Bryant University

HONORS THESIS

How Infectious is Your Twitter Feed? Disease Modeling Applied to the Dynamics of Twitter

BY Kai-Jia Yue

ADVISOR • Dr. Brian Blais

EDITORIAL REVIEWER • Dr. Suhong Li

_Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
December 2022

Table of Contents

ABSTRACT

This study aims to use compartmental disease-models to explore Twitter dynamics. Applying an epidemiology model to Twitter tweets can give deeper insights into the factors that make a tweet go viral. In addition, this study explored the differences between a stochastic and a dynamic compartmental model. This research connects the world of diseases with the internet and explored if a disease model will accurately model Twitter dynamics. We found that stochastic models were better at fitting to smaller populations of data than dynamic models were. Dynamic models ended up predicting larger populations better. Furthermore, we found that although a topic is popular does not mean that it is infectious. This study was able to show that disease modeling is able to accurately predict Twitter dynamics.

INTRODUCTION

Twitter is a microblogging and social networking site, where people send out tweets, which are messages posted to their newsfeeds and networks. People used to have to wait for the morning news or paper to find the latest news, but with Twitter, information and news is right at the tip of your fingers. Since the launch of Twitter, users and their behavior have evolved (Liu et al., 2014). The platform has also changed the way news and information is spread throughout the world (Kwak et al., 2010). Hashtags, mainly used to group topics, are a popular way to use twitter and can be categorized into two groups: slightly infectious and very infectious (Skaza & Blais, 2017). One seemingly small tweet can lead to massive ripples throughout the online community. Figure 1 shows an example of a tweet that we scraped.

This study attempts to use a compartmental model known as the SIR model, which stands for Susceptible, Infected, and Recovered, to attempt to analyze the infectiousness of tweets and how they spread. Being able to further explore the dynamics of the spread of tweets will give a better understanding of how information spreads and be able to predict what content may go viral next. Utilizing a disease modeling approach, we should be able to see how people get "infected" or persuaded by ideas and how the people you follow and interact with on twitter influence how you are receiving news. This research connects closely with the research "Modeling the infectious of Twitter hashtags", written by Jonathan Skaza and Dr. Brian Blais. Their study specifically applies the dynamic version of the compartmental model to multiple hashtags in the Twittersphere. This study also looks at Twitter dynamics with a compartmental model, however this study uses both stochastic and dynamic models. Furthermore, instead of looking at a bunch of hashtags, this study looks deeply into the dynamics of one topic.

METHODS

Data Collection

In our experiment, we used data scraped from Twitter to see what factors lead to an idea, or in this case, a tweet to become infectious. Using an API that Twitter created, we collected our tweets over a period of weeks and store them on the cloud. Each Twitter tweet that we scrape includes the following data: date, language, source, tweet_id, user_screen_name, user_location, user_description, user_lang, user_verified, user_followers, user_friends, user_listed, user_favorites, user_statuses, user_profile, user_joinDate, tweet_country, tweet_place_name, reply_screen_name, quoted_tweet_id, retweeter_screen_name, retweeter_followers, retweeter_friends, retweeter_listed, retweeter_favorites, retweeter_statuses, retweeter_joinDate, tweet_media, element, user_mentions_screen_name, element, longtitude, and latitude. We used this information to better understand the dynamics of Twitter.

We specifically scraped tweets during the 2016 Rio summer Olympics, 8/1/2016 to 8/29/2016. We specifically looked for a topic that had numerous topics with it that had varying numbers of tweets. We wanted to be able to compare and contrast the results from different size populations. The Olympics was topic that had multiple topics of varying popularity, as well as a somewhat controlled environment that we suspect made the tweeting population somewhat consistent. The topics we chose from the Olympics were modern pentathlon, swimming, gymnastics, steeplechase, and curling. After scraping the data from twitter, we then cleaned the data. We created a new column that measured the seconds from January 1st, 2016, to the time of tweet. We then converted the seconds into hours. Figure 2 shows an example of a timeline of the data over hours. Using that column, we binned the tweets into hours by converting the seconds. The binned data gave us an idea of when the topics peaked and when that run of the idea stopped. Figure 3 shows an example of the binned data by hours.

Model Training

In this study, we used a compartmental model to explore the dynamics of Twitter. We measured infectiousness by the rate of how fast the idea has spread throughout the Twittersphere.

Compartmental Model

A compartmental model divides the total population into a certain number of groups and then specifies the rates at which one moves from one compartment to another. There are two types of compartmental models: dynamic and stochastic. A dynamic model is deterministic with continuous variables, where if you run the same data, the model will return the same exact results. Furthermore, the dynamic model approximates the value for the population meaning it is only valid for large populations. In comparison, the stochastic is probabilistic with discreate variables, where the jump from susceptible and infected happens sometimes and does not happen others. In addition, the stochastic is valid for any number of tweets because it allows for an extrapolation of the population. In this research, we also looked at the differences between dynamic and stochastic models.

Disease Model

We specifically be utilized the SIR model, which stands for Susceptible (S), Infected(I), and Recovered (R) while also showing the rates of infectiousness (β) and recovery (γ). The population moves from susceptible to infected to recovered. Figure 4 shows the SIR model applied to the flu. In the case of the flu, most of the population starts in the susceptible compartment and a few in the infected. The infection then spreads by moving people from the S compartment to the I compartment at an infection rate of β. Meanwhile, the infected recover by moving from I to R at a recovery rate of γ. The β and γ the disease, the environment, and the behavior of the population.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = +\frac{\beta SI}{N} - \gamma I \qquad (Equation\ 1)$$

$$\frac{dR}{dt} = +\gamma I$$

Equation 1 shows the equations for the SIR model. $\frac{dS}{dt}$ is how much the susceptible changes over time. - $\frac{\beta SI}{N}$ means that the susceptible is decreasing at a rate β depending on the interaction of the susceptible with the infected $\frac{SI}{N}$. The same amount is added to the I, so the change in infected in time, $\frac{dI}{dt}$, is $+\frac{\beta SI}{N}$ where the amount of the recovered, γI, is subtracted out. $\frac{dR}{dt}$ is how much the recovered changes over time. γI means that the infected is decreasing at a rate of γ.

In order to apply the SIR model to social media, we adapted the model, as shown in Figure 5. The population is split up into 3 compartments, followers, tweets, and silent. Each compartment has a specific equation that determine how those populations move from one population to the next.

There are many variations of the SIR model. In this research, we used the SIRI variation of the SIR model. As depicted in Figure 6, the SIRI model has an enhanced recovery rate, which allows the population to recover at an increased rate. We utilized this model in an effort to identify which users have been "infected" by a tweet and to show who they could pass it next by who is susceptible. This helps show the trail of infection of the tweet.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = +\frac{\beta SI}{N} - \frac{\gamma SI}{N} \qquad (Equation\ 2)$$
$$\frac{dR}{dt} = -\frac{\gamma SI}{N}$$

Equation 2 shows the equations for the SIRI model. $\frac{dS}{dt}$ is how much the susceptible changes over time. - $\frac{\beta SI}{N}$ means that the susceptible is decreasing at a rate β depending on the interaction of the susceptible with the infected $\frac{SI}{N}$. The same amount is added to the I, so the change in infected in time, $\frac{dI}{dt}$, is $+\frac{\beta SI}{N}$ where the amount of the recovered, $\frac{\gamma SI}{N}$, is subtracted out. $\frac{dR}{dt}$ is how much the recovered changes over time.$\frac{\gamma SI}{N}$means that the infected is decreasing at a rate of γ depending on the interaction of the susceptible with the infected $\frac{SI}{N}$.

Markov Chain Monte Carlo (MCMC)

In order to find the best-fit β and γ, we will be utilizing the method called Markov Chain Monte Carlo, which is the experimental determination of the values of parameters that impact the model's behavior. By using parameter estimation, this will allow us to tweak the environmental variables to observe whether there is a set of parameters that will better predict the virality of the tweet and then maximize the probabilities of those parameters which will allow us to predict which factors lead more accurately to the virality of a tweet. As shown in figure 7, this algorithm not only returns the best fit, but a full probability distribution for the parameters. This will allow us to find the best-fit βs and γs, but also see the uncertainties associated with them, which is the experimental determination of the values of parameters that impact the model's behavior.

Basic Reproductive Number

The basic reproductive number, also known as R_0, is the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. The equation of R_0 is:

$$R_0 = \frac{\beta}{\gamma} \tag{3}$$

We used the R_0in order to have a comparison between actual diseases and the topics. Although the as R_0 has been derived for the SIRI model, we will be using the SIR model to get a more accurate as R_0.

RESULTS

Dynamic vs. Stochastic Models

Some sample outputs for the SIRI model ran on both dynamic and stochastic models are depicted in figure 8. The orange line represents the SIRI model and the blue line represents the plotted data. These four plots show data from the modern pentathlon and swimming. Modern pentathlon was an unpopular topic with a maximum of 175 tweets. Swimming was a popular topic with a maximum of 6000 tweets.

The dynamic models show a smooth orange line in comparison to the bumpy irregular line of the stochastic. The bumpiness of the stochastic model is due to the probabilistic nature of the model. In addition, the stochastic model was better at predicting the model for modern pentathlon than swimming was. The stochastic model, although bumpy has a higher peak and follows the data more closely.

Stochastic Model

Figure 9 shows all the βs and γs for all sports ran with the stochastic SIRI model.

In this research we used β to look at how "infectious" a topic was. Steeplechase had the highest β meaning that it is the most infectious in comparison to the rest of the sports. Steeplechase was an unpopular sport with the highest β of 0.934, whereas Gymnastics was a popular sport with the lowest β of 0.333. Swimming, which was also a popular topic, had the second highest β at 0.633. Curling, which was the least popular topic, had the lowest β of 0.496.

Swimming, which was one of the most popular topics had a β of 0.633 and a γ of 2.901, had the largest γ in proportion to the β. This means when someone gets "infected" with the idea of swimming they recovered at a significantly faster rate than normal. Gymnastics, Steeplechase, and Modern Pentathlon had very similar γs all around 1.4-1.5. Curling had the lowest γ at 0.842, however that is proportional with the low β.

Dynamic Model

Figure 10 shows all the βs and γs for all sports ran with the dynamic SIRI model.

In comparison the dynamic model βs for gymnastics, swimming, steeplechase, and curling are all lower than the best fits for the stochastic model. Furthermore, the uncertainty for all the βs with the dynamic model are bigger. Steeplechase, which still has the largest β, has a β of 1.522 in comparison to its β for the stochastic model that is 0.596. Gymnastics still has the smallest β at 0.335, but the uncertainty is significantly smaller than the stochastic model. However, the βs all follow around the same dynamic that the stochastic model set out.

Swimming still has the highest γ which is 2.757 which stays consistent with the stochastic model. Modern pentathlon has a smaller γ in comparison to the one from the stochastic model. Gymnastics had the lowest β but the second highest γ. This indicates the recovery from the topic is faster than the infection rate. Overall, the γs reflect very similarly what was found in the stochastic model.

Basic Reproductive Number

Figure 11 shows the R_0 for all topics. The best fit R_0 for steeplechase is 7.928 and an uncertainty of +3.215 and -3.433. The best fit R_0 for swimming is 2.573 and an uncertainty of +1.149and -1.338. The best fit R_0 for gymnastics is 3.878 and an uncertainty of +3.202 and -1.930. The best fit R_0 for modern pentathlon is 2.573 and an uncertainty of +1,149and -1.338. Steeplechase has the highest R_0 compared to the other topics.

DISCUSSION AND CONCLUSIONS

This research scratches the surface of the application of disease modeling on the dynamics of social media. Potentially by looking at other topics or by delving deeper into the current topics, we could uncover certain factors that actively influence how viral a topic is.

This study establishes that disease modeling can be applied to Twitter dynamics. We now know that we can map out the dynamics of a topic using disease modeling. Furthermore, we were able to conclude that although a topic is popular it does not always mean it is infectious.

FIGURES

Figure 1: An example of a tweet that was scraped. This tweet was tweeted by NBC about the Gymnastics Olympic Trials

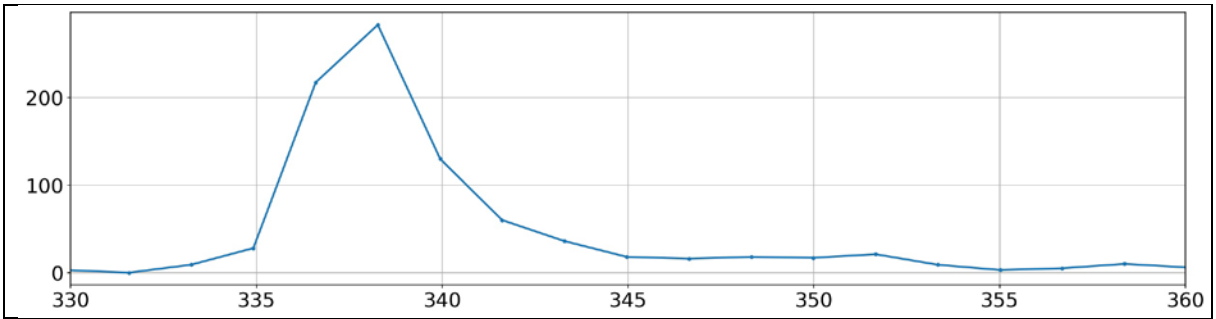

Figure 2: An example of a timeline of tweets over hours. This example specifically looks at steeplechase over a period of 30 hours where the number of tweets peaked at a little under 300.

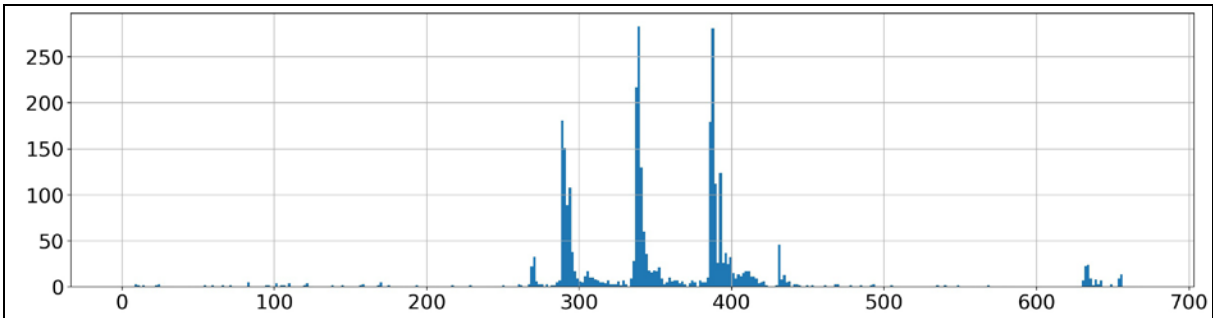

Figure 3: An example of binned tweets over time. This example specifically looks at all the steeplechase tweets binned over 700 hours.

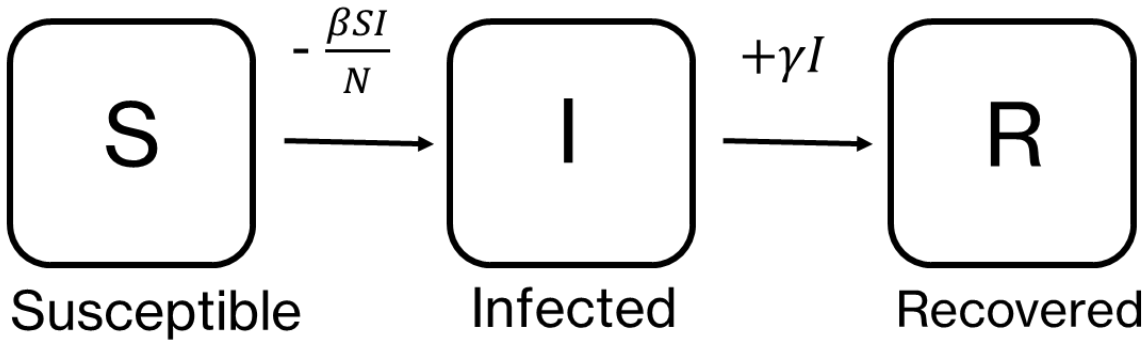

Figure 4: This shows the SIR model applied to the Flu. The population is split up into 3 compartments, susceptible, infected, and recovered. Each compartment has a specific equation that determine how those populations move from one population to the next. The infection spreads by moving people from the S compartment to the I compartment at an infection rate of β. Meanwhile, the infected recover by moving from I to R at a recovery rate of γ.

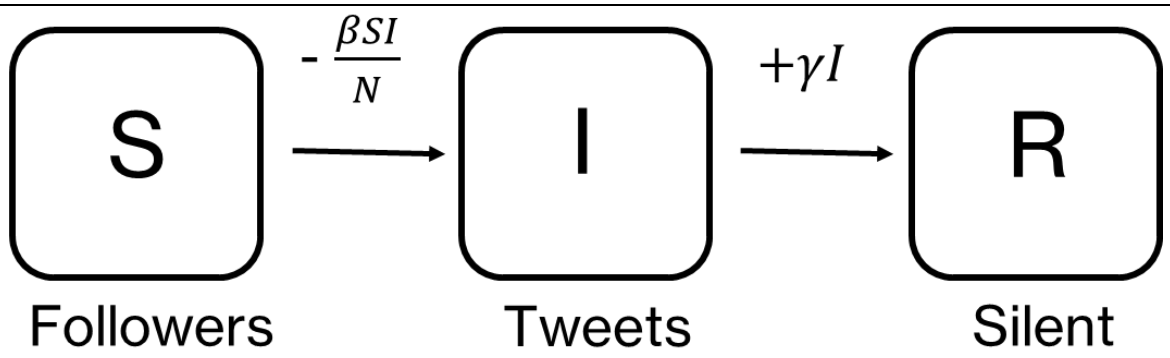

Figure 5: This shows the SIR model applied to social media. The population is split up into 3 compartments, followers, tweets, and silent. Each compartment has a specific equation that determine how those populations move from one population to the next.

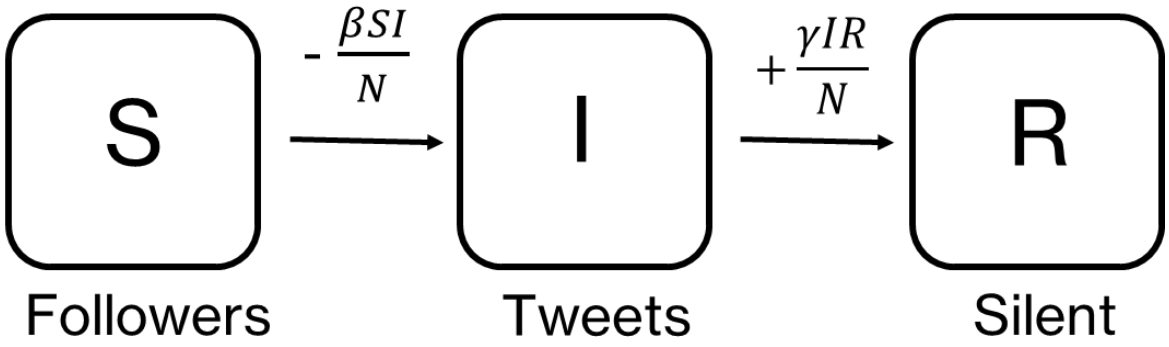

Figure 6: This shows the SIRI model applied to social media. The population is split up into 3 compartments, followers, tweets, and silent. Each compartment has a specific equation that determine how those populations move from one population to the next. The SIRI model has an enhanced recovery rate, which allows the population to recover at an increased rate.

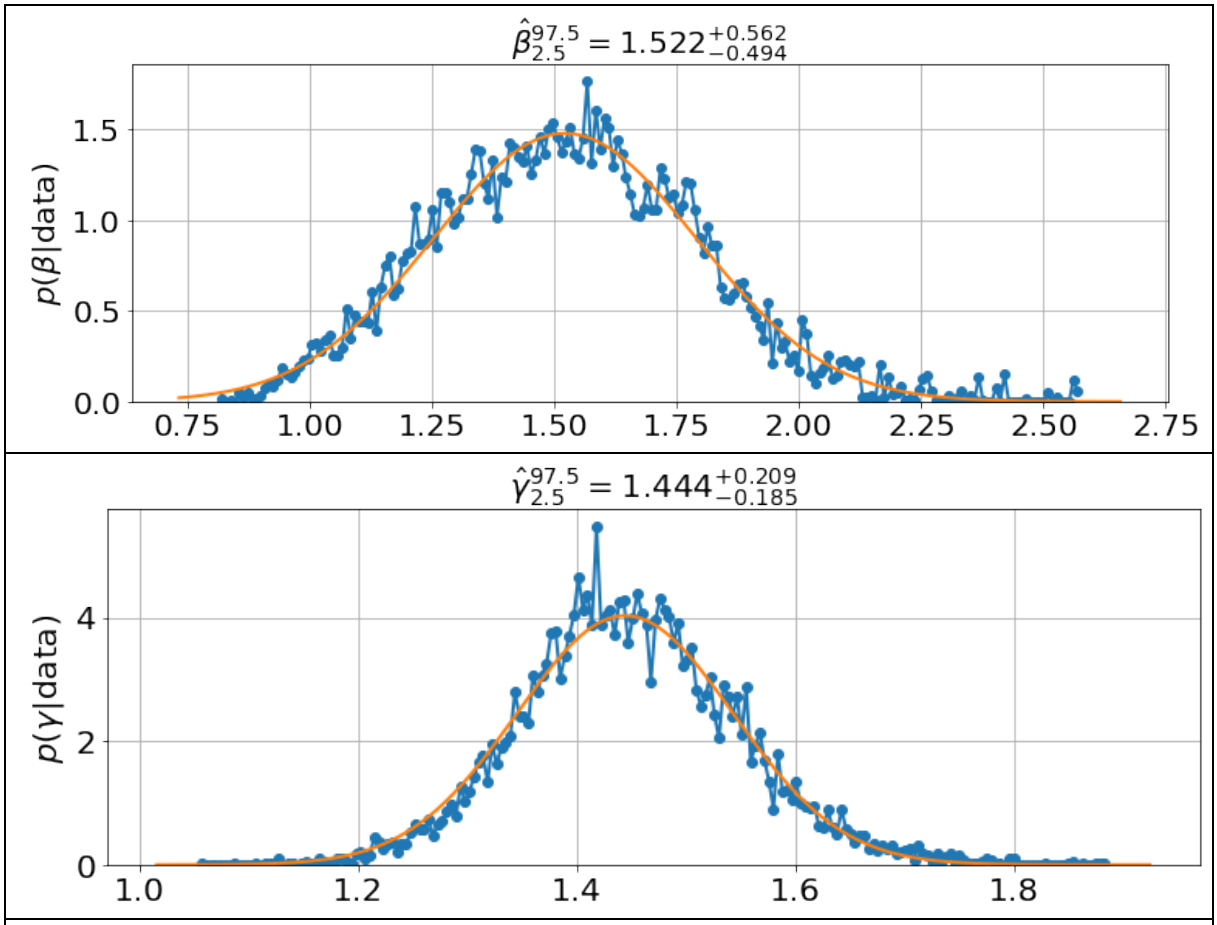

Figure 7: This is an example of an MCMC output with both β and γ. The algorithm not only returns the best fit, but a full probability distribution for the parameters. This allows us to find the best-fit βs and γs, but also see the uncertainties associated with them, which is the experimental determination of the values of parameters that impact the model's behavior. For β, in this example, the best fit is 1.522 and an uncertainty of +0.529 and -0.494. For γ, the best is 1.444 and has an uncertainty of +0.209 and -0.185.

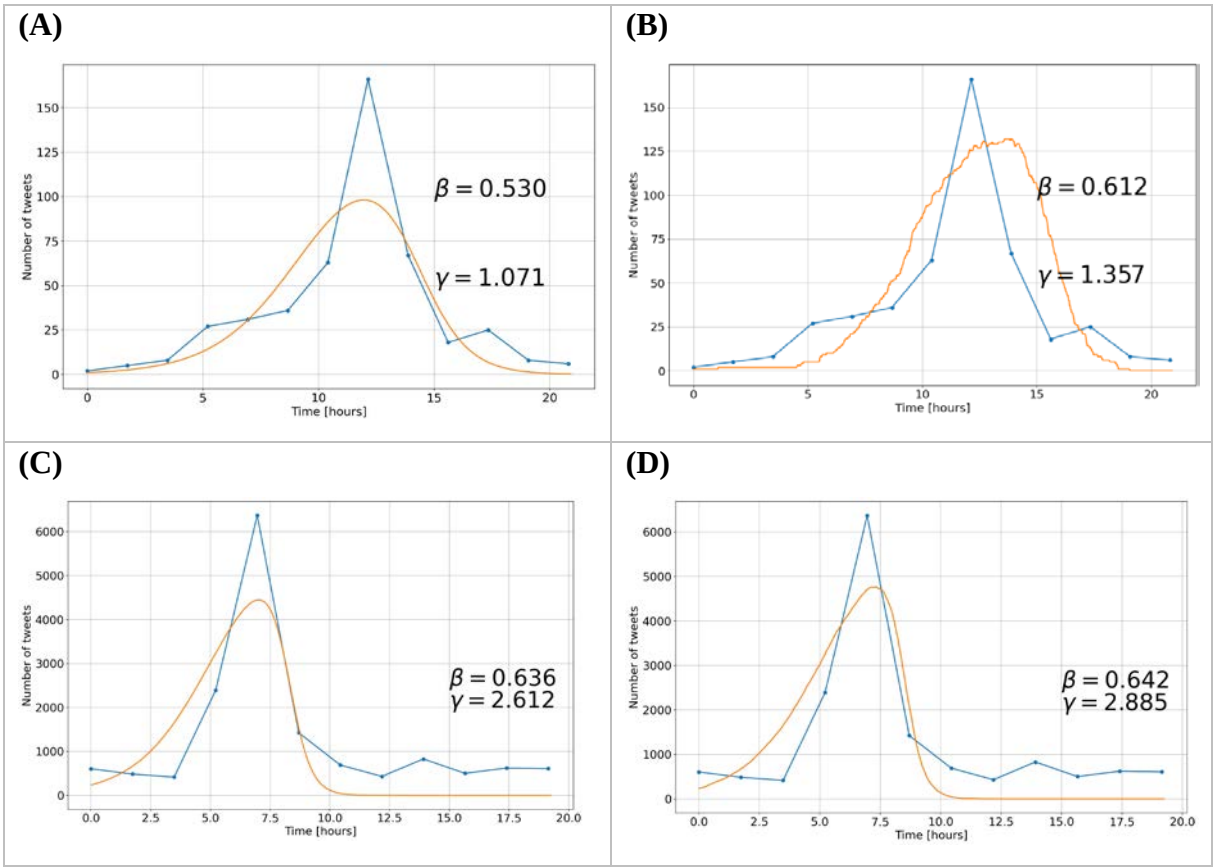

Figure 8: Some sample outputs for the SIRI model. The orange line represents the model and the blue line represents the plotted data. These four plots show data from the modern pentathlon (A) and (B) and swimming (C) and (D). The plots on the left (A and C) are the deterministic model whereas the plots on the right (B and D) are the stochastic model. Modern pentathlon was an unpopular topic with a maximum of 175 tweets. Swimming was a popular topic with a maximum of 6000 tweets.

Sport	β	γ
Gymnastics	$0.333^{+0.08}_{-0.08}$	$1.581^{+1.091}_{-0.459}$
Swimming	$0.633^{+0.158}_{-0.156}$	$2.901^{+1.457}_{-0.942}$
Steeplechase	$0.934^{+0.200}_{-0.156}$	$1.471^{+0.685}_{-0.395}$
Modern Pentathlon	$0.596^{+0.302}_{-0.188}$	$1.405^{+1.061}_{-0.364}$
Curling	$0.496^{+0.239}_{-0.200}$	$0.842^{+0.703}_{-0.200}$

Figure 9: Shows all the βs and γs for all sports ran with the stochastic SIRI model. In this model steeplechase had the highest β which was 0.934 with and uncertainty of +0.200 and -0.156. Gymnastics had the lowest β which was 0.333 with an uncertainty of +0.08 and -0.08. The highest γ was swimming which was 2.901 with an uncertainty of +1.457 and –0.942. The lowest γ Curling which was 0.842 with an uncertainty of +0.703 and – 0.200.

Sport	β	γ
Gymnastics	$0.355^{+0.28}_{-0.19}$	$1.471^{+0.45}_{-0.32}$
Swimming	$1.050^{+0.64}_{-0.46}$	$2.757^{+0.64}_{-0.61}$
Steeplechase	$1.522^{+0.56}_{-0.40}$	$1.444^{+0.20}_{-0.18}$
Modern Pentathlon	$0.478^{+0.27}_{-0.27}$	$1.013^{+0.54}_{-0.28}$
Curling	$0.476^{+0.22}_{-0.26}$	$0.834^{+0.33}_{-0.21}$

Figure 10: Shows all the βs and γs for all sports ran with the dynamic SIRI model. In this model steeplechase had the highest β which was 1.522 with and uncertainty of +0.56 and -0.46. Gymnastics had the lowest β which was 0.355 with an uncertainty of +0.28 and -0.19. The highest γ was swimming which was 2.757 with an uncertainty of +0.64 and – 0.61. The lowest γ Curling which was 0.834 with an uncertainty of +0.33 and – 0.21.

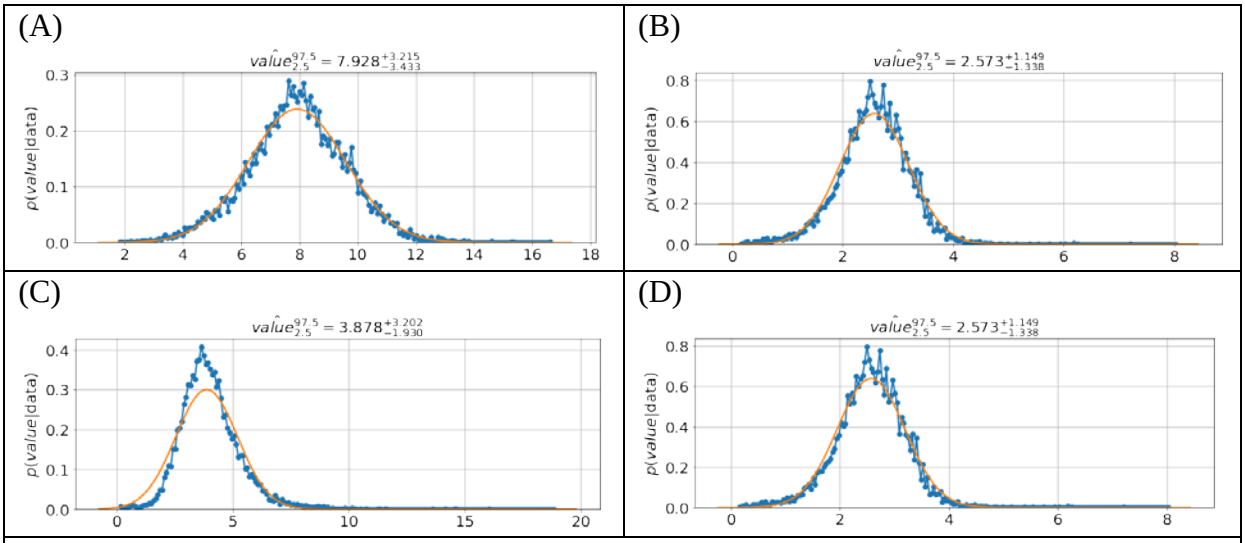

Figure 11: The R_0 for the topics we explored. These four plots show data from (A) steeplechase, (B) Swimming, (C) Gymnastics, and (D) Modern Pentathlon. The best fit R_0 for steeplechase is 7.928 and an uncertainty of +3.215 and -3.433. The best fit R_0 for swimming is 2.573 and an uncertainty of +1.149and -1.338. The best fit R_0 for gymnastics is 3.878 and an uncertainty of +3.202 and -1.930. The best fit R_0 for modern pentathlon is 2.573 and an uncertainty of +1,149and -1.338.

APPENDICES

Appendix A – Literature Review

What is "infectiousness"?

As defined by the Oxford Dictionary, infectiousness is how easily a disease is passed from one person to another (*Infectiousness*). Similarly, this study defines infectiousness as how easily an idea is passed from one person to another. When an individual is "infected" with an idea, it means that individual has adopted that idea and will in turn transmit it to others (Goffman & Newill, 1964; Jenders et al., 2013). This leads to an intellectual "epidemic" where certain ideas run rampant for a period of time (Goffman & Newill, 1964). In a study done on Twitter hashtags, they measured infectiousness by the rate of how fast the idea has spread throughout the Twittersphere. This study also uses that same definition of infectiousness. By using this idea of infectiousness, this study hopes to identify infectious people and ideas.

How tweets are spread

Twitter has fundamentally changed the way we spread, view, and produce news (Hu et al., 2012). Viral tweets become popular for different reasons; however, the fundamental components that make a tweet go viral are similar. In a study done on the ALS Ice Bucket challenge, it was found that there were three components that lend itself to a tweets virality (Pressgrove et al., 2018). The first is that the information of the tweet must give the potential tweeter social currency or make them perceive they will get social currency. For example, if family or friends participate in a challenge or idea, a person is more likely to also partake (Burgess et al., 2018; Pressgrove et al., 2018). Second, there must be a strong emotion elicited from the tweet. This will be preferably a strong emotion, like hope or sadness and there must be a high level of believability (Hu et al., 2012a; Kilgo et al., 2020; Lenoir et al., 2017; Pressgrove et al., 2018). Third, Twitter heavily relies on herd mentality, specifically actions that are easily imitable or highly visible, which makes people more likely to interact with a tweet (Pressgrove et al., 2018). These three elements help tweets spread information;

however, the people who tweet or retweet the information are important as well (Tulu et al., 2018a). The retweet option enables these tweets to reach at least 1,000 users even if they are not an "opinion leader". An "opinion leader" is a small concentrated group of individuals who first start the dissemination of information and enable the spread even further (Hu et al., 2012; Jenders et al., 2013). The removal of an "opinion leader" significantly decreases the spread of information and decreases the connection that information has with the other readers (Hu et al., 2012). These methods introduce the way and how tweets and information is spread through the Twittersphere.

Tweets used on diseases

Within the context of healthcare, existing research utilizes Twitter data to map out when a disease will peak or where the current spread of the disease is. One such study focused on influenza, scraped tweets including words about the influenza or H1N1. They trained a linear Support Vector Machine (SVM) model using the Twitter data they collected and influenza data collected by the FDA (Signorini et al., 2011). Another study utilized a real time disease surveillance system to monitor influenza and cancer activity (Lee et al., 2013). Although they were not able to predict the future of where the disease spread, they were able to provide real time estimates of how many people were currently infected. Both these studies used Twitter to perform surveillance on disease and tried to see whether Twitter could predict where the diseases were and when they would peak. These models were relatively successful in their ability to predict where the disease was and the peak would occur (Bodnar & Salathé, 2013).

Compartmental models

The compartmental model is a general modelling technique most used in the mathematical modeling of infectious diseases. It has since been adapted to analyze the spread of ideas. The simplest and thus the most common compartmental model is the SIR epidemiology model. As seen in equation 1, this model categorizes the members of interest into three groups Susceptible (S), Infected(I), and Recovered (R) while also showing the rates of infectiousness (β) and recovery (γ) (Skaza & Blais, 2017).

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = +\frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = +\gamma I \quad (1)$$

There are multiple variations of the SIR that can be used for different scenarios. Another version of the SIR model is the SIRI model, where even after the person is recovered, they can be reinfected. As seen in equation 2, this model is most used for showing the infection of ideas, as a person does not need as long of a recovery time compared to that of a real infection (Skaza & Blais, 2017).

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = +\frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = +\frac{\beta IS}{N} - \vartheta IR/N \quad (2)$$

The SEIR model adds an additional category to the original three, Exposed (E). This model introduces the idea where a person can be exposed to an idea but not entirely infected or adopt the idea completely(Abdullah & Wu, 2011). Another version is the SIS model which removed the recovery category. This model states that there is no recovery between when you are infected or susceptible to new ideas (Smyth et al., 2020). Although there are many forms of the SIR model, they all have one basic purpose, which is to be able to categorize the different types of information spreaders and receivers(Abdullah & Wu, 2011).

Network models

Another way of analyzing data is with network models, a database model where there is a flexible way of representing objects and their relationships with each other. When applying a network model to social media the traditional measurements are Degree, Betweenness, Community Based Centrality, PageRank, Eigenvector and LE(Tulu et al., 2018). These measurements aim to try and figure out why and how a person gets infected by a tweet. When using an SIR model in conjunction with the network model, it can be used to label nodes to show where the news will spread based on the nodes and their proximity to the infected node (Tulu et al., 2018). In a study that applied the network model with and SIR model, researchers found that they were able to select specific nodes that were influencing the spread of information (Tulu et al., 2018). Another study using the Spatial-Temporal Event Spread Model an SIR model looked at a case study about a gas shortage that took place in the New Jersey area (Ganti et al., 2013). This model they used was specific to the location in which the gas shortage affected. They found, that in this instance, disease modeling was not useful in the prediction and analysis of where the information spread from (Ganti et al., 2013). In this scenario, it seems like the effect of the information was too concentrated to one area and did not follow the usual SIR model.

Stochastic Models

After the application of the compartmental mode, a stochastic model must be used to help further analyze the spread of information. Stochastic models help show how the movement between S to I or I to R is probabilistic and is a random process (Hochreiter & Waldhauser, 2013; Skaza & Blais, 2017). The Markov Chain Monte Carlo algorithm (MCMC) is used to fit data and optimize the SIR model(Abdullah & Wu, 2011; Deng & Wang, 2019; Skaza & Blais, 2017). In studies conducted on twitter hashtags, researchers have combined both the SIR model and the MCMC algorithm to quantify the trendiness of a hashtag. Similarly, other studies like the previous found that tweets fall into two categories, slightly infectious or very infectious topics (Skaza & Blais, 2017). In another study, researchers also used the MCMC algorithm coupled with the SEIR model to predict trend dynamics on Twitter(Abdullah &

Wu, 2011). This study was able to successfully pinpoint when there was a change in Twitter dynamics and show that disease dynamics do not only apply to disease but to the spread of information as well (Abdullah & Wu, 2011).

Summary of Literature Review

Millions of people contribute to the Twitterverse daily and everyday tweets go viral. This study aims to pinpoint the factors that influence a tweets virality and see how ideas are spread. By using the Gillespie model along with an SIR model, we hope to categorize users and ideas that are more infectious than others. Furthermore, we hope to assess how accurately an epidemiology model can predict and identify the infectiousness of a tweet. This study will give a deeper understanding into viral internet culture.

REFERENCES

Abdullah, S., & Wu, X. (2011a). An Epidemic Model for News Spreading on Twitter. *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 163–169. https://doi.org/10.1109/ICTAI.2011.33

Abdullah, S., & Wu, X. (2011b). An Epidemic Model for News Spreading on Twitter. *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 163–169. https://doi.org/10.1109/ICTAI.2011.33

Bodnar, T., & Salathé, M. (2013). Validating models for disease detection using twitter. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, 699–702. https://doi.org/10.1145/2487788.2488027

Burgess, A., Miller, V., & Moore, S. (2018). Prestige, Performance and Social Pressure in Viral Challenge Memes: Neknomination, the Ice-Bucket Challenge and SmearForSmear as Imitative Encounters. *Sociology*, *52*(5), 1035–1051. https://doi.org/10.1177/0038038516680312

Deng, X., & Wang, X. (2019). The Application of Gillespie Algorithm in Spreading. *Proceedings of the 3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)*. Proceedings of the 3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019), Dalian, China. https://doi.org/10.2991/icmeit-19.2019.110

Ganti, R., Srivatsa, M., Liu, H., & Abdelzaher, T. (2013). Spatio-temporal Spread of Events in Social Networks: A Gas Shortage Case Study. *MILCOM 2013 - 2013 IEEE Military Communications Conference*, 713–718. https://doi.org/10.1109/MILCOM.2013.127

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, *81*(25), 2340–2361. https://doi.org/10.1021/j100540a008

Goffman, W., & Newill, V. A. (1964). Generalization of Epidemic Theory: An Application to the Transmission of Ideas. *Nature*, *204*(4955), 225–228. https://doi.org/10.1038/204225a0

Hochreiter, R., & Waldhauser, C. (2013). A Stochastic Simulation of the Decision to Retweet. In P. Perny, M. Pirlot, & A. Tsoukiàs (Eds.), *Algorithmic Decision Theory* (Vol. 8176,

pp. 221–229). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41575-3_17

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., & Ma, K.-L. (2012a). *Breaking news on twitter*. 4.

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and predicting viral tweets. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, 657–664. https://doi.org/10.1145/2487788.2488017

Kilgo, D. K., Lough, K., & Riedl, M. J. (2020). Emotional appeals and news values as factors of shareworthiness in Ice Bucket Challenge coverage. *Digital Journalism*, *8*(2), 267–286. https://doi.org/10.1080/21670811.2017.1387501

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 591. https://doi.org/10.1145/1772690.1772751

Lee, K., Agrawal, A., & Choudhary, A. (2013). *Real-time disease surveillance using Twitter data: Demonstration on flu and cancer*. 4.

Lenoir, P., Moulahi, B., Azé, J., Bringay, S., Mercier, G., & Carbonnel, F. (2017). Raising Awareness About Cervical Cancer Using Twitter: Content Analysis of the 2015 #SmearForSmear Campaign. *Journal of Medical Internet Research*, *19*(10), e344. https://doi.org/10.2196/jmir.8421

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). *The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior*. 10.

Pressgrove, G., McKeever, B. W., & Jang, S. M. (2018a). What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing*, *23*(1), e1586. https://doi.org/10.1002/nvsm.1586

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, *6*(5), e19467. https://doi.org/10.1371/journal.pone.0019467

Skaza, J., & Blais, B. (2017). Modeling the infectiousness of Twitter hashtags. *Physica A: Statistical Mechanics and Its Applications*, *465*, 289–296. https://doi.org/10.1016/j.physa.2016.08.038

Smyth, M., Buntain, C., Dwyer, D., Finn, J., Jones, J., Garland, J., & Egan, M. (2020). *Information Processing on Social Media Networks as Emergent Collective Intelligence*. 4.

Tulu, M. M., Hou, R., & Younas, T. (2018b). Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network. *IEEE Access*, *6*, 7390–7401. https://doi.org/10.1109/ACCESS.2018.2794324