Mathematical Modeling of Trending Topics on Twitter

BY Jonathan S. Skaza

ADVISORS • Brian S. Blais

Submitted in partial fulfillment of the requirements for graduation with honors in the Bryant University Honors Program

APRIL 2015

Table of Contents

Abstract

Created in 2006, Twitter is an online social networking service in which users share and read 140-character messages called Tweets. The site has approximately 288 million monthly active users who produce about 500 million Tweets per day. This study applies dynamical and statistical modeling strategies to quantify the spread of information on Twitter. Parameter estimates for the rates of infection and recovery are obtained using Bayesian Markov Chain Monte Carlo (MCMC) methods. The methodological strategy employed is an extension of techniques traditionally used in an epidemiological and biomedical context (particularly in the spread of infectious disease). This study, which addresses information spread, presents case studies pertaining to the prevalence of several "trending" topics on Twitter over time. The study introduces a framework to compare information dynamics on Twitter based on the topical area as well as a framework for the prediction of topic prevalence. Additionally, methodological and results-based comparisons are drawn between the spread of information and the spread of infectious disease.

Introduction

Twitter (`twitter.com`) is a popular social networking website that allows users to both send and read 140-character messages known as Tweets. Twitter was created in 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass and was granted corporation status on April 19th, 2007. The social networking site has approximately 288 million monthly active users that produce an average of about 500 million Tweets per day. Furthermore, the California-based website supports over 35 languages and 77% of accounts are held outside of the United States (*About Twitter, Inc.*, 2015).

Twitter serves as a place for users to share anything and everything on their minds—news stories, ideas, quotes, lyrics, etc. The company defines a Tweet as "an expression of a moment or idea…[which] can contain text, photos, and videos" (*About Twitter, Inc.*, 2015). Users of Twitter are able to embed hashtags within their tweets by using the hash character (i.e., #). Hashtags are metadata tags which allow Tweets containing the text following the hash character to be grouped together. From there, it is possible for users to query certain hashtags to see what is being discussed throughout the site. The Twitter site even contains a panel of "Trending Topics"—i.e., hashtags and topics that have become very popular in a short period of time.

Hashtags can also prove useful for researchers in need of categorizing or grouping Tweets. While it is also possible to filter Tweets by words or phrases, such an approach can be problematic. For instance, a researcher interested in exploring the degree of happiness on Twitter may search for Tweets containing the word "happy". While this strategy will return Tweets from people expressing sentiments such as "I am *happy*", it will also return messages in the category of "I am not *happy*". Sentiment analysis techniques are needed to rectify this issue (Agarwal et al., 2011). Using hashtags to filter Tweets hedges against the need to address such concerns; a person who inserts "`#happy`" into his/her Tweet is likely happy. However, the volume of Tweets meeting the specific search criteria will be reduced because not every "happy" Tweet, for example, will include "`#happy`". Nevertheless, to avoid problems with contradicting sentiments, I use hashtags as a proxy to study the prevalence and popularity of topics on Twitter.

Recognizing how ideas and information spread on Twitter is important in many respects. From a business perspective, understanding the dynamics of information diffusion and virality is important to professionals marketing products—crafting messages that can stick and persist for lengthy periods can be of tremendous value for companies. In fact, there have been several popular books concerning this topic, including Gladwell (2000) and Heath and Heath (2008). Similarly, being able to predict the lifespan of a message or Twitter hashtag could prove useful for business professionals. More generally, better understanding how people share and spread information on social networking sites such as Twitter is another step in understanding human behavior in the new age

of information.

This study attempts to quantify the spread of certain hashtags on Twitter. Using the mathematical and statistical methodology described below, one can estimate the rates of infection and recovery for a particular trending topic. Furthermore, with slight data processing, the same methodology can be used in a predictive context. This study proceeds to introduce the dynamical modeling strategy relied upon and to provide an overview of the existing literature concerning information dynamics on Twitter. Subsequently, the analytical strategy used to quantify the propagation of trending topics on Twitter is introduced. Following the overview of this approach, the analytical strategy is applied to a number of case studies. Finally, there is a discussion concerning results and potential avenues for future work.

Previous Work

Mathematical models have been used in the prediction, control, and analysis of epidemic phenomena—most notably, the spread of infectious disease throughout a population—since the advent of the susceptible, infected, and recovered (SIR) model (Kermack & McKendrick, 1927). These types of epidemic models are featured in studies concerning measles (Grais et al., 2006; Kuniya, 2006; McGilchrist et al., 1996; Tuckwell & Williams, 2007) and influenza (Coelho et al., 2011; Hooten et al., 2010; Li et al., 2009; Tuckwell & Williams, 2007), among others. The basic SIR model describes the dynamical process of disease by categorizing members of the population of interest as either susceptible (S), infected (I), or recovered (R), while incorporating rates of infectiousness (β) and recovery (γ). Figure 1 illustrates a model diagram.

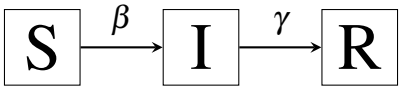


Figure 1: SIR compartmental model

Members of the population transition to and from different compartments based on the system of differential equations presented in Equation 1.

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= +\beta SI - \gamma I \\ \frac{dR}{dt} &= +\gamma I \end{aligned} \tag{1}$$

The SIR framework allows for many extensions depending on the context of the disease or problem being studied. One popular adaptation is the susceptible, infected, and susceptible (SIS) model, which is relevant when the infectious process being studied does not have any long-lasting immunity after infection. See Anderson and May (1979a), Anderson and May (1979b), Brauer and Castillo-Chavez (2001), and the *Epidemic model* and *Compartmental models in epidemiology* Wikipedia pages for further reading on the foundations of the SIR model and its extensions.

A relatively new strategy in the field of epidemic modeling is to develop a statistical model for the parameters of the dynamical model. I rely heavily upon this method in the present work. The strategy entails using Markov Chain Monte Carlo (MCMC) techniques to estimate the posterior probabilities of the epidemic model's parameters (e.g., β and γ) (Coelho et al., 2011; Witkowski & Blais, 2013).

As mentioned above, much of the previous work featuring epidemic modeling techniques concerns the transmission of an infectious disease. However, epidemic models may just as well be applied to capture the transmission of some other trait from individual to individual or group to group. A trait may take the form of a genetic characteristic, a cultural phenomenon, an addictive activity, the gain or loss of information, etc. (Brauer & Castillo-Chavez, 2001).

In this regard, Bettencourt et al. (2006) apply several epidemic modeling techniques to analyze the spread of an idea, Feynman diagrams, through the scientific literature during the mid-20th century. Bettencourt et al. (2006) utilize a basic SIR model and explore extensions germane to the spread of ideas in particular by including incubator and skeptic classes in the dynamical model. Building off the susceptible, incubator, infected, and skeptic (SEIZ) methodology introduced in Bettencourt et al. (2006), Jin et al. (2013) are the first to apply the SEIZ methodology to news and rumors on Twitter. Similarly, Zhao et al. (2012) present another extended SIR model—the susceptible, infected, hibernator, and removed (SIHR) model—which is germane to the spread of rumors in social networks.

While the literature illustrates that epidemic modeling techniques are robust when applied to studies of information propagation (specifically in regard to Twitter), others have taken different ap-

proaches when modeling this process. For example, Yang and Counts (2010) use survival analysis. Moreover, several have studied information diffusion on Twitter at the network level (Bakshy et al., 2011; Lerman & Ghosh, 2010; Weng et al., 2012, 2013; Wu et al., 2011).

Analytical Strategy

The first stage of the present study's analysis of Twitter consisted of developing a database of hashtags versus time, as displayed in the Sample Data exhibit in the Appendix. Such a database was developed using one of Twitter's application programming interfaces (APIs) and the Python programming language. Twitter has a number of different APIs; however, the class of streaming APIs was most applicable to this study, as Tweets were collected in near real-time.[1]

Hence, by accessing the streaming API and performing a filter, I collected a sample of time-stamped tweets, each of which contained the hash character. I again used Python to develop an interactive widget intended to identify topics that experienced a period of trending during the data collection phase. Figure 5 in the Widget Software section of the Appendix provides screenshots of the data processing procedure performed on a particular hashtag, `#Obama`. Figure 5a displays the `#Obama` counts in one-second bins (alternatively referred to as a one-second window) over the entire data collection period. The widget was used to obtain data corresponding to the plot in Figure 5b. In this particular example, Tweets were counted in 1,000-second bins over a much shorter timescale to more adequately capture the system's dynamics.[2] A table of the various window selections applicable to the present study can be found in the Window Selections exhibit in the Appendix.

Once a proper dataset was obtained, I proceeded to model the dynamics of several trending topics. The modeling process consisted of two-steps:

1. Specify a dynamical model to emulate the trending phenomenon.
2. Specify a Bayesian statistical model to obtain best estimates of—as well as to capture uncertainty in—the various parameters in the dynamical model.

In the present study, each dynamical model presented is of the SIR variety. Other specifications were explored, but the basic SIR model proved sufficient in capturing the dynamical process for

[1] See `https://dev.twitter.com/streaming/overview` for documentation.

[2] Specifically, there is still one datum for each second; however, the datum corresponds to the count of `#Obama` Tweets in the previous 1,000 seconds.

each hashtag considered.[3] Hence, each model is of the form detailed in Equation 1. The SIR simulations were implemented using a wrapper around the `odeint` function in Python's SciPy library.

For each dynamical model, a statistical model was attached to obtain best estimates and credible intervals for β, γ, the initial susceptible population (S_0), and the initial infected population (I_0). Uniform prior probabilities were assumed for the parameters and a Python implementation of the affine invariant MCMC ensemble sampler developed by Goodman and Weare (2010) was used to obtain posterior probabilities for the parameter values as well as the correlation between parameters (Foreman-Mackey et al., 2013).

Case Studies

As a first case study exhibiting the methodology described in the Analytical Strategy section, again consider `#Obama`, which could be categorized as a news and/or current event topic. By using the aforementioned identification software, the data were processed into 1,000-second windows and the trending period was determined to persist for approximately 191 minutes. After specifying the appropriate uniform priors for β, γ, S_0, and I_0 of the SIR model, the MCMC algorithm was run to obtain posterior probabilities. The resulting point estimates for β and γ are displayed in Table 1. A step-by-step outline of the mechanics germane to this example can be found in the #Obama Supplement section of the Appendix and a plot of the resulting simulation can be seen in Figure 2.

Parameter	Median Estimate	95% Credible Interval
β	0.4711	(0.4234, 0.5190)
γ	0.0997	(0.0913, 0.1101)

Table 1: β and γ estimates for `#Obama`

[3]This is not to say that the SIR model is suitable for every hashtag; however, the SIR model performed well on each hashtag analyzed in this study.

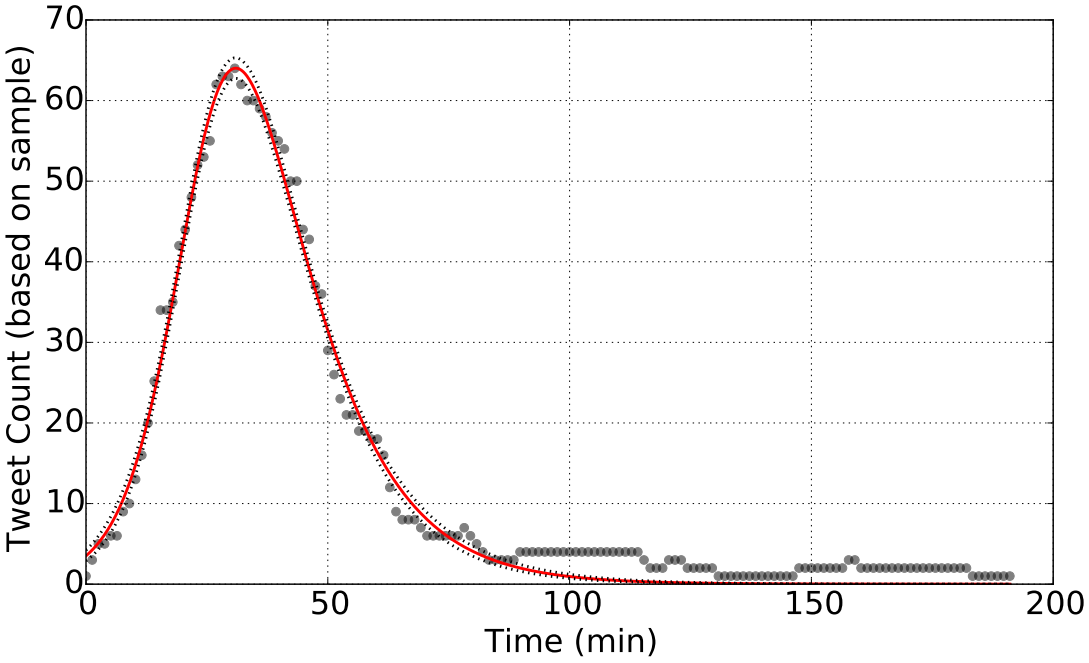


Figure 2: Simulation for `#Obama`: median estimate and 95% credible interval versus actual data

As another example, consider `#thewalkingdead`, a hashtag that has roots in the world of entertainment and pertains to the television program titled *The Walking Dead*. The data for this trending hashtag spanned only about 15 minutes. After implementing a 200-second windowing scheme and applying the dynamical and statistical methodology to the data, the estimates presented in Table 2 were obtained. A plot of the corresponding simulation is included in #thewalkingdead Simulation section of the Appendix.

Parameter	Median Estimate	95% Credible Interval
β	0.3840	(0.3263, 0.4849)
γ	1.4971	(1.1876, 1.7658)

Table 2: β and γ estimates for `#thewalkingdead`

Finally, consider `#CWC15`. This hashtag was popular in early March 2015 (the time in which the data were collected) as the 2015 Cricket World Cup (i.e., CWC) was in full effect. Hence, while the previous two examples considered hashtags relating to political news and television entertainment, respectively, `#CWC15` possesses roots in the world of sports. The trending period spanned about 475 minutes and reflects a 900-second binning strategy. Parameter estimates are included in Table 3

and a plot of the median fit of the simulation along with upper and lower bounds is included in the #CWC15 Simulation exhibit of the Appendix.

Parameter	Median Estimate	95% Credible Interval
β	0.1226	(0.0931, 0.1607)
γ	0.2435	(0.1857, 0.3208)

Table 3: β and γ estimates for `#CWC15`

2013-14 U.S. Flu Season

As discussed above, the methodology employed in this study is renowned for its epidemiological applications. Such applications include modeling the spread of infectious diseases, such as influenza. To illustrate the similarities and differences in applying the SIR methodology to information diffusion and to the spread of infectious disease, consider an SIR model used to analyze the 2013-2014 flu season and trained on Google Flu Trends data.[4]

Unlike the Twitter data, the Google Flu Trends data did not require a binning procedure to capture the system dynamics. Instead, the data were already aggregated into one-week bins with each bin representing a weekly count (from September 1st - April 6th) of Influenza-Like Illness (ILI) cases per 100,000 population. Using methodology identical to what was used to model trending topics on Twitter, I obtained estimates for β and γ as applied to the influenza dynamics spanning the period of September 1st - April 6th. The parameter estimates, obtained using the weekly data, are presented in Table 4 and simulation results are located in the 2013-14 U.S. Flu Season Simulation portion of the Appendix.

Parameter	Median Estimate	95% Credible Interval
β	0.4505	(0.3549, 0.6878)
γ	0.5592	(0.3593, 0.7192)

Table 4: β and γ estimates for 2013-14 flu season

While addressed further in the Discussion section below, it is worth noting that the dynamics of hashtag diffusion through Twitter and the dynamics of flu spread through a population are quite similar. It appears that both dynamical processes can be adequately captured using the SIR framework. The only difference is the time interval in which each process occurs.

[4]Data can be found at `https://www.google.org/flutrends/us/#US`.

Prediction Application

In addition to modeling topics that have trended in the past, it is possible to use the SIR model as a prediction tool. The ability to predict trending topics on Twitter could be a precious asset to companies. For instance, being able to predict when a slogan is about to "go viral" or, conversely, when it is most likely to fade out of the mainstream could be of great value to a marketing team. A team of researchers has developed an algorithm, intended to predict trending topics on Twitter, which is more sophisticated than the basic SIR model (Steadman, 2012). Nevertheless, the SIR model itself is viable in predictive settings.

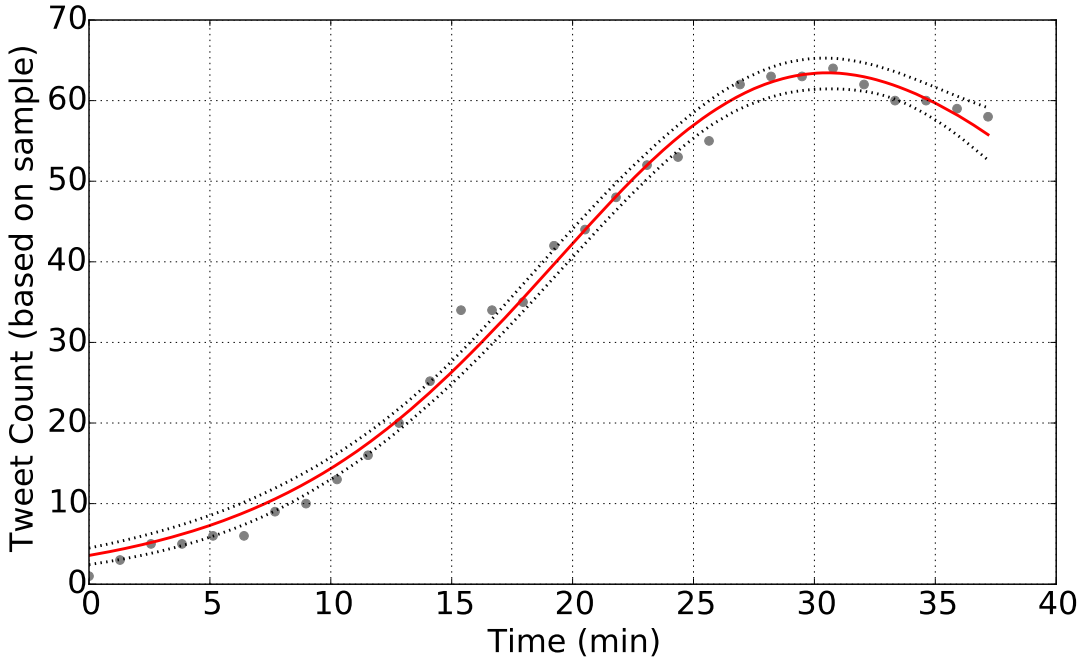
To illustrate, again consider the `#Obama` data. The SIR methodology was applied to a subset of the data, hereafter referred to as the training set. As displayed in Figure 3a, the training set contained approximately the first 38 (out of 191) minutes of `#Obama` data. Again, the dynamical and statistical models were used to parameterize the information diffusion process. Once the parameter estimates, based on the training dataset, were captured, the remaining data were used as a validation tool. Thus, after fitting the model to the 38-minute training set, the simulation was run over the full 191-minute interval. If of interest, one could then use various tests to evaluate the success of the model—the central tenet of each test being to ascertain how well the simulation run over the longer period fares when compared to the validation data. Here, I simply include a graphical assessment of the `#Obama` prediction in Figure 3b.

Discussion

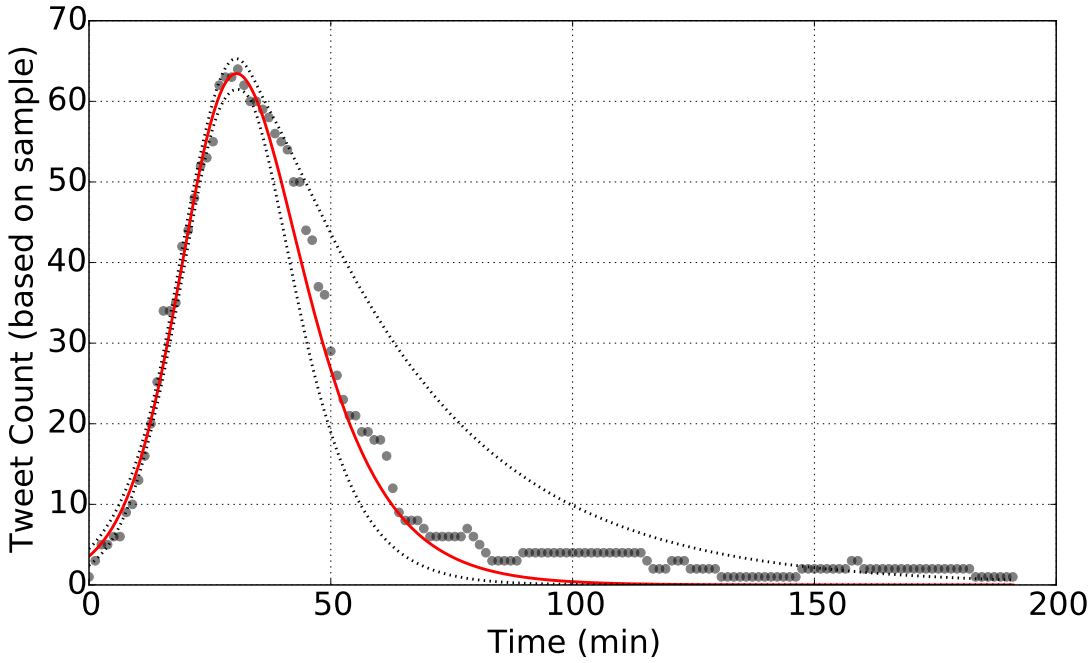
This paper presented a novel application for the widely used SIR model. The simulations for the three hashtag case studies seem to verify that the SIR methodology is indeed robust when applied to social media (particularly Twitter) meme diffusion, as Bettencourt et al. (2006) suggest. While three case studies are certainly not enough to draw Twitter-wide inferences, a few speculations can be made regarding information diffusion.

For one, it seems probable that hashtags experience a much quicker dynamic than does the flu (and likely other infectious diseases). Though the "ramp-up, peak, and ramp-down" dynamic applies to both phenomena, the process spans a much larger timeframe for the influenza data.

A second conjecture relevant to the case studies is that hashtags regarding television shows are subject to a faster dynamic than are hashtags relating to sporting events or political events. This could be attributed to the notion that television shows are constrained to 30- or 60-minute blocks, while political news or sporting events may tend to persist for longer durations.



(a) Training set fit: median estimate and 95% credible interval versus training data



(b) Simulation over longer interval: median estimate and 95% credible interval versus validation data

Figure 3: Prediction of `#Obama`

On a similar note, `#thewalkingdead` possessed an estimate for the γ parameter that was relatively large in comparison to the β estimate. This indicates that the recovery rate was much higher than was the infection rate and, consequently, that the "ramp-down" happened relatively faster than the "ramp-up". Figure 4 contains a plot which compares β and γ estimates for each hashtag. However, one should be wary when comparing the parameters of two different hashtags due to differences in the window choice.

Much more work can be done within the topic of Twitter meme diffusion. There is a need to explore different models—whether deterministic, stochastic, or of another variety—and, moreover, to compare the success of each model type. Furthermore, machine learning techniques could be very helpful in prediction problems. Additionally, it could be of interest to examine causality issues. For instance, one could ask and address the questions "Why is the recovery rate for `#thewalkingdead` what it is?" and "How could those promoting *The Walking Dead* lower the recovery rate to ensure viewers are talking about the show long past its airing?".

In reference to the methodology used in this study, there seem to be two major components in need of improvement. First, it would be useful to develop better software to identify trending hashtags. The hashtags analyzed in this paper were chosen based on visual inspection (as well as entertainment value). Second, it would be extremely useful to optimize the window selection (again, this was essentially done manually in the present study). Moreover, for the purpose of comparison, it would be useful to have a constant window for each hashtag.

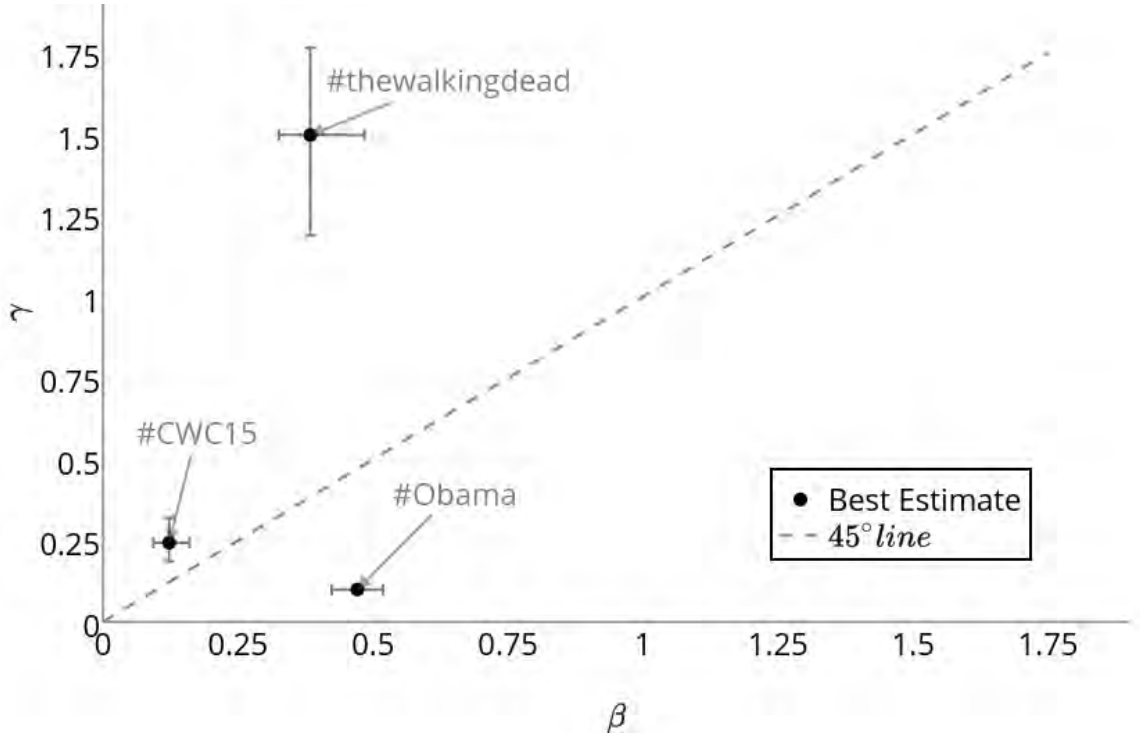


Figure 4: Plot detailing the relationship between β and γ estimates for each hashtag studied. One should avoid making absolute comparisons among different hashtags because of different window selections (refer to Window Selections).

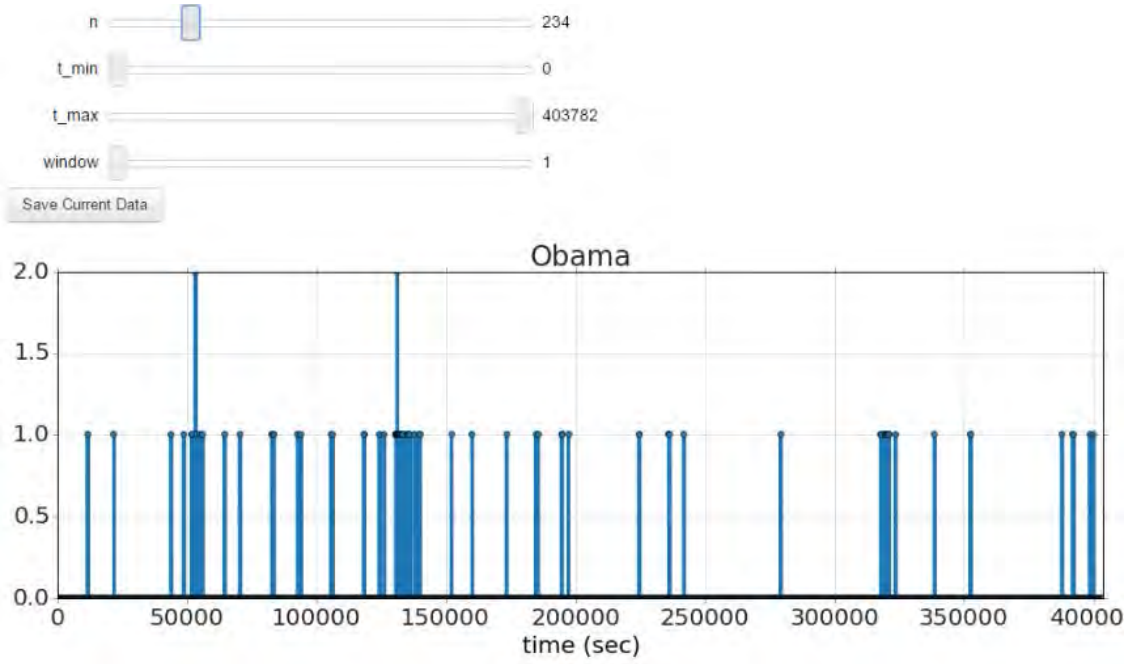
Appendix

Sample Data

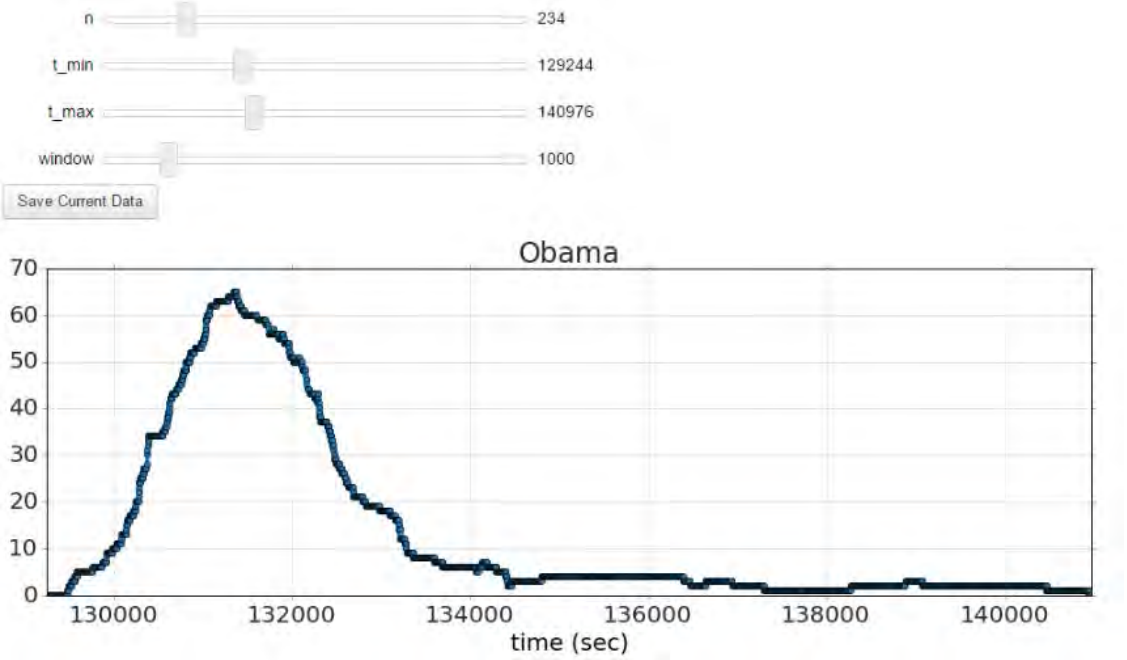
Timestamp	Hashtags
Tue Mar 10 23:02:51 +0000 2015	['iTunes', 'iPhone']
Tue Mar 10 23:02:51 +0000 2015	['Saputosoapopera', 'Y', 'Daysofourlives', 'JerrySpringer']
Tue Mar 10 23:02:52 +0000 2015	['iTunes', 'iPhone']
Tue Mar 10 23:02:53 +0000 2015	['billgates', 'wise', 'inspiration']

Table 5: A snippet of the time-stamped hashtag data

Widget Software



(a) Pre-processing



(b) Post-processing

Figure 5: Example of using the software to identify and process hashtag data

Window Selections

Hashtag	Window
`#Obama`	1,000s
`#thewalkingdead`	200s
`#CWC15`	900s

Table 6: Window selection for each hashtag

#Obama Supplement

Step 1: Process data as illustrated above in Widget Software.

Step 2: Specify SIR Model and experiment with parameters to establish reasonable bounds for uniform prior probabilities.

```
S0=500.0
sim=Simulation()
sim.add("S'=-beta*S/S0*I",S0,plot=False)
sim.add("I'=beta*S/S0*I-gamma*I",1,plot=True)
sim.add("R'=gamma*I",0,plot=False)
sim.params(S0=S0, beta=1.6,gamma=.9) # initial guess
sim.add_data(t=time,I=obama,plot=True)
sim.run(0,191)
```

Step 3: Specify prior probability distributions and run MCMC algorithm to obtain parameter estimates and credible intervals.

In this example, the prior probability distributions are as follows:

$$\beta \sim U(0,2)$$
$$\gamma \sim U(0,2)$$
$$S_0 \sim U(30,5000)$$
$$I_0 \sim U(0,10)$$

```
model=MCMCModel(sim, beta=Uniform(0,2),
               gamma=Uniform(0,2),
               initial_I=Uniform(0,10),
               initial_S=Uniform(30,5000)
               )

model.run_mcmc(10000)

# run again resampling the initial conditions
model.set_initial_values('samples')
model.run_mcmc(10000)
```

Step 4: Run simulation drawing from posterior probability distributions.

```
# 10000 simulations
for i in range(10000):
    model.draw()
    sim.run(0,191)
```

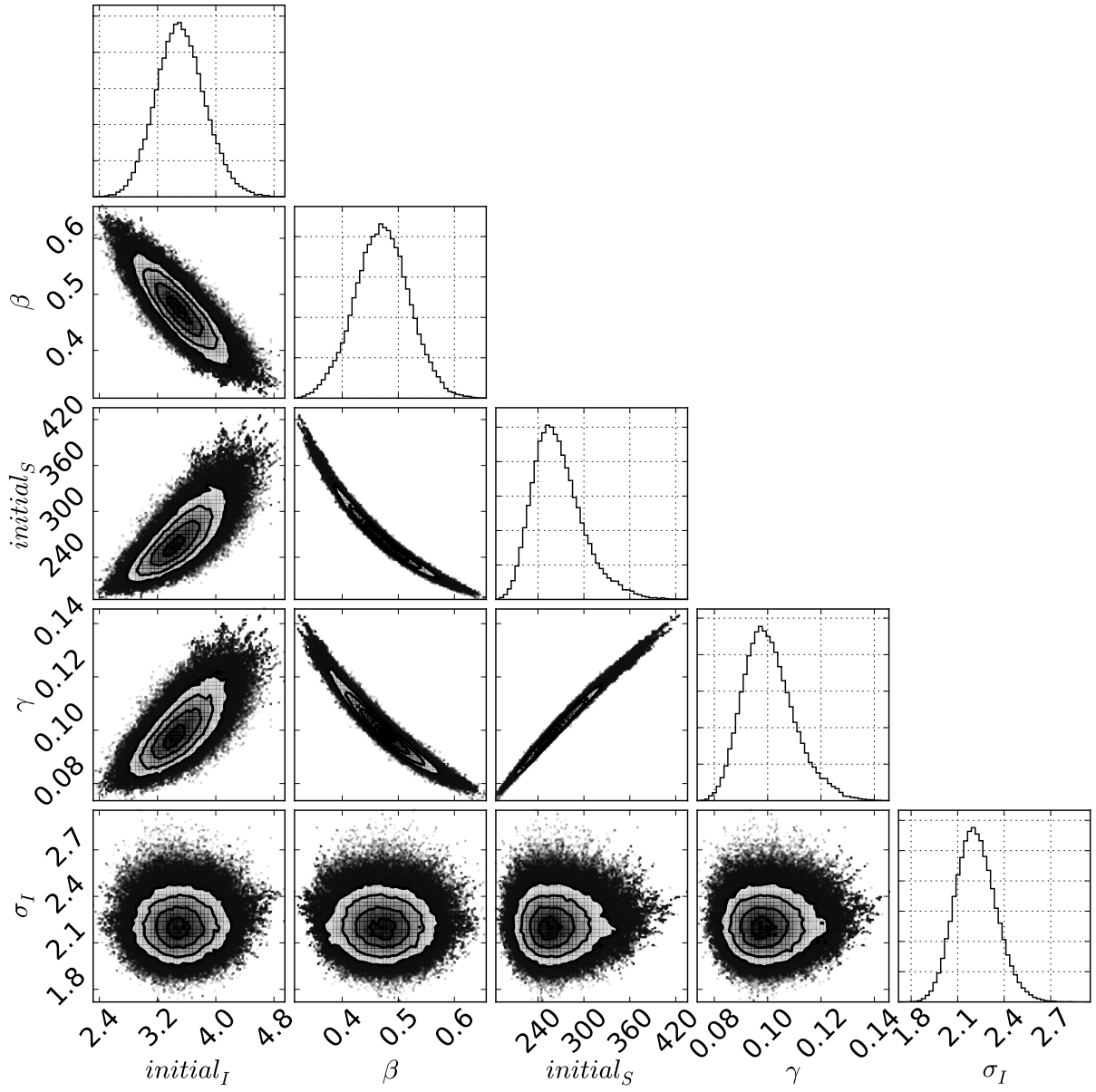


Figure 6: Corner plot illustrating posterior probability distributions as well as correlations between parameters

#thewalkingdead Simulation

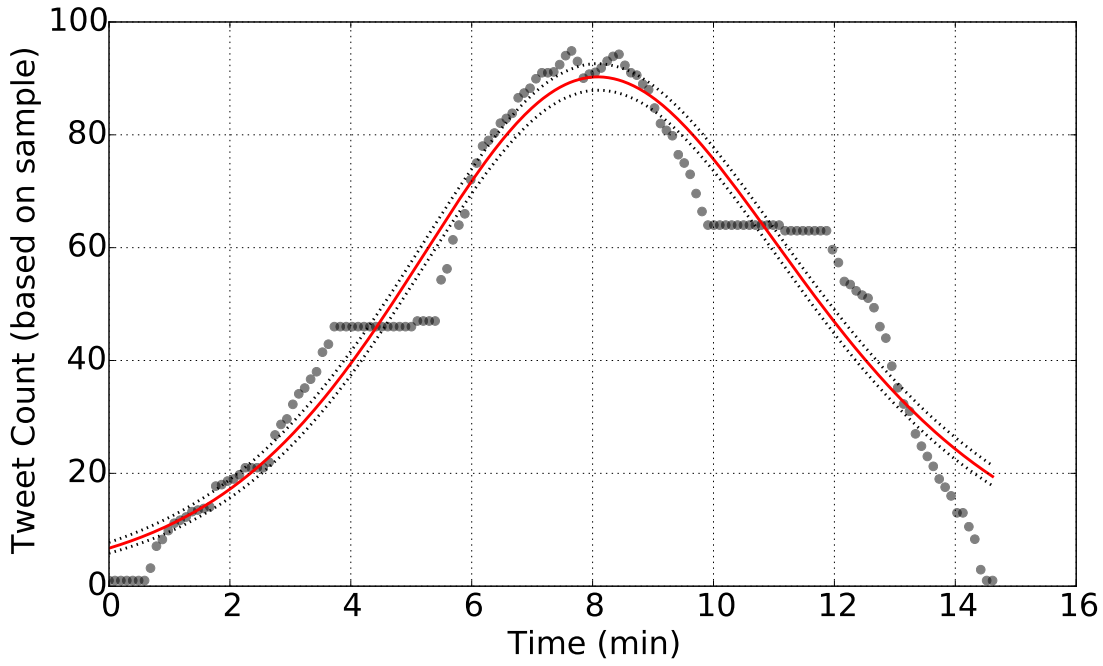


Figure 7: Simulation for `#thewalkingdead`: median estimate and 95% credible interval versus actual data

#CWC15 Simulation

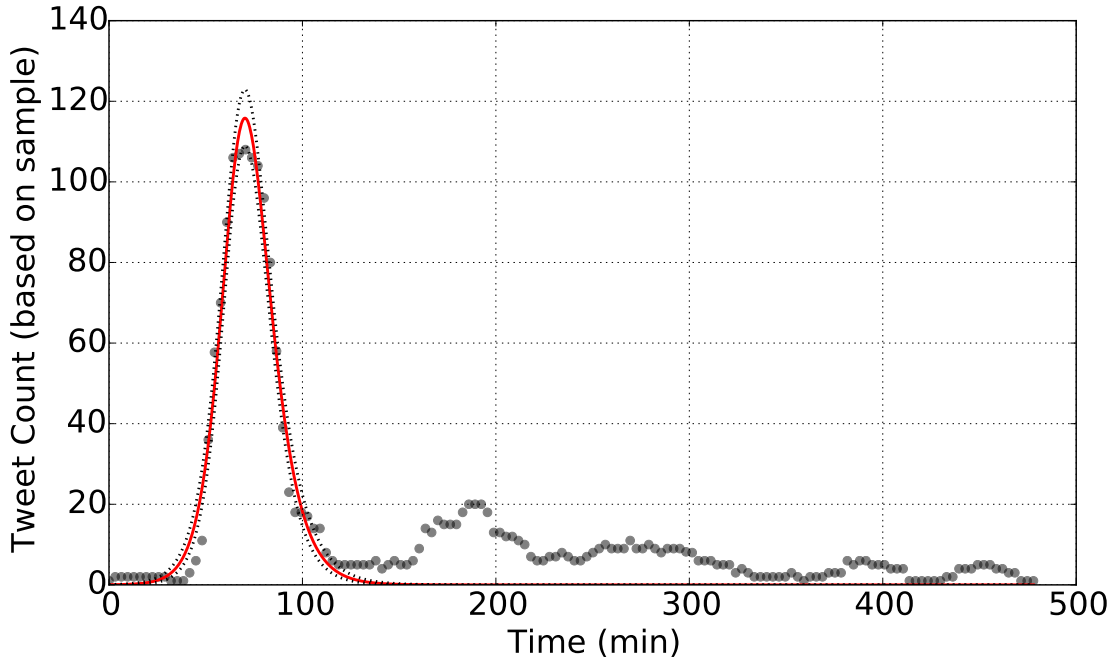


Figure 8: Simulation for `#CWC15`: median estimate and 95% credible interval versus actual data

2013-14 U.S. Flu Season Simulation

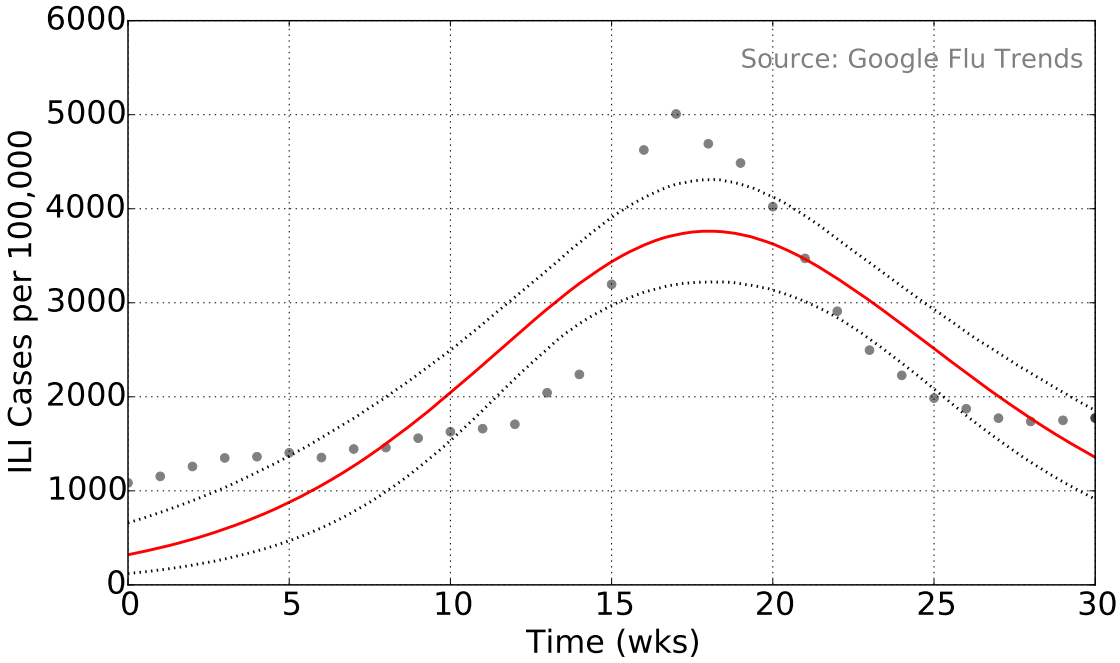


Figure 9: Simulation for 2013-14 U.S. Flu Season: median estimate and 95% credible interval versus actual data

References

About Twitter, Inc. (2015). `https://about.twitter.com`.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *LSM '11 Proceedings of the Workshop on Languages in Social Media.* Association for Computational Linguistics.

Anderson, R. M., & May, R. M. (1979a). Population biology of infectious diseases: Part I. *Nature*, *280*, 361-367.

Anderson, R. M., & May, R. M. (1979b). Population biology of infectious diseases: Part II. *Nature*, *280*, 455-461.

Bakshy, E., Hofman, J., Mason, W., & Watts, D. (2011). Everyone's an influencer: quantifying influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (p. 65-74). ACM.

Bettencourt, L. M., Cintron-Arias, A., Kaiser, D. I., & Castillo-Chavez, C. (2006). The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Physica A*, *364*, 513-536.

Brauer, F., & Castillo-Chavez, C. (2001). Mathematical models in population biology and epidemiology. In (chap. Basic ideas of mathematical epidemiology). Springer-Verlag New York, Inc.

Coelho, F. C., Codeco, C. T., & Gomes, M. G. M. (2011). A Bayesian framework for parameter estimation in dynamical models. *PLoS ONE*, *6*.

Compartmental models in epidemiology. (2015). `http://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology`.

Epidemic model. (2015). `http://en.wikipedia.org/wiki/Epidemic_model`.

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. (2013). emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, *125*. (`http://adsabs.harvard.edu/abs/2013PASP..125..306F`)

Gladwell, M. (2000). *The Tipping Point*. Little Brown.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, *5*.

Grais, R. F., Ferrari, M. J., Dubray, C., Bjornstad, O. N., Grenfell, B. T., Djibo, A., ... Guerin, P. J. (2006). Estimating transmission intensity for a measles epidemic in Niamey, Niger: lessons for intervention. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *100*, 867-873.

Heath, C., & Heath, D. (2008). *Made to Stick*. Random House.

Hooten, M. B., Anderson, J., & Waller, L. A. (2010). Assessing North American influenza dynam-

ics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, *1*, 177-185.

Jin, F., Dougherty, E., Saraf, P., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on Twitter. In *SNAKDD '13*. ACM.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, *115*, 700-721.

Kuniya, T. (2006). Global stability analysis with a discretization approach for an age-structured multigroup SIR epidemic model. *Nonlinear Analysis: Real World Applications*, *12*, 2640-2655.

Lerman, K., & Ghosh, R. (2010). Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of* 4^{th} *International Conference on Weblogs and Social Media.*

Li, X. Z., Li, W. S., & Ghosh, M. (2009). Stability and bifurication of an SIR epidemic model with nonlinear incidence and treatment. *Applied Mathematics and Computation*, *210*, 141-150.

McGilchrist, C. A., McDonnell, L. F., Jorm, L. R., & Patel, M. S. (1996). Loglinear models using capture - recapture methods to estimate the size of a measles epidemic. *Journal of Clinical Epidemiology*, *49*, 293-296.

Steadman, I. (2012). *MIT algorithm predicts Twitter trending topics up to five hours in advance.* `http://www.wired.co.uk/news/archive/2012-11/02/algorithm-predicts-twitter-trends`.

Tuckwell, H. C., & Williams, R. J. (2007). Some properties of a simple stochastic epidemic model of SIR type. *Mathematical Biosciences*, *208*, 76-97.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, *2*.

Weng, L., Menczer, F., & Ahn, Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports*, *3*.

Witkowski, C., & Blais, B. S. (2013). Bayesian analysis of epidemics - zombies, influenza, and other diseases.
(`http://web.bryant.edu/~bblais/pages/publications.html`)

Wu, S., Hofman, J., Mason, W., & Watts, D. (2011). Who says what to whom on Twitter. In *Proceedings of the* 20^{th} *international conference on World wide web* (p. 705-714). ACM.

Yang, J., & Counts, S. (2010). Predicting the speed, scale, and range of information diffusion in Twitter. *Association for the Advancement of Artificial Intelligence*.

Zhao, L., Wang, J., Chen, Y., Wang, Q., Cheng, J., & Cui, H. (2012). SIHR rumor spreading model in social networks. *Physica A: Statistical Mechanics and its Applications*, *391*, 2444-2453.