approach used to evaluate and predict the spread of various infectious diseases [14]. This is done by breaking the population down into distinct compartments and establishing parameters for the rates at which the population transitions between compartments. These parameters are obtained by looking at the relationships between each class of the population and making assumptions about the disease. With this, we can generate a set of differential equations to make predictions about the spread of the disease. A basic approach, which is more practical than the SI model, would have the population divided into three compartments: Susceptibles, S, who are the individuals at risk to the disease; Infected, I, which are those who have the disease and are capable of transmitting it; and Removed, R, which are individuals who can no longer be infected or infect anyone else, so they either died from the disease or they have recovered and are now immune [5]. A great application of this model would be to a disease like measles, where once recovered, an individual has immunity for life. One of the assumptions made in this Susceptible-Infected-Removed, or SIR, model, is that the overall population, N, is constant. So we would have

$$\frac{dN}{dt} = 0, \tag{2.1}$$

$$N = S + I + R. \tag{2.2}$$

If we assume that individuals only transition from S to I with a contact rate of $r > 0$ and from I to R with a transition rate of $a > 0$, we get the following relationship, as shown in Figure 2.1:

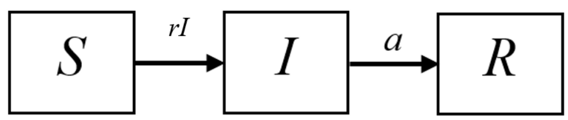

Figure 2.1: Transition Rates of *SIR* Model

These relationships may then be used to generate a system of ODEs:

$$\begin{aligned} \frac{dS}{dt} &= -rS\frac{I}{N} \\ \frac{dI}{dt} &= rS\frac{I}{N} - aI \\ \frac{dR}{dt} &= aI. \end{aligned} \tag{2.3}$$

The rate at which individuals transition from S to I is dependent on r as well as the number of infected individuals. The transition rate is dependent on I because someone can only become infected if they come into contact with an infected individual. The transition rate from I to R, however, is only dependent on a. The transition for each individual out of the Infected compartment is independent of everyone else for this model.

Another simple model which can be used is the Susceptible-Infected-Susceptible, or *SIS*, model. In this case, it is assumed that instead of dying or becoming immune, the infected individuals transition back into the susceptible class. It can be seen how this model could be used for diseases such as gonorrhea, because individuals do not gain immunity after recovering from infection. Though this model is not perfect, its simplicity can give us a good idea of what is going on. If we assume that individuals transition from S to I with a contact rate of $r > 0$ and back from I to S with a transition rate of $\lambda > 0$, we get Figure 2.2.

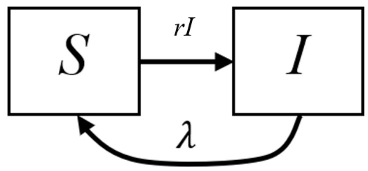

Figure 2.2: Transition Rates of *SIS* Model

With these relationships, the following system of ODEs can be generated:

$$\begin{aligned} \frac{dS}{dt} &= -rS\frac{I}{N} + \lambda I. \\ \frac{dI}{dt} &= rS\frac{I}{N} - \lambda I \end{aligned} \tag{2.4}$$

Similarly to the *SIR* model, the transition rate out of I is independent of everything except the parameter, λ. Also, the transition rate from S to I is dependent on the parameter, r, as well as the number of people in the Infected compartment.

In the *SIR* and *SIS* models above, we are assuming that once a susceptible individual contracts the disease, they immediately transition to the infected class. This is not generally the case, however, as many diseases take time to manifest during an incubation period. For these cases, an Exposed compartment, E, is introduced to denote those who have come in contact with an infected individual, but have yet to become infected themselves. From this, we can get the *SEIS* and *SEIR* models. For the purpose of information propagation, Bettencourt et al. [6] proposed the *SEIZ* model, where we have a Skeptics compartment, Z. In this model, the skeptics are those that become immune to the disease. Though it is similar to the Removed compartment, the skeptics transition directly from the susceptibles and their interaction can still affect other compartments.

With a fourth compartment added to the model, there is an increased layer of complexity in transitioning from one compartment to the next. In Figure 2.3, we have the diagram illustrating the relationship between each compartment:

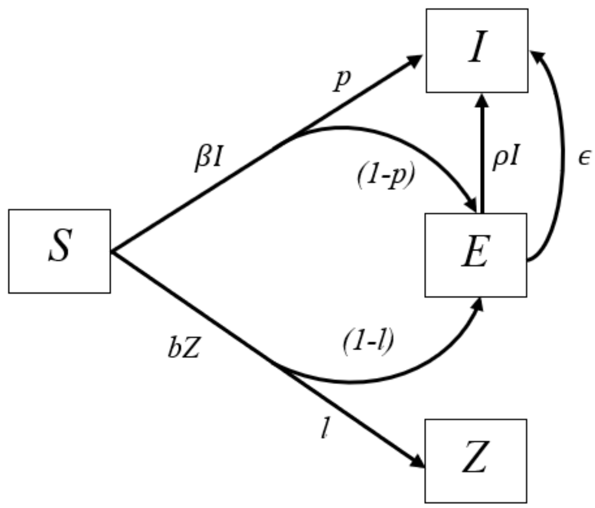

Figure 2.3: Transition Rates of *SEIZ* Model

Then in Table 2.1, we have a description of each parameter of this model. This gives a more intuitive look into the model and relates the relationships above with the actual equations:

Table 2.1: Description of Parameters in *SEIZ* Model

Parameter	Meaning	Units
β	Rate of contact between S and I	Per Unit Time
b	Rate of contact between S and Z	Per Unit Time
ρ	Rate of contact between E and I	Per Unit Time
p	Transition rate of $S \to I$, given contact with I	Unit-less
$1-p$	Transition rate of $S \to E$, given contact with I	Unit-less
l	Transition rate of $S \to Z$, given contact with Z	Unit-less
$1-l$	Transition rate of $S \to E$, given contact with Z	Unit-less
ε	Transition Rate of $E \to I$	Per Unit Time

With the relationships between each compartment described by the parameters above, we have the following set of ODEs:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta S\frac{I}{N} - bS\frac{Z}{N} \\
\frac{dE}{dt} &= (1-p)\beta S\frac{I}{N} + (1-l)bS\frac{Z}{N} - \rho E\frac{I}{N} - \varepsilon E \\
\frac{dI}{dt} &= p\beta S\frac{I}{N} + \rho E\frac{I}{N} + \varepsilon E \\
\frac{dZ}{dt} &= lbS\frac{Z}{N}
\end{aligned} \tag{2.5}$$

Jin et al. [1] studied the *SEIZ* and *SIS* models for the spread of news on Twitter and found the *SEIZ* model produced much better results. When the models were optimized against Twitter data from 4 news stories and 4 rumors, it was found that the *SEIZ* model consistently had a much lower relative error than the *SIS* model. Also, they looked at a ratio of parameters from the model, which was the ratio of transition rates entering the Exposed compartment to those leaving this compartment. The ratio, which we will call the Exposed ratio, is given by

$$R_{SI} = \frac{(1-p)\beta + (1-l)b}{\rho + \varepsilon}. \tag{2.6}$$

The Exposed ratio may be able to indicate a difference between news stories and rumors. A value greater than 1 for R_{SI} means that the transition rates from S to E were greater than those from E to I. If the ratio is less than 1, than the exit rates are greater than the entry rates. For the cases studied in [1], it was shown that the news stories typically had $R_{SI} > 1$ and the rumors had $R_{SI} < 1$. There was one borderline case for the news stories, however this particular case had elements of a rumor associated with it. We have obtained data from Twitter on our own stories, and will be applying the $SEIZ$ model to these data sets to further verify its low error and the ability of R_{SI} to differentiate between news and rumors. When applying this model to a specific news story, we have the following compartments: the Susceptibles relate to individuals who have not encountered a tweet related to the story; the Exposed relate to those who have encountered such a tweet, and have a delay before retweeting or posting a tweet themselves; the Infected relate to those who have tweeted about the story; the Skeptics relate to those who have chosen not to tweet about a story at all, but can influence the Susceptibles into not believing a story. For the purpose of this thesis, the number of infected individuals refers to the total number of tweets. The other compartments, S, E, and Z, simply aid us in obtaining an accurate fit for I, and thus may not translate directly to a real-world application.

Chapter 3

Case Studies

From the limited data set of tweets provided from [7], stories were chosen based on the date of the event and the ability of the story to appear and spread on social media. For example, an event such as the outcome of the Superbowl would not be the ideal case to model. Since most people watch the Superbowl live, there would be a massive spike in tweets as soon as the news broke out, rather than a propagation of tweets similar to the infections of a disease. The data set provided contains tweets from June 1, 2009 to December 31, 2009. Several cases of real news stories and rumors which gained traction were analyzed, with one story having as few as 2,000 tweets, and the largest having over 150,000 tweets. Table 3.1 gives a look at some of the important numbers for each event, along with criteria used to identify relevant tweets such as keywords and date restrictions. In this table we also include the time span, which is the number of days a story remained active.

Table 3.1: Summary of Data Sets

Event	Date	Duration	Tweets	Type	Keywords
Death of Michael Jackson	06-25-2009	5 days	150,000	News	Michael Jackson, #MichaelJackson
Fort Hood Shooting	11-05-2009	5 days	20,000	News	Fort Hood, #FortHood
Tiger Woods Car Accident	11-27-2009	2 days	10,000	News	Tiger Woods, #TigerWoods
Death of Britney Spears	06-28-2009	2 days	1,500	Rumor	Britney Spears, #BritneySpears
Kanye West Car Accident	10-21-2009	4 days	20,000	Rumor	Kanye West, #KanyeWest
Sarah Palin Divorce	08-01-2009	2 days	10,000	Rumor	Sarah Palin, #SarahPalin

3.1 News Stories

Here we will discuss the 3 cases of real news. The stories chosen were the death of Michael Jackson, the 2009 shooting at Fort Hood, and a car accident involving Tiger Woods.

3.1.1 Death of Michael Jackson

Michael Jackson passed away after suffering from cardiac arrest on Thursday, June 25, 2009. The paramedics were called to his home at 12:21 P.M. PST, and he was pronounced dead at the hospital at 2:26 P.M. PST. The news broke out on Twitter almost immediately, with mentions of his name jumping from about 2 times per minute to over 200 per minute within the first hour. Over the next 5 days, nearly 150,000 tweets were posted mentioning his name, with nearly half of those tweets coming in the first 24 hours [8].

3.1.2 Fort Hood Shooting

On November 5, 2009 at 1:34 P.M. CST, a mass shooting occurred at Fort Hood, a military base in Texas. Over 40 people were either killed or injured in this tragedy. There were less than 10 mentions per day of Fort Hood or the gunman on Twitter, but when the news broke out it jumped to over 1,000 mentions per hour. This continued for several hours, as the total mentions reached about 15,000 in the first day [9].

3.1.3 Tiger Woods Car Accident

On November 27, 2009 around 2:30 A.M. EST, Tiger Woods crashed his car near his house. Although the cause of the accident wasn't released, there was already news circulating that Woods may be cheating on his wife. This news along with the suspicious car accident made this news story gain traction quickly. Similarly to Michael Jackson, there was very little activity on Twitter regarding Woods until the accident, where mentions of his name jumped up to several thousand per hour [10].

3.2 Rumors

In this section we discuss the 3 cases of rumors. The stories used here were the death of Britney Spears, Sarah Palin's divorce, and a deadly car accident involving Kanye West. Fake tweets or articles came out which caused these rumors to spread, and they would slow down once official statements were released to deny them.

3.2.1 Death of Britney Spears

After the death of Michael Jackson, many rumors began to spread about the deaths of other celebrities. On June 28, 2009, Britney Spears' Twitter account was hacked, and someone posted a tweet stating that she had passed. This caused a minor spike in mentions of the singer for the next few days. Her representatives quickly posted another tweet revealing that this was untrue, but the rumor still spread with several thousand tweets being posted about the event [11].

3.2.2 Kanye West Car Accident

On October 20, 2009, articles were released saying Kanye West had died in a big car accident. In the article, it was reported that either West or the other driver was intoxicated. There was already drama surrounding West at this time, so this story was believable and many people would spread the information without checking sources. Nearly 20,000 tweets mentioning Kanye West were posted within 2 days of this story breaking out [12].

3.2.3 Sarah Palin Divorce

Articles had come out on August 1, 2009 that Sarah Palin would be getting a divorce from her husband. Just a few days earlier, she had stepped down as governor of Alaska. Again, the recent drama around Palin allowed this story to be more believable and spread. Palin went from being barely mentioned on Twitter to having hundreds of tweets per hour. She released a statement later on that the reports were false, but similarly to the other rumors, it had already spread and took some time before settling down [13].

Chapter 4

Approach

A data set of approximately 467 million tweets was provided for this thesis. These tweets ranged from June 2009 to December 2009, and they are estimated to be about 20-30% of all tweets during this time. This was because the data set contained only tweets posted from public accounts and also did not contain tweets from every language. We used these tweets to gather total tweet counts, which is the total number of tweets and retweets, for each story. Then we fit the Infected compartment of the $SEIZ$ model, I, to the tweet counts. Once the fit is complete, we obtain each compartment as a function of time ($S(t), E(t), I(t)$, and $Z(t)$).

4.1 Gathering and Filtering the Data

The Stanford Network Analysis Project (SNAP) provided the 467 million tweets that were used [7]. A script was written in Python (shown in Appendix B) to extract data for each story from this large data set. The data set presented each tweet in three lines, and was given in the following form:

T: Date and time in UTC

U: The URL linking to the user's account

C: The content of the tweet

The script reads through every line of the file, and stores these three lines in a temporary list. Then, we check if tweet contains any of our desired keywords. If the tweet matches our keywords and was posted in the appropriate time frame, we then take the difference in time between the tweet and the start of the story and save it to a new file. We then keep track of the number of these desired tweets over time. A cumulative total is taken every minute, 15 minutes, or hour. Once the time frame for our story has passed, we break out of the loop, save the file, and proceed to Matlab for the analysis.

4.2 Fitting the Model

The scripts used for fitting the model and minimizing the error between $I(t)$ and the actual number of tweets were given by Jin et al. [1]. There were three functions written and small modifications were made to fit this problem. Each of the functions lists the parameters as global variables, so that they may be given values later on. Function 1 defines some of the variables and the system of ODEs, given by Equation 2.5. Function 2 integrates the Equation 2.5 using the Forward Euler Method, and computes the error between $I(t)$ and the actual number of tweets. The inputs for Function 2 are the parameters and initial conditions we want optimized. Function 3 gathers the data, minimizes the error produced by Function 2, computes the error and R_{SI}, and plots the results. This process is repeated for each story and for different periods of time. These scripts are all shown in appendices A.1-A.3.

4.2.1 Function 1 - ODEs

Here, some of the variables and the system of ODEs are defined. The inputs for the function are the initial conditions of each compartment (S, E, I, Z), and we define the total population as $N = S + E + I + Z$. Then, we have the system of ODEs, Equation 2.5. The initial conditions and total population are unknown, so they are found in the next two functions to produce the most accurate model.

4.2.2 Function 2 - Computing the Model's Error

In this function, we input the initial conditions for each compartment and the values of each parameter. The optimization function in Function 3 obtains these values for us. These values are then used in the Forward Euler Method to approximate the solution to the system of ODEs given in Function 1. Although the built-in Matlab function **ode45** can be used, this method is much more efficient and produced similar results for several test cases. The Forward Euler Method is discussed below:

Suppose we have a system of ODEs, $y'(t) = F(y(t))$. Then, we can approximate the solution with the first two terms of the Taylor series using an initial guess $f(t_0, y_0)$ and a time step Δt:

$$y(t_n + \Delta t) = y(t_n) + \Delta t * F(y(t_n)) \tag{4.1}$$

Equation 4.1 can then be used iteratively over a specified time span to give us an approximate solution to the system.

Once we have an approximate solution to the *SEIZ* model, we compute the error between the actual number of tweets (our data) and the estimated number of tweets, $I(t)$. In this function, the error is defined as

$$Err_{abs} = |I(t) - tweets(t)|. \tag{4.2}$$

This error is output as a vector which we then aim to minimize in Function 3.

4.2.3 Function 3 - Optimization of Parameters

Here, we extract and analyze the data, optimize the model, and plot our results. First we open and read the desired file, separate the data into two columns (time and tweet count), and then place this information into arrays. These data give the cumulative total of tweets from the start of the event, either per hour or 15 minutes. Next, we take the final number of tweets, and use this value to determine the upper and lower bounds of the parameters we look to optimize. The final number of tweets will be referred to as I_{total}, and will be discussed in the next chapter. The desired parameters are the initial states of each compartment, $S_0, E_0, I_0,$ and Z_0, and the parameters specified in Table 2.1, $\beta, b, \rho, p, l,$ and ε.

The goal of optimizing these parameters is to minimize Equation 4.2. Some of the constraints to this minimization problem are that all of the parameters from Table 2.1 must be non-negative. Thus, we have

$$\beta, b, \rho, \varepsilon \geq 0 \text{ and } 0 \leq p, l \leq 1.$$

Similarly the compartments must also be non-negative. Upper bounds were added to the compartments' initial conditions as well, as without them the model would produce unrealistic results. The upper bounds chosen for this thesis are shown in Appendix A.3. These were the values that produced the most stable results, but there are most likely other bounds which will give results that are similar or potentially even more stable. Since the initial conditions can take on a range of values, the total population, N, is fixed for a single fit, but is not fixed for each event. Changing the constraints or the amount of data fed into the model could change the total population as well. Although the minimization problem itself appears simple, the large number of parameters makes the solution very sensitive and the need to solve a system of ODEs every iteration makes this problem computationally heavy.

We use a built-in Matlab function called **lsqnonlin** to optimize the parameters in this minimization problem. The function uses a trust-region-reflective algorithm to obtain a solution. This function then outputs the optimal values, which we can now use to run the Forward Euler Method once again and obtain the optimal estimation of $I(t)$. We then compute R_{SI} and the final error and save this information. Finally, all 4 compartments of the $SEIZ$ model are placed together on a plot, and another plot is made comparing the estimate, $I(t)$, with the true data, or tweet counts.

A modification was made to this script to fit the model for each case over multiple time spans. The number of hours of data fed into the model was restricted, and once the values of all parameters were found, the model is run again from scratch with more data. We did this to see the optimal parameters for each case as a function of time, and attempt to draw some conclusions from this information.

Another version of this function restricted the amount of data fed into the model, and used this restricted model to predict the future number of tweets. The approximation of the model's solution using the optimal parameters found with restricted data was applied over a longer duration and compared to the actual number of tweets. For example, 24 hours of data may have been fed into the model, and the optimal parameters for this case were found. Then, using these same optimal parameters, we approximate $I(t)$ over 48 hours and compare it to $tweets(t)$ over 48 hours. We then calculate the relative error for the entire model, as well as over just the new batch of data.

Chapter 5

Results and Discussion

The primary goal of this thesis was to recreate the work in [1], and apply it to different data sets. We were successfully able to apply the model and display the results for 3 news stories and 3 rumors. We then proceeded to look at the Exposed ratio, R_{SI}, proposed in [1] and found that we were unable to draw any conclusive results from it. Next, we proposed our own combinations of parameters for differentiating between news and rumors. Again the results appear to be inconclusive, but there was potential shown in these values. We then attempted to stabilize the parameters over time by reducing the complexity of the optimization problem. Finally, we looked at the ability of the model to predict future numbers of tweets and discussed the applications of these results.

5.1 Accuracy of $I(t)$

In Table 5.1 we have the relative error for each story between $I(t)$ and the total tweet counts. The durations and time intervals chosen below were those that best captured the data for each story, while keeping all the durations relatively close.

Table 5.1: Here we have a brief summary of fitting the model for each case. The duration is the length of time given for each case, the time interval refers to how often the tweet count is obtained, and the error is given in Equation 5.1.

Event	Type	Duration	Time Interval	Error
Michael Jackson	News	2 days	15 minutes	0.88%
Fort Hood	News	2 days	1 hour	1.57%
Tiger Woods	News	36 hours	15 minutes	3.57%
Britney Spears	Rumor	2 days	15 minutes	1.47%
Kanye West	Rumor	2 days	1 hour	1.18%
Sarah Palin	Rumor	1 days	15 minutes	1.44%

The relative error was calculated by

$$Err_{rel} = \frac{||I(t) - tweets(t)||}{||tweets(t)||} * 100, \tag{5.1}$$

where $tweets(t)$ is the number of cumulative tweets taken at every time interval, and $I(t)$ is the infected compartment of the *SEIZ* model evaluated at each of those time intervals. These two variables end up as vectors of the same length, so we take the 2-norm of their difference, divide by the 2-norm of the data vector, and multiply by 100 to get the error as a percentage. This error indicates how far the model deviates from the actual number of tweets, relative to the values in our data.

With these results, the accuracy of this model is further verified as we have a largest relative error of about 3.57%. This means that for a story with 10,000 tweets, we should be able to estimate this number within about 350 tweets. Below we have comparisons between the model and the actual data for the Michael Jackson and Tiger Woods stories, which had the lowest and highest errors, respectively.

It can be seen seen from Figures 5.1 and 5.2 that even with the slightly larger error, both cases follow the trend of the data quite closely, and the model accurately measures

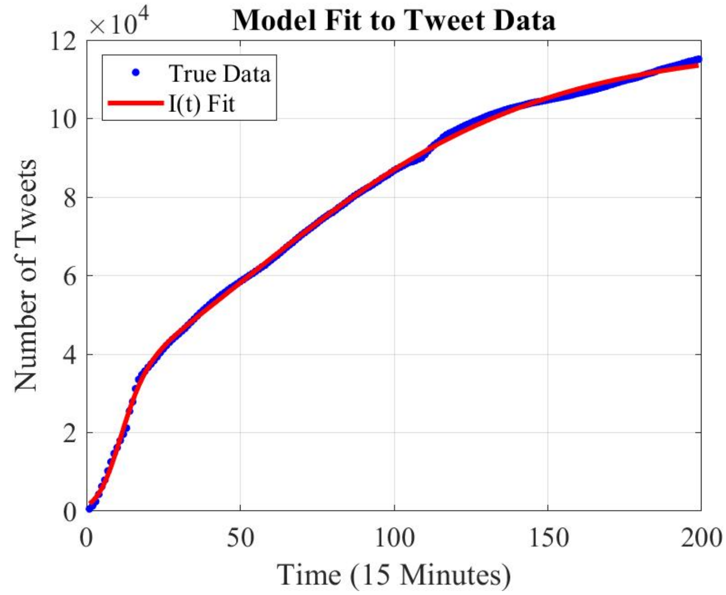

Figure 5.1: Estimation for the Michael Jackson case compared against the actual number of tweets. Here, we obtained the number of tweets every 15 minutes for 2 days.

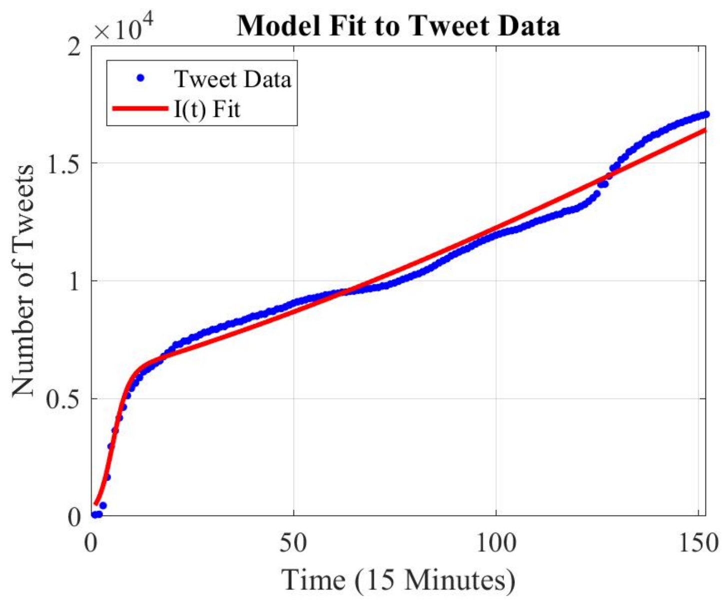

Figure 5.2: Estimation for the Tiger Woods case compared against the actual number of tweets. We have data every 15 minutes for about 36 hours.

the spread of information on Twitter. This was shown to be true both for real news stories and the rumors. Although data was given every 15 minutes in these cases, it was observed that the model successfully kept a low error even when data was given every hour or even 2 hours.

5.2 Rumor Detection

The next step was to examine the model's ability to differentiate between real news stories and rumors. In [1], the Exposed ratio, R_{SI}, showed potential to provide some distinction. The news stories had $R_{SI} > 1$, the rumors had $R_{SI} < 1$, and the case with both elements of truth and rumors had $R_{SI} < 1$, but was closer to 1 than the rumors. We then looked at all the parameters and proposed two new values to detect rumors.

5.2.1 Exposed Ratio

Although promising results were shown in [1] for R_{SI}, our results were inconclusive when using the cases detailed here. It was observed that no clear distinction could be found in the ratio at first, so we varied the duration of each event and ran the model again in an attempt to find a trend. In Table 5.2 we look at the Exposed ratio for each case on multiple time intervals.

It can be seen that the Exposed ratio does not show a trend, and almost appears to be completely random. It was observed that even if the duration is held constant, slight changes to the constraints would cause these large jumps in the parameters. This could be caused by the nonlinear nature of the model and the number of parameters used in the optimization function. Thus, we have concluded that for these cases, the Exposed ratio either has no significant meaning or there are multiple solutions to the minimization

Table 5.2: Exposed ratio, R_{SI}, for each case over various time intervals. The "Duration" column represents the number of hours from the start of the story, with the cumulative tweet count recorded every 15 minutes.

	Real News			Rumors		
Duration (Hours)	MJ	Fort Hood	Tiger Woods	Britney	Kanye	Palin
10	11.026	89.942	120.69	52.977	42.146	49.232
12	11.624	8.803	103.55	126.13	43.166	4.1261
14	11.994	103.15	41.163	36.905	39.797	33.944
16	12.308	60.655	64.911	114.8	36.695	0.88764
18	10.49	18.43	71.26	98.39	38.58	55.78
20	12.79	17.92	307.3	71.73	31.52	35.34
22	13.75	17.479	302.3	3.2218	44.363	27.786
24	7.5639	13.427	76.489	28.358	56.328	21.809
26	13.608	18.5	294.27	56.103	98.797	23.31
28	106.83	19.194	503.35	8.1636	86.838	26.015
30	13.212	12.602	47.941	54.701	83.567	21.391
32	13.972	15.002	9.929	8.5945	75.074	27.663
34	30.615	20.915	2.2123	7.2421	127.85	10.455
36	98.924	15.199	5.4404	13.437	70.878	24.524
38	146.35	20.211	6.4545	8.8266	29.594	26.575
40	67.809	27.402	2.5858	9.5009	11.861	25.814
42	20.218	31.518	3.2823	13.052	11.593	24.696
44	12.846	312.33	3.5684	5.699	11.657	34.608
46	42.934	246.46	38.316	57.956	9.2003	27.783
48	58.573	103.57	47.362	20.474	14.942	31.22

problem.

In Figures 5.3-5.5, we have plots of the entire *SEIZ* model for the Michael Jackson case. Here, the story is analyzed with data every hour, for 1 day, 2 days, and 5 days. Looking at the different plots as a whole, the S, E, I compartments all maintain the same shape. On the other hand, the impact of the Z compartment is constantly changing. Each of the stories may have a different shape, but the shape is consistent within each story despite the changing R_{SI} values. This can be seen in Figures 5.6 and 5.7 as well, where we see the same happen for the Kanye West rumor. Since all of the stories showed similar patterns, the next step will be to look at parameters outside of the Z compartment.

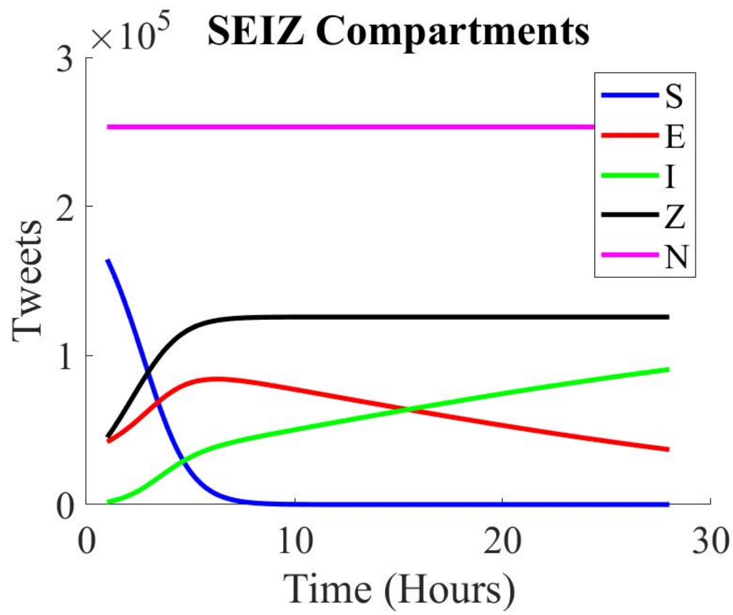

Figure 5.3: Michael Jackson case over 1 day

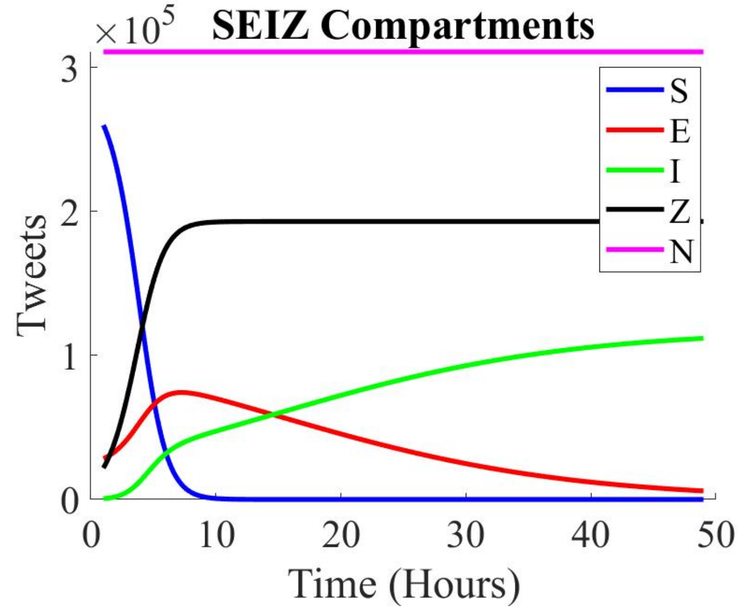

Figure 5.4: Michael Jackson case over 2 days

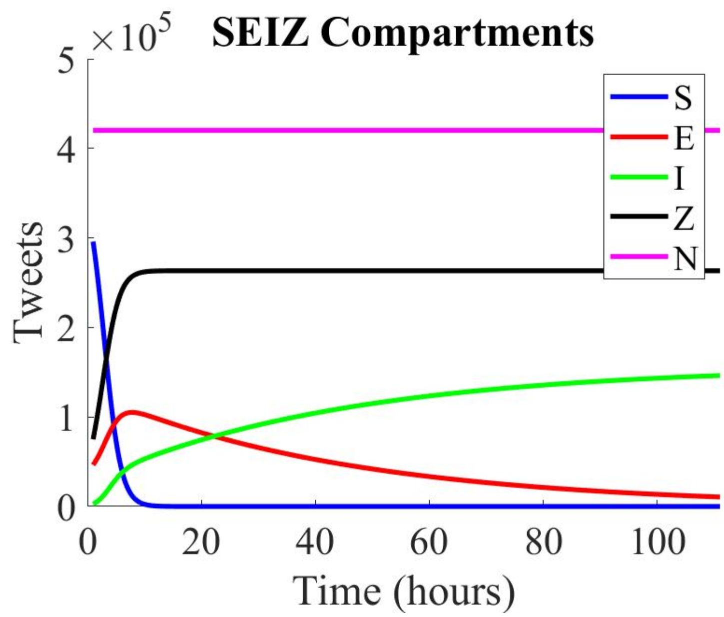

Figure 5.5: Michael Jackson case over 5 days

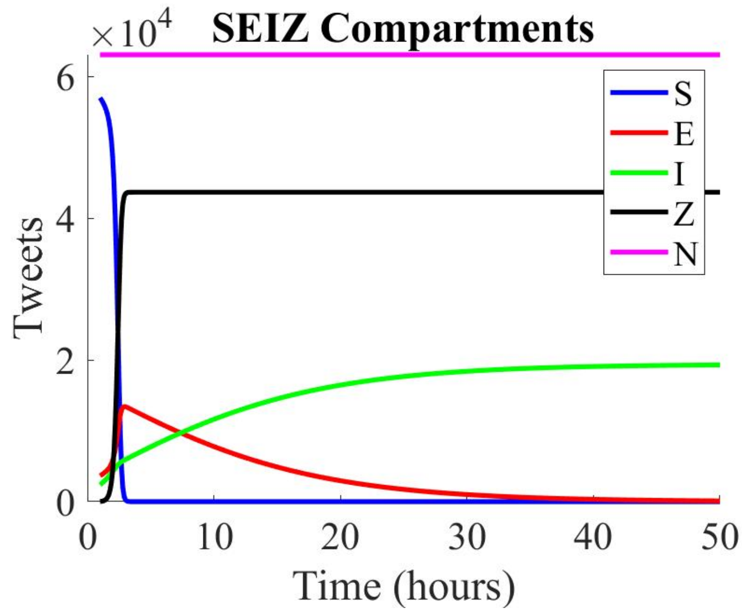

Figure 5.6: Kanye West rumor over 2 days

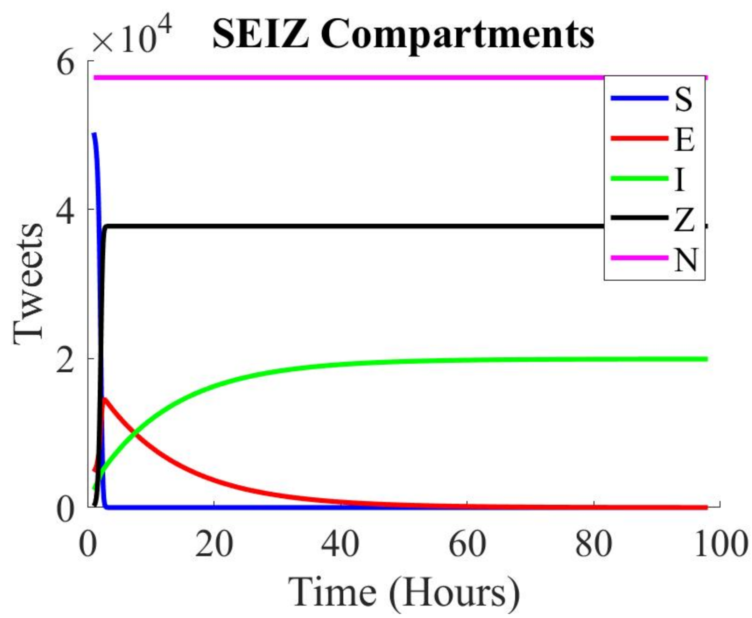

Figure 5.7: Kanye West rumor over 4 days

5.2.2 Other Parameters

In Table 5.3, we have a look at every parameter that our Matlab function optimizes.

Table 5.3: Optimal initial conditions and parameters for the *SEIZ* model. Each case was analyzed with data points every 15 minutes for 2 days in these results. The I_{total} variable refers to the total number of tweets at the end of the 2 days.

	Real News			Rumors		
	MJ	Fort Hood	Tiger Woods	Britney	Kanye	Palin
$\frac{S_0}{I_{total}}$	2.37E+00	2.02E+00	2.07E+00	1.57E+00	1.61E+00	2.09E+00
$\frac{E_0}{I_{total}}$	2.58E-01	2.71E-01	2.46E-01	2.29E-01	3.54E-01	2.11E-01
$\frac{I_0}{I_{total}}$	5.10E-03	8.65E-09	5.63E-02	1.73E-01	5.52E-03	3.50E-10
$\frac{Z_0}{I_{total}}$	1.99E-01	2.61E-01	1.28E-01	2.95E-01	4.99E-01	3.28E-01
β	2.15E+00	1.82E+00	8.49E-01	1.18E+00	5.69E-01	5.12E-01
b	1.06E+00	6.94E+00	9.60E-01	9.00E+00	3.75E+00	9.61E+00
ρ	7.56E-01	3.66E-01	5.49E-01	5.06E-01	5.26E-01	2.44E-02
p	8.05E-01	9.19E-01	7.60E-01	8.43E-01	8.03E-01	7.48E-01
l	2.25E-01	1.59E-01	2.98E-02	2.14E-01	1.05E-01	0.00E+00
ε	2.00E-05	1.49E-02	2.20E-05	5.94E-03	1.43E-02	8.45E-03

Over multiple runs, The two values which showed there may be a distinction between real news stories and rumors are I_0/I_{total}, or the scaled I_0, and ε. In this situation (data at 15 minute intervals for 2 days) and several others, it was observed that for real news cases, the scaled I_0 was typically 10-100 times smaller and ε would be 10-1000 times smaller than the values found for rumors. These two parameters have some relation, as ε is the incubation rate at which members transition from the Exposed compartment to the Infected. As more time spans were analyzed, it became apparent that these values would sometimes jump around, but more often than not least one of the parameters would be able to distinguish between news and rumors. Thus, we computed two simple combinations of the two parameters which could hopefully produce better results. We can define the Infected-Incubation ratio, R_ε, as

$$R_\varepsilon = \left(\frac{I_0}{I_{total}}\right)/\varepsilon. \tag{5.2}$$

We may define the Infected-Incubation number as

$$N_\varepsilon = \left(\frac{I_0}{I_{total}}\right) * \varepsilon. \tag{5.3}$$

For 5 different time spans, it was observed that if $R_\varepsilon > 100$, the case was a real news story, and if $R_\varepsilon < 100$, the case was a rumor. To completely verify this claim though, the model was run every hour, from the beginning of the story until the end of the 2 days. Figures 5.8 and 5.9 display the results of those trials:

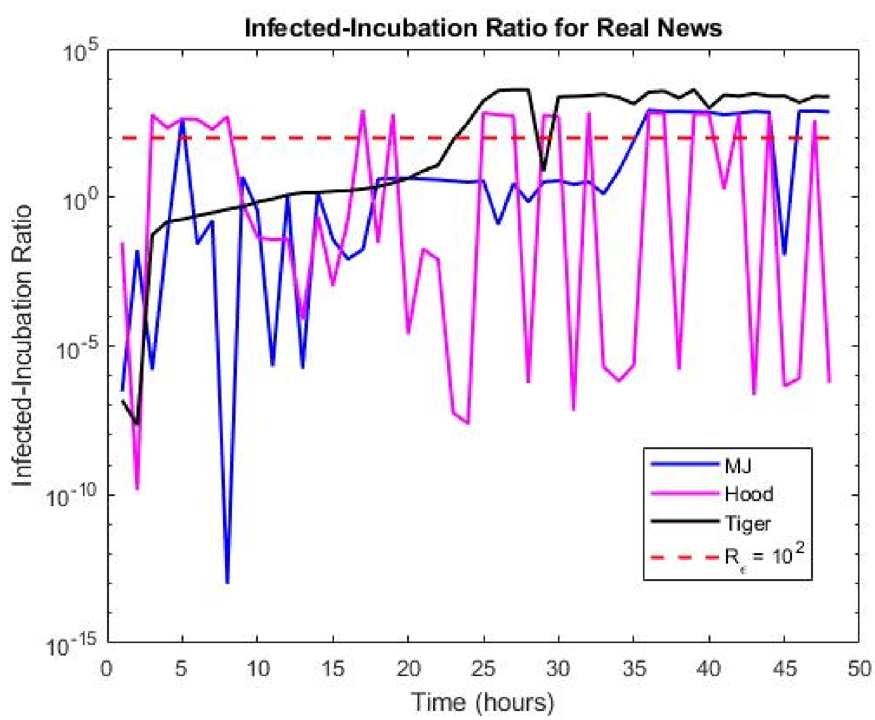

Figure 5.8: R_ε on a log scale for each of the real news stories, run every hour for 2 days. The erratic behavior of this ratio in the Fort Hood case makes it difficult to draw any conclusions.

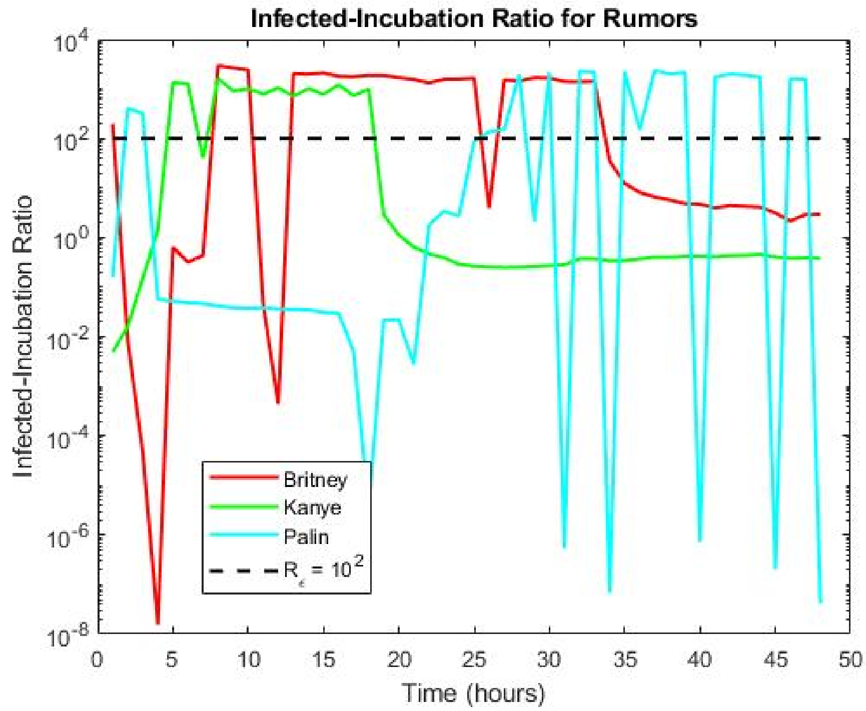

Figure 5.9: R_ε on a log scale for each of the rumors, run every hour for 2 days. Similarly to the previous plot, the behavior of the ratio in the Sarah Palin rumor makes it difficult to draw any conclusions.

We applied a similar analysis to the Infected-Incubation number, hoping to obtain better results. Here, it was observed for 5 different time spans that if $N_\varepsilon < 10^{-5}$ the case was a real news story, and if $N_\varepsilon > 10^{-5}$, the case was a rumor. Figures 5.10 and 5.11 display the results of N_ε over the same trials done for R_ε.

In the analysis of the real news stories, the Infected-Incubation ratio began to settle above 100 for the Michael Jackson and Tiger Woods cases. After about 1 day of the story breaking out, these cases began to relax and produce the consistent results seen in the scaled I_0 and ε parameters. Similarly, both the Kanye West and Britney Spears rumors began to settle well below 100 after about 24 hours worth of data. The limiting cases here were the Fort Hood and Sarah Palin stories. In both cases, R_ε was quite random and would not settle above or below any particular value. When looking at the Infected-Incubation number, N_ε, we had a similar relationship. The news stories started to become stable as the duration increased, with all three trending below 10^{-5}. The same two rumors had a trend above 10^{-5}. The Fort Hood and Sarah Palin cases were again unstable, but the Fort Hood story was ultimately able to stay below the threshold set for N_ε. Though these results are again inconclusive, these proposed quantities may prove useful when looking at an increased number of cases and more recent events.

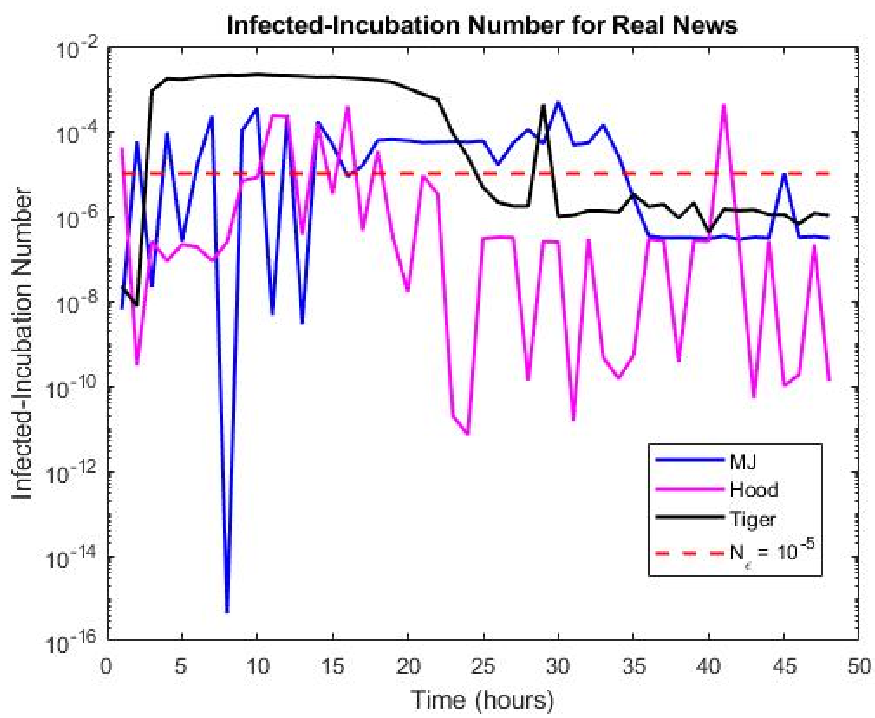

Figure 5.10: N_ε on a log scale for each of the real news stories, run every hour for 2 days. Though the number does not settle for the Fort Hood case, its value does appear to ultimately stay below the threshold.

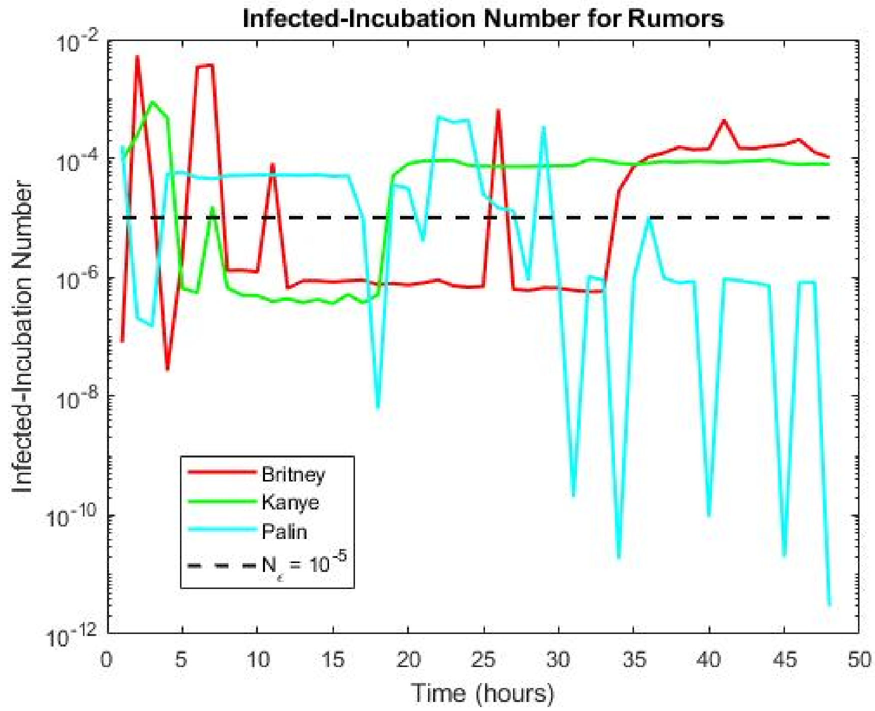

Figure 5.11: N_ε on a log scale for each of the rumors, run every hour for 2 days. When looking at this quantity, the Sarah Palin behaves the opposite to how we had hoped, and does not allow us to draw any conclusions.

5.2.3 Stabilizing the Parameters

It can be seen from the Sections 5.1.1 and 5.1.2 that the optimal parameters were incredibly unstable. Adding slightly more data would cause massive jumps in some parameters, leading to all of these inconclusive results. Thus, we will apply different methods to obtain the initial conditions of the compartments in an attempt to reduce the complexity of the model and stabilize the results when looking at different time spans.

First, we looked at the initial guess of the initial conditions. Rather than using guesses dependent on I_{total}, we used the optimal conditions from previous time spans as the initial guess. For example, we first obtain the optimal initial conditions for S_0, E_0, I_0, and Z_0 for fitting a case over 14 hours. Then, we use those initial conditions as the initial guess for these compartments in the 16 hour fit. This process is repeated every 2 hours over the next 48 hours. In Tables C.1-C.6 (Appendix C), we have the results of this method for each case. We looked at the two ratios and the number discussed earlier, along with the relative error from this method and the old one.

It can be seen that even with the same constraints, the initial guess impacts the optimal parameters as well! The Exposed ratio and the Infected-Incubation ratio still appear to have the same sensitivity as before, even when these initial guesses should be more local to the solution. The Infected-Incubation number did settle a bit more, but at different values than before. This number still did not have any trends from which we could distinguish news and rumors.

Our next approach was to try to stabilize some of the other parameters. We did this by taking the initial conditions of the compartments out of the optimization function itself, reducing the potential number of solutions. In Tables C.7-C.12 (Appendix C), we have the results of all of the optimized parameters, the Exposed ratio, and the relative error. We do not look at R_ε and N_ε, as the scaled I_0 is a fixed value.

The first thing we notice is that fixing the initial conditions to a value dependent on I_{total} and removing it from the optimization caused a huge spike in the error. This is expected, as we aren't optimizing all of the available parameters. The error is typically below 10% though, so we can say that the model maintains some accuracy. Next we see that some parameters are now stable, and maintain the same order throughout the different time spans. This translates over to the Exposed ratio, as it can be seen for some cases that its values are less random. The Exposed ratio settles down, but still does not allow us to distinguish between news and rumors. It would be interesting to find a method of stabilizing the parameters while still being able to examine R_ε and N_ε.

5.2.4 Estimating Future Tweets

Although no conclusions were made about the *SEIZ* model's ability to detect rumors on Twitter, we do know that the model can successfully estimate the number of tweets that will be posted. In Figures 5.12 and 5.13, we restricted the last parts of the data from the model. Then, we used the optimal parameters to approximate the solution to the ODEs for the final day, and compared those results to the actual data.

For each of the 6 cases, the model was able to estimate the number of tweets for the given time frame plus at least an additional 12 hours. The Sarah Palin and Tiger Woods cases would fail completely if we tried to predict more than 12 hours of tweets, with relative errors nearing 90%. When looking at 12 hours or less of the predicted numbers of tweets, these cases managed to maintain a lower error. Both of these cases

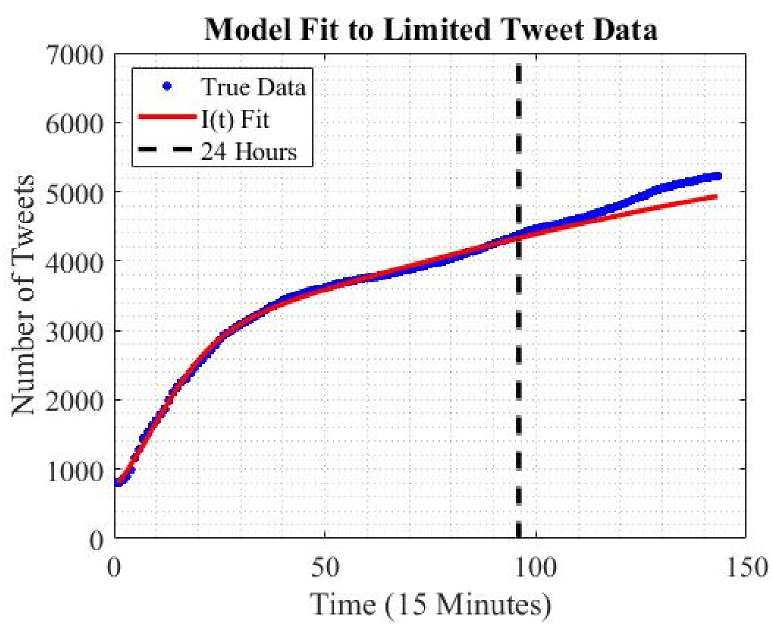

Figure 5.12: The model was fit to the Sarah Palin rumor over 24 hours, then was compared against the real data for 36 hours. The relative error here was only 1.80% over the entire case, and 2.7% after 24 hours.

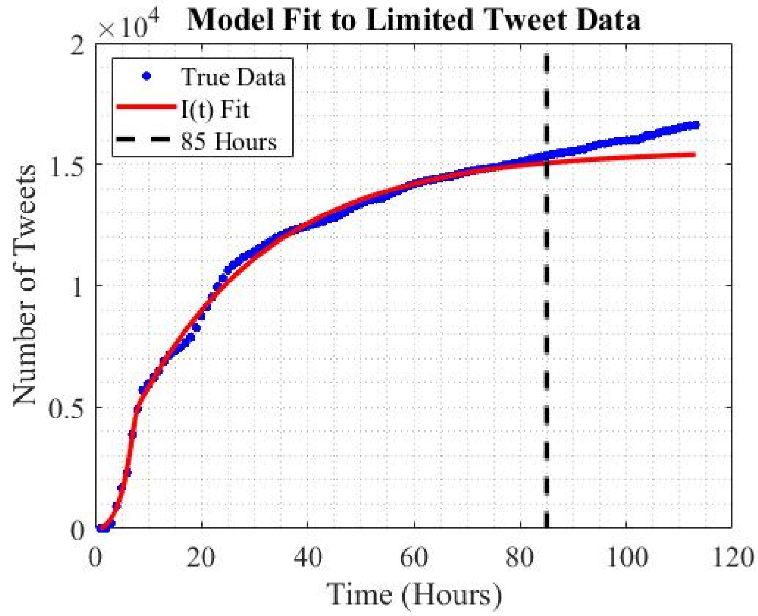

Figure 5.13: Here, the model was fit to the Fort Hood story over 85 hours, which is just about 3.5 days. This fit still had a relative error of just 3.29% overall, and 6.9% over the final 30 hours.

could predict the next 12 hours with an error less than 6%! The Britney Spears and Fort Hood stories were able to predict much further than the previous two cases, but with slightly less accuracy. Restricting 30 hours from a 5 day period, we could predict the final 30 hours of the Fort Hood case with a relative error of 6.9%. The Britney Spears rumor produced the largest error with 18 hours of data restricted, having 8.7%. The most impressive predictions came from the Kanye West rumor and the Michael Jackson story. Even when the entire second half of the data was removed, the model could still estimate the number of tweets with a relative error of less than 3%!

5.3 Applications

With all of the discussion around "fake news" recently, it can be seen how this topic is gaining in relevance. With there being so many more users now, accounts being more susceptible to hacking, or even popular social figures simply posting rumors, false information can be spread easier now more than ever. If we can confirm a parameter or a combination of these parameters to distinguish between real news and rumors, disclaimers could be attached to tweets which mention stories that could be rumors. This could reduce the number of retweets and end a rumor before it takes off, almost like introducing a vaccine. Though every rumor may not be identified and some real news stories made end up with a false disclaimer, this could encourage users to look into the credibility behind more controversial tweets. We could also have further analyzed this to see when real news stories really gain traction. This way, we could take advantage of these techniques to get information out to more people quicker in emergency situations.

The ability of the model to predict the number of tweets with suppressed data could also be beneficial. The analysis was performed rather quickly, and could be used real-time in industries such as marketing or politics. The insights provided by having an accurate estimate may be useful when looking at trending topics and big news stories.

Chapter 6

Conclusions and Future Considerations

It can be seen from this thesis how the spread of news and rumors is related to compartmental models from epidemiology. By examining more data sets, we were able to further strengthen the claim made in [1] that the *SEIZ* model is a great model for news propagation on Twitter. We observed that the Exposed ratio, R_{SI}, is too sensitive to changes in the functions used to gain useful information from.

It was also observed that no conclusions may be drawn from the newly proposed Infected-Incubation ratio, R_{ε}, or the Infected-Incubation number, N_{ε}. These values have shown potential though, and should be analyzed with more cases and with more recent data. It could be seen from Figures 5.8-5.11 that the rumors would trend in one direction while the real news stories would trend in the other. We could also see the potential in other parameters to distinguish between the two, but the limited data did not allow for further analysis. There are many factors which determine the optimal parameter values, so finding new ways to reduce these factors could stabilize these ratios and lead to more significant results. Extending previous work, we were also able to look at the model from various points in the story's timeline, and it showed us that we can use the model in real time to predict the tweet counts of the near future.

Going forward, we would like to apply this model on even more data sets, but this time with much more current data. This way, we could do a complete analysis on the three values discussed in this thesis and come to a more confident conclusion. We would also dig deeper into the effect of changing the constraints on these values and the model itself. There may be a set of unknown constraints that can stabilize these parameters over time. Additionally, we would like to look at an increased number of parameters and ratios to see if any of them could differentiate between news and rumors, the way we hoped R_{SI}, R_{ε}, or N_{ε} would.

Bibliography

[1] Jin, F., Dougherty, E., Saraf, P., Cao, Y., Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on Twitter. Proceedings of the 7th Workshop on Social Network Mining and Analysis, 1-9.

[2] U.S. population social media penetration 2018. (n.d.). Retrieved from https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/

[3] Global social media ranking 2018 (n.d.). Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[4] Twitter Usage Statistics. (n.d.). Retrieved from http://www.internetlivestats.com/twitter-statistics/

[5] Haran, M. (2009, December). An introduction to models for disease dynamics. Retrieved from http://www.unc.edu/ rls/s940/samsidisdyntut.pdf

[6] Bettencourt, Cintrn-Arias, Kaiser, Castillo-Chvez. (2006). The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A: Statistical Mechanics and Its Applications, 364(C), 513-536.

[7] J. Yang, J. Leskovec. Temporal Variation in Online Media. ACM International Conference on Web Search and Data Mining (WSDM '11), 2011.

[8] Michael Jackson dead at 50 after cardiac arrest. (2009, June 25). Retrieved from http://www.cnn.com/2009/SHOWBIZ/Music/06/25/michael.jackson/index.html

[9] Army major kills 13 people in Fort Hood shooting spree. (n.d.). Retrieved from https://www.history.com/this-day-in-history/army-major-kills-13-people-in-fort-hood-shooting-spree

[10] DiMeglio, S. (2009, December 03). Woods crash did $3,200 damage to hydrant, tree. Retrieved from https://usatoday30.usatoday.com/sports/golf/pga/2009-12-02-woods-crash-damage_N.htm

[11] Celebrity death rumors spread online. (2009, July 01). Retrieved from http://www.cnn.com/2009/TECH/07/01/celebrity.death.pranks/

[12] MacRory, S. (2009, October 21). Kanye West Not Dead, Despite Net Rumor Frenzy. Retrieved from https://www.billboard.com/articles/news/266982/kanye-west-not-dead-despite-net-rumor-frenzy

[13] Nudd, T. (2009, August 03). Sarah Palin Denies Divorce Rumors. Retrieved from https://people.com/celebrity/sarah-palin-denies-divorce-rumors/

[14] Murray, J. (2002). Mathematical biology (3rd ed., Interdisciplinary applied mathematics). New York: Springer.

Appendices

Appendix A

Matlab Code

A.1 Function 1

```
function [dy] = dSEIZ( t, y )

    global beta b p l rho eps;

    S = y(1);
    E = y(2);
    I = y(3);
    Z = y(4);

    N = S+E+I+Z;

    dy(1,1) = -1*beta*S*(I/N) - b*S*(Z/N);
    dy(1,2) = (1-p)*beta*S*(I/N) + (1-l)*b*S*(Z/N) - rho*E*(I/N) -
```

```
    eps*E;
    dy(1,3) = p*beta*S*(I/N) + rho*E*(I/N) + eps*E;
    dy(1,4) = l*b*S*(Z/N);
end
```

A.2 Function 2

```
function y = SEIZFitFunction(params)

    global data dt beta b p l rho eps;

    time=data(1,:)';                % Time
    trueI=data(2,:)';               % Actual Tweets from data

    S0 = params(1); E0 = params(2); I0 = params(3); Z0 = params(4);
    beta = params(5); b = params(6); p = params(7);
    l = params(8); rho = params(9); eps = params(10);

    [T,Y] = forward_euler(@dSEIZ, dt, [time(1) time(end)],[S0 E0 I0
    Z0]);
    I = Y(:,3);

    same = ismember(T, time);
    I_subset = I(find(same==1));
```

```
    y = abs(trueI - I_subset);                          % Fit error

end
```

A.3 Function 3

```
function SEIZ_fit()

    clear; clc;
    global data dt beta b p l rho eps;

    % Open data file.
    file = '/TigerWoods.txt';
    fid = fopen(file);

    % Load data.
    C_text = textscan(fid, '%s', 2, 'delimiter', '\t');
    C_data = textscan(fid, '%f %f');
    tweet_count = C_data{2};

    %Used to Limit Data
    \%tweet_count2 = tweet_count(end-100:end) = [];
    \%time2 = (1:length(tweet_count2));

    pop = round(tweet_count(end), 2, 'significant');
```

```
IC = [2*pop 0.3*pop 0.3*pop 0.3*pop 1 1 0.5 0.5 1 1];
LB = [0 0 0 0 0 0 0 0 0 2e-5];
UB = [3*pop 0.5*pop 0.5*pop 0.5*pop 10 10 1 1 10 10];

% Compute time array and dt.
time = (1:length(tweet_count));
dt = 0.1;

% Set data matrix for system fitting.
data = zeros(2, length(time) );
data(1,:) = time;
data(2,:) = tweet_count;

% Fit the Rumor (tweet) data to SEIZ model.
% Parameter order: S0 E0 I0 Z0 beta b p l rho eps
options = optimset('MaxFunEvals',1E8,'MaxIter',1E8,'TolFun',1e-8,
'TolX',1e-8);
fit=lsqnonlin('SEIZFitFunction', IC, LB, UB, options);

% Get fitting results.
S0 = fit(1); E0 = fit(2); I0 = fit(3); Z0 = fit(4);
beta = fit(5); b = fit(6); p = fit(7);
l = fit(8); rho = fit(9); eps = fit(10);
```

```
[T,Y] = forward_euler(@dSEIZ, dt, [time(1) time(end)],[S0 E0 I0
Z0]);
%[T,Y] = ode45(@dSEIZ1,[time(1) time(end)],[S0 E0 I0 Z0]);
S = Y(:,1);  E = Y(:,2);  I = Y(:,3);  Z = Y(:,4);
N = S+E+I+Z;

%Exposed Ratio
Rsi = ((1-p)*beta + (1-l)*b)/(rho+eps);
fprintf('Rsi = %f\n', Rsi);



% Display results
figure(10); hold on;
set(gca, 'FontName', 'Times New Roman', 'FontSize', 20)
plot(T, S, 'b', 'LineWidth', 2.5);
plot(T, E, 'r', 'LineWidth', 2.5);
plot(T, I, 'g', 'LineWidth', 2.5);
plot(T, Z, 'k', 'LineWidth', 2.5);
plot(T, N, 'm', 'LineWidth', 2.5)
legend({'S' 'E' 'I' 'Z' 'N'})
xlabel('Time (hours)'); ylabel('Tweets');
title('SEIZ Compartments');
saveas(gcf, 'SEIZ-Tiger-96hours.jpg')

% Plot true tweet data to fit results (I).
```

```
    figure(20); hold on;
    set(gca,'FontName','Times New Roman','FontSize',15)
    plot(time, tweet_count, 'b.',  'MarkerSize',14);
    plot(T, I, 'r-', 'LineWidth',2.5);
    xlabel('Time (hours)'); ylabel('Cumulative Tweet Volume');
    hleg1 = legend({'Tweet Data' 'Fit'});
    set(hleg1,'Location','NorthWest');
    title('Model Fit to Tweet Data');
    box on; grid on;
    saveas(gcf, 'I-Tiger-96hours.jpg')

    % Compute Error
    same = ismember(T, time);
    I_subset = I(find(same==1));
    err_norm = norm(I_subset-tweet_count) / norm(tweet_count);
    fprintf('Error = %f\n', err_norm);

end
```

Appendix B

Python Data Extraction

```
from dateutil import parser
import datetime
import time
import math

t_init = parser.parse('2009-11-27 2:00:00')
t_final = parser.parse('2009-12-01 2:00:00')

t_init = time.mktime(t_init.timetuple())
t_final = time.mktime(t_final.timetuple())

f1 = open('tweetsNov.txt')
f2 = open('tweetsDec.txt')
datafile = open('TigerWoods.txt', 'w')
```

```
t1 = t_init

data = []

count = 0

minute = 0

t_delta = datetime.timedelta(minutes=1)

strings = ('Tiger Woods', '#TigerWoods', '#Tiger')


for f in [f1,f2]:

for line in f:

data.append(line)

if len(data) > 3:

del data[0]


if any(s in line for s in strings):

twttime = data[0].split('', 1)[1]

twttime = parser.parse(twttime)

twttime = time.mktime(twttime.timetuple())


if twttime > t_final:

break


if twttime >= t_init:

count += 1

if twttime > t1:

hours = int(math.floor((twttime - t_init)/3600))
```

```
datafile.write(str(hours)+'\t'+str(count)+'\n')
t1 = t1 + math.floor((twttime - t1 + 3600)/3600)*3600


f2.close()
f1.close()
datafile.close()
```

Appendix C

Additional Tables

Table C.1: Proposed rumor identifiers for the Michael Jackson case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_{ε}	N_{ε}	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	0.13	3.81E-05	8.14E-07	2.63	2.60
16	28.62	5.29E-01	3.52E-04	2.16	2.58
18	37.98	1.87E+00	1.95E-04	1.78	2.06
20	64.79	2.50E-07	4.87E-10	2.64	1.51
22	30.24	4.90E-01	2.27E-04	1.70	1.33
24	103.92	2.85E+00	9.46E-05	1.25	3.35
26	73.86	4.09E+00	7.14E-05	1.15	1.14
28	70.10	4.47E+00	6.30E-05	1.04	1.03
30	84.00	3.80E+00	5.41E-05	0.94	0.94
32	62.12	4.05E+00	7.66E-05	0.90	0.87
34	14.81	3.69E+00	4.70E-05	0.84	2.63
36	57.72	2.88E+00	4.85E-05	0.84	3.12
38	142.18	2.48E-01	2.46E-05	2.46	3.21
40	0.01	3.78E-02	3.66E-05	0.74	0.78
42	8.57	8.33E+00	2.81E-05	0.83	0.82
44	23.52	2.11E+01	1.72E-05	0.91	0.88
46	0.03	2.89E-02	1.84E-05	0.88	0.93
48	14.50	7.31E+02	4.57E-07	0.94	0.93

Table C.2: Proposed rumor identifiers for the Fort Hood case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_{ε}	N_{ε}	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	15.57	6.95E-02	3.68E-04	2.35	2.71
16	27.61	2.00E-02	7.96E-05	2.59	1.25
18	21.11	1.06E+03	5.04E-07	2.08	4.01
20	45.02	1.08E+03	4.63E-07	2.28	5.25
22	81.58	1.43E-06	9.45E-10	6.02	6.02
24	28.15	5.21E+02	2.58E-07	2.45	6.21
26	24.00	8.20E+02	3.96E-07	2.21	5.90
28	21.81	6.45E+02	2.63E-07	2.45	2.43
30	31.55	1.19E+03	4.78E-07	2.34	5.16
32	0.44	1.23E+00	7.44E-05	1.18	2.49
34	0.00	7.67E+02	3.23E-07	2.40	2.38
36	0.07	1.93E+00	7.17E-05	1.26	4.29
38	176.00	1.88E-06	4.40E-10	4.14	2.24
40	12.38	9.08E+02	4.69E-07	2.06	3.96
42	0.07	3.69E+00	4.94E-05	1.29	2.10
44	17.69	6.05E+02	3.01E-07	2.01	2.01
46	0.53	9.45E+00	2.51E-05	1.47	1.95
48	10.93	4.60E+02	1.99E-07	1.97	3.29

Table C.3: Proposed rumor identifiers for the Tiger Woods case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_{ε}	N_{ε}	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	0.15	2.22E-03	2.89E-04	0.87	1.36
16	0.23	8.81E-18	1.11E-18	0.78	1.37
18	0.20	8.50E-03	8.98E-04	0.91	1.58
20	0.11	6.83E-02	5.07E-03	1.16	1.74
22	0.15	4.16E-02	3.33E-03	0.94	1.62
24	0.16	3.89E-02	3.09E-03	0.84	1.51
26	0.01	1.87E-01	8.94E-03	1.33	1.71
28	0.15	1.25E-01	6.32E-03	1.65	2.32
30	108.11	1.35E+01	4.53E-04	2.71	2.77
32	244.43	3.77E+02	2.11E-05	2.55	2.55
34	83.54	4.16E+03	2.17E-06	2.63	2.54
36	201.45	7.85E+02	8.42E-06	2.19	2.23
38	82.02	2.43E+03	1.07E-06	2.93	2.91
40	127.32	1.12E+03	4.47E-07	4.43	4.33
42	178.43	2.12E+03	1.02E-06	3.94	4.44
44	124.35	2.12E+03	1.06E-06	3.91	4.21
46	56.28	2.38E+03	1.07E-06	4.27	3.37
48	157.73	2.76E+03	1.21E-06	4.14	3.89

Table C.4: Proposed rumor identifiers for the Britney Spears case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_{ε}	N_{ε}	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	0.13	3.21E-08	5.98E-08	2.89	2.87
16	0.02	3.35E-05	3.58E-04	3.71	2.92
18	0.01	5.11E+00	5.50E-04	2.49	2.53
20	24.24	2.16E+03	1.17E-06	2.20	4.38
22	18.87	2.06E+03	8.59E-07	2.04	1.99
24	21.99	2.06E+03	8.65E-07	1.83	1.84
26	17.45	2.18E+03	9.54E-07	1.80	1.67
28	20.69	1.53E+03	8.75E-07	1.52	1.52
30	20.30	1.55E+03	8.39E-07	1.41	1.41
32	0.03	1.78E+00	5.38E-04	1.26	1.30
34	12.54	1.61E+03	7.51E-07	1.26	3.44
36	14.33	1.39E+03	6.50E-07	1.23	1.20
38	0.18	1.45E+00	5.13E-04	1.06	1.12
40	0.20	1.55E+00	4.69E-04	1.07	1.10
42	0.19	2.12E+00	3.47E-04	1.10	1.11
44	110.49	6.18E+00	7.67E-04	2.83	1.19
46	50.34	5.32E+00	1.28E-04	1.19	1.17
48	0.02	2.41E+00	2.63E-04	1.15	1.15

Table C.5: Proposed rumor identifiers for the Kanye West case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_{ε}	N_{ε}	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	22.95	1.25E+03	6.35E-07	1.41	1.19
16	21.23	1.20E+03	5.14E-07	1.18	1.30
18	0.04	9.22E-16	1.55E-17	0.40	1.27
20	9.78	1.30E+03	5.97E-07	1.19	1.10
22	12.47	8.61E+02	3.59E-07	1.00	1.01
24	0.02	9.16E-04	1.42E-05	0.89	1.01
26	0.02	1.92E-03	3.08E-05	1.03	1.11
28	13.64	1.02E+00	7.22E-05	1.60	1.60
30	0.02	1.95E-03	5.09E-05	1.82	1.82
32	0.01	3.86E-03	1.15E-04	1.81	1.80
34	0.01	6.85E-07	2.75E-08	1.73	1.72
36	0.01	9.19E-07	4.32E-08	1.63	1.61
38	188.72	2.69E-01	7.34E-05	1.54	1.54
40	0.04	5.29E-01	2.15E-03	1.79	1.50
42	0.21	6.65E-02	2.31E-03	1.60	1.46
44	0.01	2.84E-07	1.29E-08	1.43	1.43
46	0.01	3.33E-07	1.07E-08	1.39	1.39
48	0.02	3.61E-08	1.83E-09	1.36	1.34

Table C.6: Proposed rumor identifiers for the Sarah Palin case when updating initial guesses. The errors from this method are shown with the errors from the original method.

Duration (Hours)	R_{SI}	R_ε	N_ε	$\%Err_{rel}$	$\%Err_{rel}$ (Original)
14	0.28	4.29E-05	1.48E-05	1.47	0.98
16	0.02	2.27E-02	2.81E-04	0.84	0.83
18	0.39	3.86E-06	4.52E-07	1.21	0.72
20	141.15	7.28E-02	1.03E-04	0.64	0.64
22	0.03	6.07E-03	4.42E-04	0.61	0.58
24	33.31	8.47E-07	1.43E-09	0.76	0.70
26	0.29	6.07E-12	1.97E-14	1.07	1.11
28	5.38	1.89E-01	1.94E-04	1.71	1.84
30	7.76	1.68E+00	5.52E-04	2.10	2.05
32	42.22	6.60E+01	3.41E-05	1.88	2.32
34	73.17	2.20E+03	8.98E-07	1.66	1.68
36	19.14	2.22E+00	3.19E-04	2.58	1.54
38	13.64	2.05E+03	9.89E-07	1.70	1.75
40	30.92	2.32E+03	1.04E-06	1.82	1.83
42	32.27	2.23E+03	9.87E-07	1.68	5.41
44	87.22	2.32E+02	6.53E-06	1.42	1.43
46	21.10	2.03E+03	9.90E-07	1.65	1.58
48	20.61	1.83E+03	8.74E-07	1.62	4.31

Table C.7: All optimized parameters for the Michael Jackson case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	2.71E-08	6.70E-01	4.85E-09	6.87E-01	4.52E-01	2.00E-05	0.46	11.11
16	1.27E-08	7.24E-01	3.43E-06	7.05E-01	3.72E-01	2.02E-05	0.57	10.36
18	3.92E-06	7.88E-01	7.62E-06	7.08E-01	2.91E-01	2.30E-03	0.79	9.78
20	5.76E-06	6.96E-01	1.24E-04	6.87E-01	1.43E-01	1.67E-02	1.36	9.14
22	1.75E-10	5.93E-01	7.12E-03	6.19E-01	2.82E-05	2.91E-02	7.76	8.77
24	6.01E-06	5.85E-01	7.89E-05	6.66E-01	3.96E-02	2.16E-02	3.20	8.54
26	1.69E-06	6.76E-01	3.77E-14	5.87E-01	1.85E-08	1.88E-02	14.84	8.20
28	4.59E-06	7.14E-01	1.27E-12	5.65E-01	1.31E-07	1.58E-02	19.61	7.87
30	1.37E-06	7.39E-01	4.09E-14	5.47E-01	8.98E-09	1.36E-02	24.61	7.58
32	5.59E-02	8.56E-01	7.20E-06	5.42E-01	1.52E-06	1.17E-02	38.38	7.34
34	6.54E-06	7.35E-01	1.88E-08	5.29E-01	1.22E-07	1.09E-02	31.85	7.08
36	5.37E-06	7.56E-01	2.31E-14	5.06E-01	5.69E-09	9.46E-03	39.47	6.91
38	5.10E-06	7.68E-01	3.14E-14	4.90E-01	6.14E-09	8.59E-03	45.59	6.66
40	2.41E-01	9.45E-01	3.17E-13	5.02E-01	1.17E-06	7.63E-03	93.28	6.37
42	2.65E-08	7.65E-01	4.18E-14	4.87E-01	2.09E-11	8.37E-03	46.86	5.96
44	1.66E-06	7.34E-01	2.53E-14	5.00E-01	8.58E-09	8.64E-03	42.49	5.65
46	6.28E-08	5.33E-01	1.31E-06	6.17E-01	8.77E-03	8.35E-03	11.91	5.98
48	2.79E-07	7.23E-01	4.60E-14	6.51E-01	2.22E-02	6.82E-03	8.72	5.76

Table C.8: All optimized parameters for the Fort Hood case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	4.12E-05	4.98E-01	5.11E-05	5.97E-01	3.26E-01	2.01E-05	0.62	14.00
16	9.14E-08	4.42E-01	9.08E-05	6.77E-01	3.68E-01	2.06E-05	0.39	12.50
18	7.76E-12	4.48E-01	1.55E-06	6.96E-01	3.39E-01	2.00E-05	0.40	11.30
20	1.68E-09	4.64E-01	1.88E-06	7.03E-01	3.16E-01	2.00E-05	0.44	10.23
22	2.28E-06	4.81E-01	8.86E-05	7.18E-01	2.89E-01	2.00E-05	0.47	9.88
24	2.32E-11	5.23E-01	3.09E-07	7.39E-01	2.45E-01	2.00E-05	0.56	10.51
26	3.37E-08	4.76E-01	3.11E-06	6.21E-01	1.04E-09	2.13E-02	8.47	10.72
28	3.40E-08	5.64E-01	2.36E-14	5.52E-01	7.13E-11	1.57E-02	16.16	10.23
30	1.23E-06	6.21E-01	8.56E-10	4.95E-01	3.76E-12	1.12E-02	27.90	10.12
32	3.13E-06	6.63E-01	2.78E-14	4.64E-01	6.02E-07	1.05E-02	33.92	9.28
34	1.78E-05	6.53E-01	2.35E-14	4.70E-01	1.28E-07	1.06E-02	32.60	8.58
36	1.58E-09	5.58E-01	3.58E-07	5.20E-01	1.29E-10	1.02E-02	26.35	8.57
38	7.05E-08	5.33E-01	6.62E-09	5.37E-01	3.04E-09	1.06E-02	23.24	8.04
40	9.78E-08	2.49E+00	3.76E-03	6.69E-01	6.46E-02	3.08E-03	12.18	8.12
42	1.46E-07	4.81E-01	2.46E-09	5.69E-01	3.33E-06	1.17E-02	17.78	7.21
44	7.20E-08	5.90E-01	5.12E-11	6.39E-01	2.90E-02	8.67E-03	5.65	6.92
46	2.66E-09	4.55E-01	2.07E-03	6.75E-01	2.70E-02	8.94E-03	4.11	7.05
48	1.25E-06	4.59E-01	2.81E-11	6.81E-01	3.04E-02	8.61E-03	3.76	6.76

Table C.9: All optimized parameters for the Tiger Woods case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	1.63E-05	6.37E+00	3.91E-14	7.01E-01	3.90E-14	1.14E-01	16.69	5.29
16	1.04E-04	7.66E+00	3.01E-09	7.07E-01	4.18E-14	9.96E-02	22.56	4.91
18	1.41E+00	1.00E+01	2.05E-06	7.53E-01	6.95E-10	8.57E-02	45.23	4.89
20	3.46E-01	5.89E-01	7.18E-01	8.39E-01	1.42E-05	2.36E-02	8.15	3.94
22	3.46E-01	5.70E-01	6.91E-01	8.56E-01	3.02E-06	2.34E-02	8.08	3.65
24	3.28E-01	5.65E-01	6.99E-01	8.63E-01	4.45E-06	2.42E-02	7.28	3.42
26	3.10E-01	5.12E-01	6.81E-01	8.58E-01	2.32E-06	1.81E-02	9.46	3.40
28	2.08E-01	1.22E+00	9.97E-01	7.02E-02	5.76E-03	1.14E-03	164.10	3.71
30	4.53E-01	4.10E-01	4.30E-01	7.28E-01	1.96E-02	2.26E-05	18.78	3.71
32	1.94E-01	5.24E-01	8.37E-01	5.21E-01	1.91E-02	2.17E-05	14.81	3.86
34	5.03E-01	4.66E-01	3.45E-01	4.40E-01	1.31E-02	2.15E-05	45.12	3.59
36	2.83E-01	8.82E-01	5.58E-01	1.33E-01	1.05E-02	2.04E-05	84.98	3.59
38	7.23E-01	2.89E-01	1.91E-01	7.75E-01	1.39E-02	2.08E-05	46.60	4.10
40	7.85E-01	1.14E+00	2.50E-01	3.70E-01	1.42E-02	2.02E-05	92.07	5.76
42	2.32E+00	2.70E+00	1.79E-01	2.89E-01	1.34E-02	2.20E-05	285.33	6.44
44	1.75E+00	1.61E+00	1.55E-01	4.65E-01	1.52E-02	2.27E-05	153.33	6.11
46	1.27E+00	1.15E+00	1.37E-01	2.27E-01	1.20E-02	2.28E-05	165.62	5.24
48	9.34E-01	3.31E-01	1.10E-01	4.26E-01	1.15E-02	2.28E-05	88.35	4.74

Table C.10: All optimized parameters for the Britney Spears case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	8.85E-13	9.66E-01	1.42E-01	7.07E-01	5.40E-01	2.00E-05	0.52	6.97
16	1.26E-04	9.82E-01	2.29E-07	7.25E-01	3.63E-01	1.17E-02	0.72	7.47
18	2.68E-13	8.72E-01	3.36E-04	6.44E-01	3.58E-14	4.36E-02	7.11	7.46
20	5.54E-07	9.69E-01	8.86E-04	6.33E-01	9.45E-05	3.22E-02	11.01	7.39
22	4.72E-01	1.44E+00	1.81E-09	6.24E-01	1.14E-08	2.04E-02	49.69	7.26
24	1.79E-08	1.17E+00	4.17E-14	5.61E-01	2.45E-11	1.84E-02	27.92	6.83
26	1.18E+00	1.94E+00	1.10E-10	5.82E-01	7.36E-06	1.24E-02	160.81	6.79
28	2.70E-07	1.32E+00	5.53E-10	4.84E-01	3.45E-10	1.20E-02	56.50	6.33
30	1.07E+00	1.66E+00	1.27E-08	5.44E-01	1.05E-06	9.55E-03	191.01	6.14
32	1.31E+00	1.65E+00	5.60E-11	5.44E-01	7.24E-10	8.38E-03	246.68	5.92
34	4.98E-01	1.29E+00	1.56E-13	5.18E-01	5.49E-07	8.85E-03	126.37	5.54
36	2.09E-01	9.67E-01	5.95E-12	5.41E-01	4.19E-07	9.67E-03	67.46	5.27
38	1.64E-06	9.20E-01	2.87E-14	5.20E-01	6.49E-10	9.62E-03	45.92	5.10
40	2.72E-06	8.31E-01	4.15E-09	5.79E-01	6.00E-03	9.46E-03	22.62	4.98
42	3.05E-06	8.16E-01	8.54E-09	6.00E-01	9.34E-03	9.05E-03	17.77	4.81
44	3.80E-08	8.16E-01	4.35E-14	5.38E-01	1.83E-09	9.13E-03	41.33	4.56
46	2.77E-07	5.93E-01	2.42E-14	5.87E-01	1.70E-08	8.75E-03	28.02	4.85
48	4.82E-05	6.25E-01	4.44E-12	6.06E-01	3.92E-03	8.46E-03	19.90	4.63

Table C.11: All optimized parameters for the Kanye West case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	2.56E+00	2.46E+00	1.70E-09	4.55E-01	1.33E-05	2.21E-02	176.38	8.19
16	7.98E-04	4.50E+00	2.94E-14	1.81E-01	1.60E-02	1.14E-02	134.41	7.41
18	2.39E-05	2.46E+00	2.70E-14	3.69E-01	2.75E-02	1.17E-02	39.43	6.59
20	1.20E-08	1.71E+00	1.30E-08	5.73E-01	7.59E-02	9.40E-03	8.57	6.08
22	4.69E-07	1.50E+00	1.18E-06	6.50E-01	1.20E-01	4.77E-03	4.21	5.75
24	4.18E-08	1.47E+00	8.06E-06	6.56E-01	1.26E-01	4.49E-03	3.86	5.11
26	1.35E-08	1.11E+00	1.65E-07	6.91E-01	1.37E-01	3.38E-03	2.44	5.07
28	7.88E-09	6.49E-01	4.75E-02	6.99E-01	9.67E-02	8.58E-03	1.85	5.19
30	1.75E-06	2.01E+00	3.78E-03	6.92E-01	1.16E-01	2.87E-03	5.24	5.07
32	1.05E-06	3.92E-01	2.04E-06	6.73E-01	2.16E-02	1.80E-02	3.23	4.92
34	6.47E-08	4.02E-01	1.28E-09	6.50E-01	6.81E-03	1.89E-02	5.48	4.59
36	2.34E-14	3.93E-01	4.19E-14	6.41E-01	9.11E-08	1.96E-02	7.19	4.29
38	3.09E-07	3.99E-01	1.38E-10	6.52E-01	8.34E-03	1.88E-02	5.13	4.07
40	1.28E-04	3.65E-01	3.49E-14	6.50E-01	2.54E-06	2.07E-02	6.19	3.87
42	1.93E-06	4.50E-01	1.45E-10	6.68E-01	2.88E-02	1.57E-02	3.36	3.73
44	1.56E-06	4.04E-01	2.71E-14	6.66E-01	1.93E-02	1.77E-02	3.65	3.57
46	2.15E-06	3.80E-01	2.41E-14	7.01E-01	3.15E-02	1.57E-02	2.41	3.72
48	3.55E-09	3.84E-01	6.73E-10	7.02E-01	3.36E-02	1.54E-02	2.34	3.58

Table C.12: All optimized parameters for the Sarah Palin case when fixing the initial conditions. The errors from this method are shown along with the Exposed ratio.

Duration (Hours)	β	b	p	l	ρ	ε	R_{SI}	$\%Err_{rel}$
14	2.87E-09	1.06E+00	7.51E-07	6.71E-01	3.45E-01	2.01E-05	1.01	10.33
16	3.89E-14	1.05E+00	5.45E-05	6.83E-01	3.14E-01	2.30E-05	1.05	8.89
18	4.69E-05	9.24E-01	3.11E-06	6.94E-01	2.66E-01	4.07E-03	1.05	8.09
20	7.33E-06	8.52E-01	1.19E-06	6.95E-01	2.24E-01	8.46E-03	1.12	7.31
22	5.89E-06	7.94E-01	2.03E-04	7.01E-01	1.96E-01	1.15E-02	1.14	6.79
24	3.20E-12	7.31E-01	1.20E-01	7.06E-01	1.62E-01	1.54E-02	1.21	6.44
26	4.99E-07	7.35E-01	9.31E-04	7.13E-01	1.58E-01	1.42E-02	1.23	6.30
28	2.75E-06	5.52E-01	2.27E-10	7.05E-01	2.47E-02	3.22E-02	2.86	6.43
30	6.27E-13	5.65E-01	7.97E-12	7.03E-01	3.11E-14	3.09E-02	5.42	6.61
32	2.50E-06	1.30E+00	1.59E-06	7.02E-01	3.39E-02	1.94E-02	7.26	6.93
34	2.86E-07	7.23E-01	5.10E-13	6.85E-01	2.61E-09	2.34E-02	9.74	6.59
36	1.62E-06	2.45E+00	9.82E-08	6.67E-01	2.58E-05	1.75E-02	46.79	6.99
38	1.20E-06	2.23E+00	1.33E-04	6.69E-01	2.18E-05	1.58E-02	46.74	7.03
40	1.45E-06	2.58E+00	2.49E-03	6.54E-01	1.96E-05	1.40E-02	63.78	6.99
42	6.24E-12	1.13E+00	1.09E-01	6.48E-01	2.73E-14	1.34E-02	29.68	6.76
44	5.12E-08	1.22E+00	3.63E-06	6.39E-01	2.37E-08	1.22E-02	36.12	6.61
46	6.09E-07	1.25E+00	9.38E-04	6.43E-01	2.10E-06	1.18E-02	38.09	6.41
48	1.48E-07	1.27E+00	2.54E-05	6.36E-01	1.45E-07	1.09E-02	42.33	6.22