2011 23rd IEEE International Conference on Tools with Artificial Intelligence

An Epidemic Model for News Spreading on Twitter

Saeed Abdullah and Xindong Wu
Department of Computer Science, University of Vermont
Burlington, Vermont, USA
sabdulla, xwu@uvm.edu

***Abstract*—In this paper, we describe a novel approach to understand and explain news spreading dynamics on Twitter by using well-known epidemic models. Our underlying hypothesis is that the information diffusion on Twitter is analogous to the spread of a disease. As mathematical epidemiology has been extensively studied, being able to express news spreading as an epidemic model enables us to use a wide range of tools and procedures which have been proven to be both analytically rich and operationally useful. To further emphasize this point, we also show how we can readily use one of such tools — a procedure for detection of influenza epidemics, to detect change of *trend* dynamics on Twitter.**

I. Introduction

Recently, micro-blogging and specifically Twitter, have become one of the fastest growing trends with an exponentially increasing user base. Because of its real-time nature and plurality of end user clients, Twitter has become an effective method for information dissemination — helping it to be formalized into a big media as both an influencer and a reflector of real-time news.

Twitter, as a micro-blogging platform, lets users (*twitterers*) publish statuses (*tweets*) limited by 140 characters. A user can follow other users. Being a follower on Twitter means that the user receives tweets from those the user follows. But, unlike most social networks, the following or followed relationship does not require reciprocation — the user being followed does not need to follow back. A well-defined markup culture has evolved for responding to tweets: RT stands for retweet, '@' followed by a user identifier address, and '#' followed by a word represents a hashtag.

Because of its unique popularity and real-time characteristics, a growing number of researchers have focused on Twitter lately. Java et al. [7] investigated the motivation of twitterers and the social network that ensues. Weng et al. [24] focused on *homophily* [15] to find influential twitterers, and Cha et al. [4] provided an empirical study on user influence on Twitter. To take advantage of its real-time nature and large user base, there are also on-going efforts to use twitterers as social sensors to detect events. Sakaki et al. [19] proposed an algorithm to detect earthquakes in Japan by monitoring tweets. Starbird and Palen [22] showed that during mass emergency, retweets are more likely than non-retweets about events. Lerman and Ghosh [11] conducted an empirical study of user activities on Digg and Twitter which results in finding some similarity of between these two sites in news spreading. Another important direction of study has been to discover how trends on Twitter can be used as reflective indicators of the real-world sentiment — ranging from opinion poll in elections to collective action campaign [5].

As we can see, there has been a considerable amount of research about social aspects and usability of Twitter. But, to the best of our knowledge there has not been any attempt to mathematically model news spreading on Twitter. Being able to do so is critical to understand information-sharing behavior and dynamics on Twitter, which is necessary for the effective uses of Twitter from both personal users' and corporate users' views.

This paper aims to build a mathematical model to explain how news actually spreads on Twitter. Specifically, in this study we investigate how a Twitter activity can be described by using well-known deterministic models for infectious diseases. This deterministic model is simple enough to be operationally useful and can also help us to identify the control factors of the persistence and stability of popularity of a particular news trend. There have been no research efforts on connecting well-known epidemic models and news spreading, and furthermore the potential results obtained from such a study will help us gain an understanding in information dissemination dynamics on Twitter.

II. Epidemic Models

Traditional epidemiology studies diseases with linear models that consider individuals as independent units of observations. This process is based on Newtonian physics — however complicated the disease mechanic may be, the relation of causes to effects is straightforward. After introduced by the French mathematician Henri Poincaré at the start of the 20^{th} century, a new paradigm of complexity has been introduced into epidemiology. More recently, chaos theory [17] has been developed that highlights the importance of nonlinear phenomena in infection disease processes.

Usually two types of models are used in the study of infectious diseases: stochastic and deterministic models. Stochastic models rely on among-individual chance variations in risks of exposure, diseases and other factors — allowing heterogeneity in population. But stochastic models are very hard to set up and need more data and many simulations to yield a useful prediction. Deterministic models, also known as compartmental models, attempt to describe and explain what happens on the average at the population scale.

1082-3409/11 $26.00 © 2011 IEEE
DOI 10.1109/ICTAI.2011.33

Most models of infectious disease processes used are deterministic because they require less data, and are relatively easy to set up. In this paper, we focus on deterministic models only as we concentrate on the behavior of a large scale population. For small populations, stochastic models should be used [3].

A. Deterministic Modeling

Deterministic models categorize individuals into different subgroups (compartments). The SEIR model, for example, includes four compartments represented by Susceptible, Exposed, Infectious and Recovered. The models also specify the transition rate between the compartments. For modeling a disease, it is necessary to have a realistic representation of the biology of the disease — duration of ineffective period, incubation period, immune status after infection and so on.

For example, the SEIR model considers the infected phase accounting for a latent period of disease — when infected individuals (exposed) go through a latent period before being infectious. On the other hand, the SIR model assumes that individuals are infectious as soon as infected — no latent period to be taken care of. Some models assume long lasting immunity after infection (SIR and SEIR) while other models posit recoverds become susceptible again (SIRS and SEIRS). To analyze deterministic models, they are usually represented by differential equations describing the transitions between the different disease compartments using continuous time steps. For example, the SIR model can be represented in Figure 1

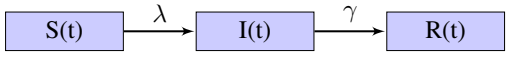

Fig. 1. SIR model.

where

- $S(t)$ = number of susceptible at time t.
- $I(t)$ = number of infectious at time t.
- $R(t)$ = number of recovered at time t.
- λ = the rate of infection per unit time.
- γ = the rate at which an infectious individual recovers per unit time.

Then, the differential equation system corresponding to the SIR model is:

$$\frac{dS}{dt} = -\lambda \cdot S(t).$$
$$\frac{dI}{dt} = \lambda S(t) - \gamma \cdot I(t).$$
$$\frac{dR}{dt} = \gamma \cdot I(t).$$

Here, $\frac{dS}{dt}$ means change in S per small unit time dt. To be more specific the equation

$$\frac{dS}{dt} = -\lambda \cdot S(t)$$

means that the compartment of susceptible depletes itself by $\lambda \cdot S(t)$ as susceptible become infectious by the time interval dt. Similarly, for the infectious compartments, new $\lambda \cdot S(t)$ individuals are being added and $r \cdot I(t)$ individuals become recovered by the time interval dt, and so on.

As the propagation of a disease depends only on the ability of infectious agents to transmit the disease to the susceptible, the number of the newly infected at each time step depends on the contacts between infectious and susceptible individuals. So, if we know the probability of an effective contact β, then the rate of infection can be effectively expressed as

$$\lambda = \beta \cdot I(t).$$

In that case the SIR model can be re-written as follows:

$$\frac{dS}{dt} = -\beta \cdot S(t) \cdot I(t).$$
$$\frac{dI}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t).$$
$$\frac{dR}{dt} = \gamma \cdot I(t).$$

Depending on disease biology and available data, we can build a more complex model to have a better understanding of how a particular epidemic sets up in a population. For example, if we want to allow entries of the new susceptible by birth and mortality in the course of time as shown in Figure 2, where

- μ = birth rate per unit time.
- θ = death rate per unit time.

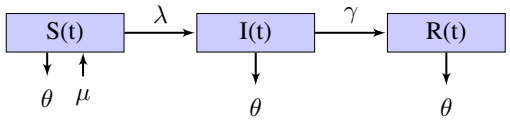

Fig. 2. A SIR model with birth and death rate.

Then the corresponding SIR models can be represented by the following differential equations:

$$\frac{dS}{dt} = -\beta \cdot S(t) \cdot I(t) + N(t) \cdot \mu - \theta \cdot S(t).$$
$$\frac{dI}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) - \theta \cdot I(t).$$
$$\frac{dR}{dt} = \gamma \cdot I(t) - \theta \cdot R(t).$$

where $N(t) = S(t) + I(t) + R(t)$, is the size of population at time t.

The potential of infection in a population depends on the *basic reproduction number* R_0 that is defined as the average number of persons directly infected by an infectious individual during his entire infectious period when he enters a totally susceptible population.

The threshold theorem established by Kermack and McKendric [9] says that if R_0 gets smaller than 1, the disease eventually disappears from the population because, on average, each infectious individual cannot ensure transmission of the disease to one susceptible resulting in lesser amplitude of the disease spreading phase comparing to preceding ones. If

R_0 equals to 1 then, the disease remains endemic as one infectious on the average spreads the disease to one susceptible individual. On the other hand, if R_0 is greater than 1, an epidemic ensues. This also explains why the introduction of infectious individuals into a community of the susceptible does not automatically give rise to an epidemic outbreak.

III. News Spreading on Twitter

In our approach, there is a basic similarity between the news dissemination on Twitter and the transmission of an infectious disease among the individuals. In other words, each news topic on Twitter spreads like 'a contagious disease', where

- The infectious are the twitterers who have participated in news spreading by tweeting about that topic.
- The susceptible are the set of twitterers who follow the infected twitterers as they receive those tweets (infectious contacts) on their stream and as a result they too can tweet about that topic (risk of being infected).
- As recency is an important issue in news spreading, to penalize older contents, we assume that infectious individuals lose their ability to spread news after a certain amount of time — becoming the recoverd in epidemiological terms.

To develop an epidemiological model for news spreading on Twitter, it is necessary to pick a model and its corresponding parameters that portray a complete and realistic picture. We choose the SIR model because of the following observations.

- To highlight the importance of recent tweets, we put more emphasis on newly infectious individuals. In other words, infectious individuals can not remain infectious for ever, which excludes SI-related models from consideration.
- When individuals tweet about a topic, it appears on the streams of the susceptible immediately. So, there is no latent period of the spreading, excluding SEIR-related models from consideration.
- Usually participating on a news spreading is a one-time shot per news cycle, making SIRS models unusable for this case.

We also allow the entry of a new susceptible similar to the birth rate in traditional epidemiology as tweets from infectious individuals reach to their followers' stream causing the population size to grow. But, unlike in traditional epidemiology, new susceptibles can be introduced only from infectious individuals. So, our proposed model can be represented by the following differential equations:

$$\frac{dS}{dt} = -\beta \cdot S(t) \cdot I(t) + I(t) \cdot \mu$$
$$\frac{dI}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t)$$
$$\frac{dR}{dt} = \gamma \cdot I(t)$$

with parameters comparing to traditional epidemiology shown on Table I.

A. Parameter Selection

As different diseases have different dynamics determined by the demographic and biological characteristics (transition rates), the next step after selecting a model is to collect data and find appropriate values of parameters that can explain dynamics of disease spreading.

One important assumption in this model is, all new infectious individuals can arise from the susceptible set. But, as a news item becomes a more mainstream media topic, this assumption may not reasonably hold in our case, as individuals can also get infected from the outside of the twitter population thus becoming an infectious without ever being in the susceptible set. To measure this effect, we focus on three kinds of trends.

- Events internal to Twitter — events that arise and die away within Twitter without any external interference. For this, we focus on 'Follow Friday' trend, where on every Friday users suggest other user(s) to follow. This has been a recurring event on Twitter after introduced by a user on January 16^{th}, 2009 [1].
- Real time news events, in other words, the traditional news. For this, we focus on the games in the world cup soccer between USA and Ghana on June 26^{th}, 2010.
- Social events — which are not news in the traditional sense, but, as each important social event usually becomes a trending topic on Twitter because of the sheer number of tweets related to it, we decided to track one such an event. For example, the Memorial Day in the USA.

1) Data set: For each of these events, we maintained a set of infectious, susceptible and recovered individuals over time by using Twitter API.

- We used stream API to track a particular keyword and in every time epoch Δt, we updated the set of infectious individuals $I(t)$ by retrieving the users who tweeted about that topic in the $[t - \Delta t, t]$ interval. Though, unlike traditional epidemiology — where the duration of epochs usually ranges from weeks to months, in our case the duration of an epoch is understandably much smaller — ranging from one to four hours.
- We then retrieve all the followers F of each infectious individual $i \in I(t)$, and after filtering out any followers who are also in $I(t)$, add them to the susceptible set $S(t)$
- We also remove infectious individuals from $I(t)$ who have not tweeted about that topic in the $[t - 2\Delta, t]$ interval to mimic a recovering process and add them to the recovered $R(t)$ set.

For the Memorial Day event, we collect $546,320$ tweets starting from 28^{th} of May over 7 days. Similarly, for 'Follow Friday' we collected $115,300$ tweets starting from 20^{th} of May to 22^{nd} of May. And, for the match between USA vs. Ghana, we collect $165,779$ tweets starting from 25^{th} of June for 3 days.

B. Simulation Results

Given the above data set, our objective is to determine appropriate values of parameters that reasonably explain the

	Epidemiology	**Information diffusion on Twitter**
S(t)	Set of susceptible individuals at time t.	Set of users who have received tweets from infectious individuals at time t.
I(t)	Set of infectious individuals at time t.	Set of individuals who tweeted about that topic at time t.
R(t)	Set of individuals who have recovered at time t.	Set of infectious individuals who have been inactive for a pre-defined period of time by not tweeting about that topic.
β	Force of infection: Infection rate.	Spreading rate.
μ	Birth rate.	Number of new followers who receive tweets from infectious individuals per unit time per infectious individual.
γ	Recovery rate.	1/average duration of infectiousness.

TABLE I
MODEL PARAMETERS IN EPIDEMIOLOGY VERSUS NEWS SPREADING.

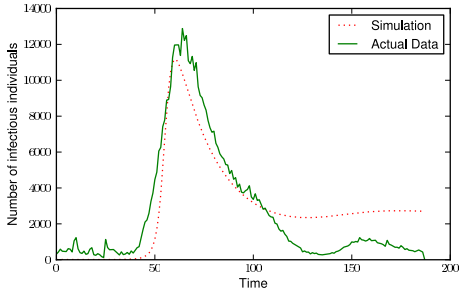

Fig. 3. Memorial Day Simulation with $\mu = 4.33 \cdot 10^{-05}$, $\beta = 5.38 \cdot 10^{-02}$, $\gamma = 1.02 \cdot 10^{-01}$.

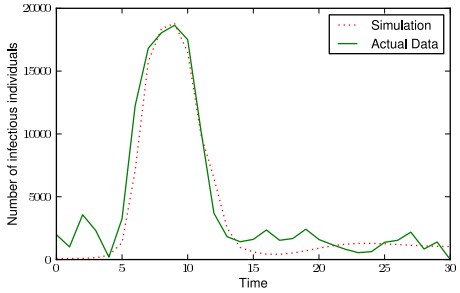

Fig. 4. World Cup USA Simulation with $\mu = 1.4 \cdot 10^{-04}$, $\beta = 3.43 \cdot 10^{-01}$, $\gamma = 6.635 \cdot 10^{-01}$.

spreading of news. To do so, we perform a multiparameter least-square fit by using the optimize module provided by SciPy [8].

From Figure 3, we can see that our model does fairly well except in the later region. That's because, Memorial Day — being a social event, does not follow our assumption that infectious individuals can only arrive from the susceptible set. In other words, the number of infectious individuals entered from outside of the population is quite high. But the point is, even though the assumption of mass action principle is arguable here, our model predicts the rise of the trend quite well.

In contrast to the Memorial Day, the event of world cup match being played between USA and Ghana is more of a traditional real-time news. As we can see in Figure 4, our

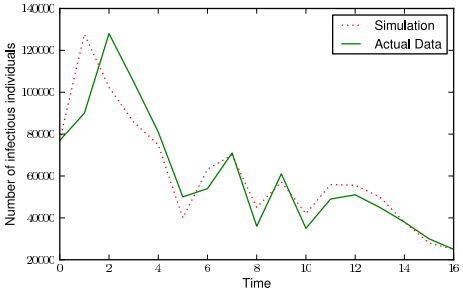

Fig. 5. Follow Friday Simulation with $\mu = 1.612$, $\beta = 5.859$ and $\gamma = 6.274$.

model performs much better than the Memorial day event. We believe that the better performance of our model here is due to the fact that the population (the susceptible, infectious and recoverd) in this case is more connected by either spatial or user-preference similarity. In other words, "reciprocity" is high between infectious and susceptible twitterers causing the presence of homophily and less sporadic outside interference.

In Figure 5, we can see that the number of incidences for the Follow Friday event is quite irregular contradicting the common trend of spreading dynamics for both the Memorial Day and the world-cup event as shown in Figure 3 and Figure 4, respectively. We assume that this happens because of the time difference on different time zones as we did not restrict our query to track tweets only from the USA. But, even in this highly irregular case, the performance of our model is quite good which is expected as this event arises and dies down within Twitter without having almost any external influence.

Though, we can not compare our model performance with any existing work, as to the best of our knowledge, this is the first approach in modeling twitter dynamics, but from the above figures, it seems quite plausible that a system of differential equations can be used to successfully emulate the trend spreading dynamics on Twitter for a range of events. This observation supports our hypothesis that information diffusion dynamics on Twitter can be explained by using an epidemic model. In other words, we can infer that the life cycle of trends on Twitter is similar to how epidemics set up in a population — initially, the number of infectious I and recoverds R are 0. As a trend begins to spread I, R

increases and the dynamics of S change depending on both the force of infection and birth-rate. After the number of infectious individuals who participated to spread news increases to reach a peak, it decreases and eventually the trend dies down, as infectious individuals become less effective in spreading the trend showing that recency is very important in trend popularity.

So, by drawing analogy between disease spreading and information diffusion on Twitter, we can explain and predict trend dynamics by plugging appropriate parameters into our model.

IV. Detecting change of trend dynamics

One important aspect of being able to express trend spreading on Twitter as an epidemiological model is, it enables us to use a wide number of tools and techniques to use for both predicting and explaining trend behavior. For example, surveillance of an infectious disease is a quite well-studied subject. Modern surveillance systems emphasize on detection of an epidemic as early as possible as it enables prompt intervention which is very important due to the threat of new infections, as well as modified strains of old infections.

In this section we show that change of trend dynamics on Twitter can be detected by extending one such surveillance system used for detection of influenza epidemics.

A. Influenza Surveillance System

Influenza is an acute respiratory illness caused by an influenza virus. Influenza epidemics occur almost every year with peak prevalence in winter. According to WHO [16], each year Influenza causes three to five million cases of severe illness and resulting in $250,000$ to $500,000$ deaths.

The traditional approach to influenza surveillance focuses on determining a baseline distribution of the susceptibles to death cases during a non-epidemic period from which an alert threshold can be established. Both Serfling's method [21] used by the Centers for Disease Control and Prevention *(CDC)* and historical limit methods used by European Influenza Surveillance Network *(EISN)* are examples of this approach. But in our case, to detect trend changing on Twitter, this methodology has a major drawback — the lack of predefined epidemic and non-epidemic periods to model the baseline distribution for each trend.

As the same issue of lacking historical data arises while detecting the outbreak of a new disease or old disease with changed behavior, a number of alternatives to the traditional baseline method have been suggested. LeStrat and Carrat [10] introduced a Bayesian perspective that observations are supposed to be independent given our belief about the epidemics, and they used a hidden Markov model to segment influenza data series into epidemic and non-epidemic series. This Bayesian perspective has been extended by Rath et al. [18], Madigan [13] and Sebastiani et al. [20]. For our purpose, we focus on a more recent approach by Martínez-Beneito et al. [14] which uses a Markov switching model to determine the epidemic and non-epidemic phases from surveillance data which consists of a series of differenced incidence rates.

B. Data

In this approach, we use a series of differenced incidence rates which enables us to use autoregressive modeling to analyze data. In particular, a first order differenced series, being stationary — as shown in Figure 6, allows us to restrict our study to its variability at each moment as it has a zero mean. In other words, non-epidemic dynamics are characterized by small random changes around zero, while in the epidemic phase, dynamic changes are greater and interrelated. So, depending on whether the system is in an epidemic or non-epidemic phase, it can be modeled as a first-order autoregressive process or with a Gaussian white noise process, respectively.

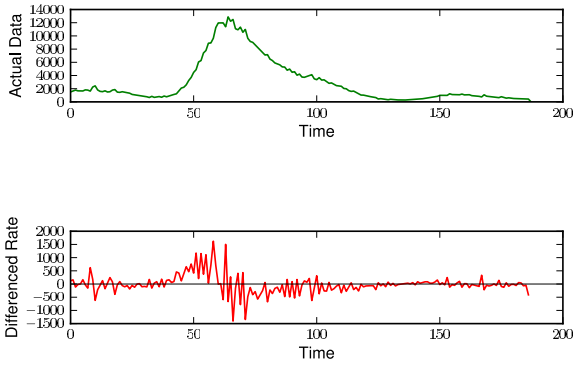

Fig. 6. Original incidence rate (on top) and first order differenced rate (bottom).

The same approach of using differenced series to distinguish between epidemic and non-epidemic phases has also been used by Barón [2].

C. Hidden Markov Model

The main idea here is to determine epidemic and non-epidemic phases from the series of differences using a two-stage Markov model. Let I_i denote the difference of rates between time epochs $i+1$ and i. E_i is an unobserved random variable which indicates the current phase of the system — 1 for being in epidemic; 0 for being in non-epidemic. And the transition probabilities of E_i are given by

$$P_{k,l} = P(E_{i+1} = l | E_i = k),$$

where $k, l \in \{0, 1\}$. Then we can model the conditional distribution of I_i either as an autoregressive process of order 1 or as a Gaussian white noise process depending on the current value of E_i as follows:

$$\begin{aligned} I_1|(E_1 = 0) &\sim N(0, \sigma_0^2). \\ I_1|(E_1 = 1) &\sim N(0, \sigma_1^2). \\ I_i|(E_i = 0) &\sim N(0, \sigma_0^2). \\ I_i|(E_i = 1) &\sim N(\rho \cdot I_{i-1}, \sigma_1^2). \end{aligned}$$

where variance σ_k^2 depends on the phase of the system.

The next step is to specify the prior distribution of parameters. Following the suggestion of Gelman [6], we choose uniform distributions for standard deviations of random effects. In addition to that, we also note that σ_1^2 is related to the epidemic phase where σ_0^2 denotes a random variation, so the later value should be lower which encourages the use of the following hierarchical steps to describe our prior knowledge:

$$\begin{aligned}
\theta_{low} &\sim \text{Unif}(a, b).\\
\theta_{mid1} &\sim \text{Unif}(\theta_{low}, b).\\
\theta_{mid2} &\sim \text{Unif}(\theta_{mid1}, b).\\
\theta_{high} &\sim \text{Unif}(\theta_{mid2}, b).\\
\sigma_0 &\sim \text{Unif}(\theta_{low}, \theta_{mid1}).\\
\sigma_1 &\sim \text{Unif}(\theta_{mid2}, \theta_{high}).
\end{aligned}$$

Here, a and b are hyper-parameters. By using this hierarchical structure also enables us to avoid the identifiability problem [23] between periods in the MCMC process. We also associate the following usual non-informative prior for $\rho, P_{0,0}, P_{1,1}$:

$$\begin{aligned}
\rho &\sim \text{Unif}(-1, 1).\\
P_{1,1} &\sim \text{Beta}(0.5, 0.5).\\
P_{0,0} &\sim \text{Beta}(0.5, 0.5).
\end{aligned}$$

D. Results

We employ Markov Chain Monte Carlo (MCMC) methods to infer analytical estimation. For that, we use OpenBUGS [12] to carry out the inference.

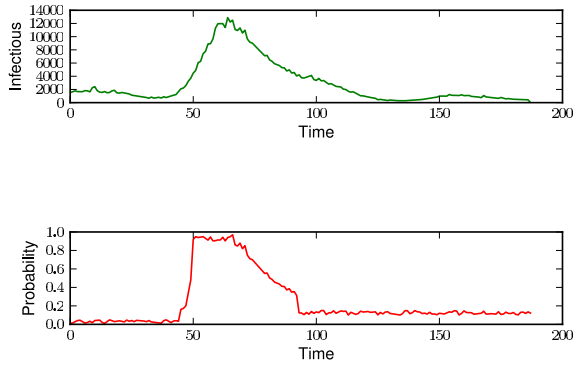

Fig. 7. Posterior probability of being in the epidemic state (at the bottom) corresponding to the actual incidence rate(at the top).

Figure 7 and Table II show the results of inference for the Memorial Day incidence rate data. The results have been obtained by using 4 independent chains of simulations with 20000 iterations, discarding the first 10000. Values for a and b have been set to be 20 and 1622 respectively, as the biggest difference of rates is 1622 and a is set to 20 to make sure that the posterior distribution of standard deviations does not converge to 0.

In Figure 7, we show the posterior probability of being in an epidemic phase — corresponding to the posterior mean of state variable E_i. These probabilities allow us to quantify the possibility of being in an epidemic state during each epoch — enabling for detection of trend change. For example, when for any topic the probability of being in an epidemic state exceeds 0.5, we can assume that the particular topic is trending.

Table II shows the posterior mean and the 95% posterior credible interval of the parameters. One important point is the strictly positive value of ρ, which justifies the autoregressive process used for the epidemic phase. From the high values of $P_{0,0}$ and $P_{1,1}$, it is apparent that the model is more likely to be in the same phase as it was during the previous epoch. Another point to note here is the distribution of $\theta_{low}, \theta_{mid1}, \theta_{mid2}$ and θ_{high} lies in seemingly different intervals, which validates the use of the current disjoint hierarchical structure as the parent of variances for different phases.

Parameter	Posterior Mean	95% Credible Interval
θ_{low}	266.0	$[26.77, 574.2]$
θ_{mid1}	346.1	$[43.65, 1184.0]$
θ_{mid2}	476.2	$[65.67, 1299.0]$
θ_{high}	557.5	$[127.8, 1384.0]$
ρ	0.9934	$[0.9849, 0.9994]$
$P_{0,0}$	0.9731	$[0.9407, 0.9941]$
$P_{0,1}$	0.02688	$[0.0059, 0.0592]$
$P_{1,0}$	0.04763	$[0.0105, 0.1052]$
$P_{1,1}$	0.9524	$[0.8948, 0.9894]$

TABLE II
POSTERIOR MEAN AND THE 95% CREDIBLE INTERVAL OF THE PARAMETERS.

From the above results, it is apparent that our model can be used as a notification system which raises an alarm when there is a change of Twitter dynamics. As shown in previous works [19, 22, 5], such a notification system can be used in a number of situations – ranging from mass emergency responses to public sentiment measurement. Though, an important novelty with respect to existing works is that unlike existing works, our model is not limited by the assumption that only a single instance of the target event may exist at any time — for example the method proposed by Sakaki et al. [19] can not differentiate between two or more earthquakes happening almost simultaneously. Our method can differentiate between recurring events happening in quick succession because of a direct consequence of the Markovian behavior defined in our model which enables it to adapt to quick phase-switching of trends by allowing any number of changes in subsequent epochs.

V. Conclusions

This paper focused on building a mathematical model of news spreading on Twitter. To do so, we have shown that well-known deterministic compartmental epidemic models can be extended to explain dynamics of trend spreading for various

types of trends including real-time news as well as social events. As epidemiology has been extensively studied, it is quite useful to be able to express a process as an epidemic model, which opens up an array of analytically rich tools that are known to work in real life situations. We also pointed out this advantage by showing that one such tool can be readily extended for detecting change in trend dynamics on Twitter.

For future work, we plan to focus on two possible extensions. One is to build a real-time online system for early detection of attention gathering trends from streams of tweets. Another potential extension to our work would be modeling the information spreading on Twitter as a stochastic epidemic model. As stochastic models can consider the among-individual variation in risks of exposure — enabling us to introduce heterogeneity in our modeling. These can lead to a better controlled and more realistic system as we can then consider the influence of users in spreading news.

ACKNOWLEDGMENT

This research has been supported by US National Science Foundation (NSF) under grant CCF-0905337.

REFERENCES

[1] Micah Baldwin. #followfriday: The anatomy of a twitter trend. Found at http://mashable.com/2009/03/06/twitter-followfriday/.

[2] M. Baron. Bayes and asymptotically pointwise optimal stopping rules for the detection of influenza epidemics. *Case Studies in Bayesian Statistics*, 6:153–163, 2002.

[3] F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*. Springer Verlag, 2001.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA, May 2010.

[5] M. Cheong and V. Lee. Twittering for Earth: A Study on the Impact of Microblogging Activism on Earth Hour 2009 in Australia. *Intelligent Information and Database Systems*, pages 114–123, 2010.

[6] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian analysis*, 1(3):515–533, 2006.

[7] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[8] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python. http://www.scipy.org/, 2001–.

[9] W.O. Kermack and Mckendrick. A G. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London*, volume 115 of *Series A*, pages 700–72. Royal Society of London, 1927.

[10] Y. Le Strat and F. Carrat. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, 18(24):3463–3478, 1999.

[11] K. Lerman and R. Ghosh. Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proceedings of the 4th International Conference on Weblogs and Social Media*. The AAAI press, 2010.

[12] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

[13] D. Madigan. Bayesian data mining for health surveillance. In A.B. Lawson and K. Kleinman, editors, *Spatial and syndromic surveillance for public health*, pages 203–221. John Wiley & Sons Inc, 2005.

[14] Miguel a Martínez-Beneito, David Conesa, Antonio López-Quílez, and Aurora López-Maside. Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in medicine*, 27(22):4455–68, September 2008.

[15] M. McPherson, L. Smith-Loin, and J.M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

[16] World Health Organization. Influenza fact sheet. Available at http://www.who.int/mediacentre/factsheets/fs211/en/.

[17] R. Pool. Is it chaos, or is it just noise? *Science(Washington)*, 243(4887):25–25, 1989.

[18] T. Rath, M. Carreras, and P. Sebastiani. Automated Detection of Influenza Epidemics with Hidden Markov Models. *Advances in Intelligent Data Analysis V*, pages 521–532, 2003.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[20] P. Sebastiani, KD Mandl, P. Szolovits, IS Kohane, and MF Ramoni. A Bayesian dynamic model for influenza surveillance. *Statistics in medicine*, 25(11):1803–1816, 2006.

[21] R.E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78(6):494, 1963.

[22] C. Starbird and L. Palen. Pass it On? Retweeting in Mass Emergencies. In *Proceedings of Conference on Information Systems on Crisis Response and Management (ISCRAM 2010)*, 2010.

[23] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[24] J. Weng, E.P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.