Social Media Epidemiology (I) – Microblog Emerging Outbreak Monitoring

Victor W. Chu Raymond K. Wong

University of New South Wales, Australia
{wchu,wong}@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-201328
October 2013

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

A recent study on collective attention in Twitter shows that epidemic spreading of hashtags only plays a minor role in hashtag popularity and is predominantly driven by exogenous factors. Although a standard epidemic model is insufficient to explain the diffusion patterns of hashtags, we show that a time-series form of susceptible-infectious-recovered (SIR) model can be extended to monitor emerging outbreaks in microblog. In particular, we focus on disturbance analysis in Twitter. Different from other research work on hashtag analysis, we introduce a notion of disturbance; which is defined by a probability distribution over a common vocabulary. We investigate the disturbances which have already been identified by a community, e.g., topics learned from hashtagged messages, to focus on interpretable results. The use of probabilistic definition of disturbances overcomes small usable sample space problem in hashtag analysis, such that related tweets can be included by inference. This report presents a Bayesian online parameter mining method to monitor the diffusion of emerging disturbances in Twitter by combining a semi-supervised topic learning model with an enhanced SIR time-series model, which covers both endogenous and exogenous factors. By monitoring the estimated effective-reproduction-number ($\hat{R}$) of disturbances, one can profile and categorize the disturbances based on their levels of contagiousness and so as to generate alerts on potential outbreaks.

1 Introduction

Although epidemiology is now commonly referred to the study of causes and effects of health and disease conditions and their patterns in defined populations, it can be literally interpreted as "the study of what is upon the people" according to the meaning of its root words[1]. Social networks are booming, which provide various platforms for people to connect with each others at an unprecedented scale. Microblog is one of the medium for short message communications. In this report, our investigation of microblog epidemiology is to formulate an online monitoring mechanism to identify emerging outbreak of "disturbances" by using Bayesian parameter estimation approach [10, 33].

Kermack and McKendrick's deterministic models [23, 24, 25] on epidemic are commonly discussed in literature, e.g., [17] and [20], to form an introduction to mathematical epidemiology. A simple form of their models is namely susceptible-infectious-recovered (SIR) [11], in which a population is divided into three mutually exclusive compartments: susceptible, infectious, and recovered. SIR explains the propagation of a disease within a population by a set of differential equations assuming that the population is large and closely connected. Kermack and McKendrick obtain an epidemic threshold result that a parameter R_0, commonly known as basic-reproduction-number [23, 24, 25], must exceed parity in order for an epidemic outbreak to occur.

Lehmann *et al.* [26] recent work on collective attention in Twitter social network (Twitter) shows that epidemic spreading of hashtags in Twitter plays a minor role in hashtag popularity and suggests that it is mostly driven by exogenous factors. Their study focuses on tweets (messages in Twitter) with specific hashtags. In this report, we approach the problem from a different perspective by considering disturbances and to identify whether the disturbances exhibit any emerging epidemic-like propagation behavior, where disturbances are topics identified from datasets with the assistance of hashtags.

By identifying disturbances in discussion within Twitter, we can include tweets without hashtags provided that their contents match with the respective profile of the topics. As the portion of tweets with hashtags can be as low as 3.3% out of the total number of tweets in a sample (as used by [26]), the impact of this approach can be significant as the sample size will be much larger as shown in our experiments. On the other hand, Lehmann *et al.* identify four groups of temporal patterns of collective attention from their dataset: 1) activity concentrated before and during the peak, 2) activity concentrated during and after the peak, 3) activity concentrated symmetrically around the peak, and 4) activity almost totally concentrated on the single day of the peak. However, only the third group is close to an epidemic spread pattern as shown in Figure 1.2(a). This, *prima facie*, indicates that a SIR model may not be best to model the propagation of collective attention.

Twitter data are messy and uncontrolled – the lengths of messages are short, language variation and misspellings are high, non-standard acronym uses are common; hence, they are difficult to model. However, the benefit of mining information out of Twitter data overweighs these problems as they are dynamic, quick-to-the-public and up-to-the-minute, e.g., speed can be a big advantage when tracking epidemics and emerging diseases in real life as stated in [9]. In

[1] `http://en.wikipedia.org/wiki/Epidemiology`

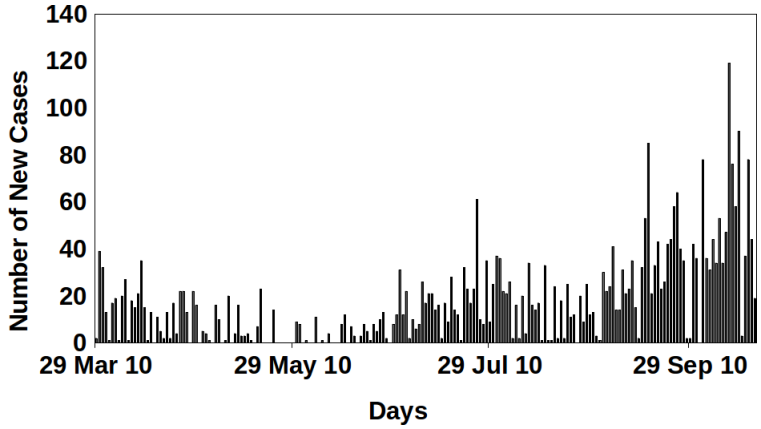


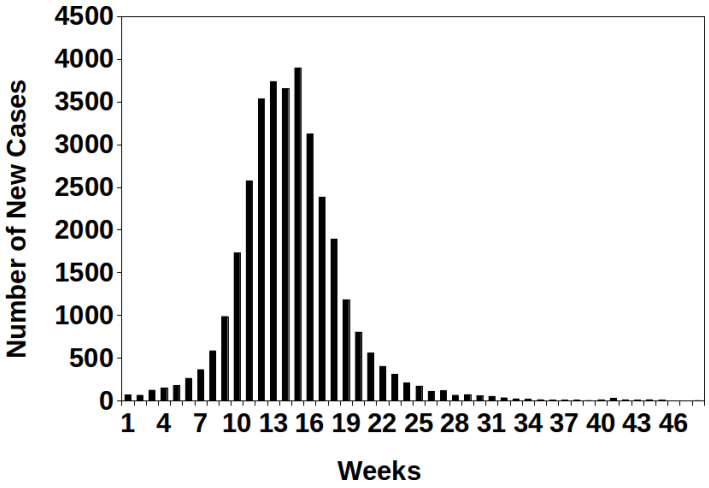
Figure 1.1: Propagation pattern of new cases of hashtag #ZodiacFacts from Twitter large corpus (described in Section 5), where the number of new cases represents the number of users first adopt the hashtag after he or she has posted a tweet containing it at a particular time

this report, we look into the epidemic potential of emerging disturbances by considering both endogenous and exogenous factors by studying the changes in the estimated effective-reproduction-number $\hat{R}$ over time, which resemble the study of the evolution of emerging infectious diseases by zoonotic virus [4].

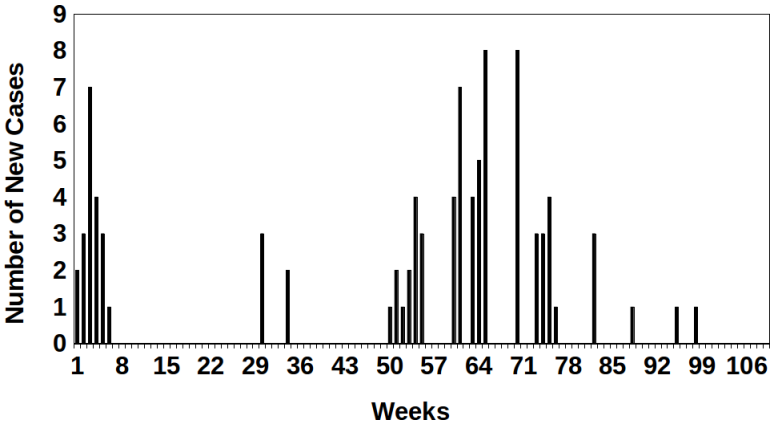
The data are expected to be stochastic, sometimes sparse, and their temporal patterns are anticipated to be close to emerging infectious diseases. The new cases propagation patterns of a hashtag, a standard epidemic and an emerging infectious disease are shown in Figures 1.1, 1.2(a) and 1.2(b) as examples respectively. As indicated by Lehmann *et al.* [26] that hashtag popularity is mostly driven by exogenous factors, in this report, we further develop a SIR time-series model considering both endogenous and exogenous factors to better explain emerging disturbances. In addition, we are aware that the adoption of hashtags has been improved, e.g., from 3.3% of sample posted in 2008-2009 [26] to 11% of sample posted in 2010 (Twitter large corpus, described in Section 5). This also justifies our work to revisit hashtags analysis, although from a different approach of emerging disturbances.

To summarise, our primary contributions are:

- We propose to identify disturbances by using probability distributions over a common vocabulary to overcome the problem of low hashtag adoption rate.
- By analysing disturbance data, we compare the diffusion patterns of disturbances with seasonal diseases and emerging epidemic; and find that disturbance time-series exhibit patterns closer to the latter.
- We enhance an emerging epidemic model which considers both endogenous and exogenous factors.
- We propose an online outbreak monitoring framework by using Bayesian parameter estimation approach for dynamic systems. By monitoring the changes of $\hat{R}$, one can predict the change of dynamics in social groups.



(a) A *Seasonal* infectious disease – A(H3) influenza virus isolated by WHO/NREVSS Collaborating Laboratories 2012-2013 season week 35 report (ending August 31, 2013) [8]



(b) An emerging infectious disease – laboratory confirmed new human H5N1 avian influenza cases from WHO reports in Vietnam (from January 2004 to June 2006) [4]

Figure 1.2: Examples of new cases propagation patterns of infectious disease

The reminder of this report is organized as follows: Section 2 presents related work. Section 3 presents our proposed emerging outbreak mining model for disturbances, including discussion of disturbance discovery and SIR model. Section 4 discusses the online disturbance mining and monitoring framework based on Bayesian parameter inference. Section 5 presents our experiment results and discuss our findings. Finally, Section 6 summarizes this report and briefly outlines our future work.

2 Related Work

Lehmann *et al.* [26] recent work on collective attention in Twitter shows that epidemic spreading of hashtags in Twitter plays a minor role in hashtag popularity and suggest that it is mostly driven by exogenous factors. In this report we further investigate the problem by using disturbances, defined as probability distributions over a common vocabulary. Loosely speaking, our definition of a social outbreak is the occurrence of cases of a disturbance, e.g., latest breaking news, new discoveries, new inventions, etc., in excess of what would normally be expected in a defined community.

Epidemiology has long been contributing to policy decisions making [12], evidence-based medicine [38], risk factors identification [3], and early intervention of susceptible populations [7]; whereas mathematical epidemiology uses mathematical models to investigate the dynamics of infectious diseases [17]. Similar applications can be done on microblog epidemiology, though in a neutral position treating control and spread indifferent. In epidemic studies, one of the early examples of mathematical epidemiology was done by Daniel Bernoulli in 1760 to evaluate the effectiveness of variolation (deliberate infection) of healthy people with smallpox virus through a puncture on the skin [17]. In recent time, infectious diseases modeling has been playing a prominent role in emerging infectious diseases prevention and control, e.g., Severe acute respiratory syndrome (SARS) [28] and influenza H1N1 (swine flu) [14], and to study the development and spread of drug resistant bacteria, e.g., Methicillin-resistant Staphylococcus aureus(MRSA) [32].

Susceptible-infectious-recovered (SIR) model [11, 23, 24, 25] divides a closed population into three mutually exclusive compartments, whereas the dynamics of this closed system is modeled by a set of differential equations. Basic-reproduction-number R_0 [23, 24, 25] must exceed unity in order for the occurrence of an epidemic outbreak. Despite that the assumptions used in the model (to be explained in Section 3.2) are very restrictive, SIR model has been shown to fit well with epidemics data, e.g., the main wave of the 1918-1919 pandemic in England and Wales [20, 31], and the recent A(H3) influenza seasonal time-series as shown in Figure 1.2(a). In particular, it best fits the cases where the modeling time frame is short as well as the size of the population is large and dense. R_0 is used as a key metric for the comparison of the propagation of diseases in a community [17]. It is important to note that R_0 is not only disease specific, it also depends on the behavior within a social group being measured, i.e., R_0 can be quite different between two social groups with different behaviors while having the same disease [22].

In addition to SIR model, we also explore Latent Dirichlet allocation (LDA) [6] for the definition of disturbances. Although LDA is commonly used topic

model, Zhao *et al.* [45] suggest that it may have difficulty in handling short messages, likes tweets which are only 140 characters in length maximally. There are several proposals in recent literature trying to alleviate the difficulty from different directions, such as 1) partially supervised learning model, e.g., Labeled LDA (L-LDA) [35, 36] and Partial LDA [37]; 2) single topic model, e.g., Twitter-LDA [45]; and 3) pooling schemes such as [19, 30, 42, 44]. Another problem of LDA is that sometimes the topics identified may be hard to be interpreted and difficult to associate the distribution of vocabulary with real-world objects [36]. On the other hand, a scalable solution is always desirable; in particular, to address big data problems. Hoffman *et al.* [18] recent proposal on an online approach to variational inference for LDA is one of the examples trying to address the computational difficulty in meeting real-world application requirements.

Besides, Mehrotra *et al.* [30] propose a hashtag labeling algorithm. The algorithm firstly pools all tweets by existing hashtags; then, if the similarity score — cosine similarity using T-F and TF-IDF vector space representations — between an unlabeled and a labeled tweet exceeds a certain confidence threshold, assign the hashtag of the labeled tweet to the unlabeled tweet.

3 Emerging Outbreaks Mining

Traditional hashtag analysis suffers from low usable sample space problem. This shortcoming also complicated by a relative short length of Twitter messages making the analysis very difficult. To address this problem, we propose to firstly identify disturbances by using semi-supervised topic learning method with hashtags as labels. Based on the signatures of disturbances, we can retrieve and group all the related tweets no matter whether they carry hashtag or not. Secondly, we modify a SIR time-series model to make it suitable to plug into an online mining framework, which we will discuss in Section 4. Before that, we firstly present the disturbance mining method, SIR model, and then our proposed emerging outbreak mining model for disturbances in the following subsections.

3.1 Disturbances Discovery

In the context of social media, we define $\mathcal{D}$ to represent a set of disturbances $\mathfrak{d}_j$, or topics as a term commonly used in topic discovery and topic model [6, 29, 34, 45]. For example, news events such as "the Haiti earthquake", entities such as "Michael Jackson" and subjects such as "global warming", etc. In the context of microblogging, #haiti, #michaeljackson and #globalwarming are popular hashtags representing the three items in social media in the past, respectively. Please note that we adopt $\mathfrak{d}_j \in \mathcal{D}$ as a notation to represent both diseases and disturbances in this report.

Latent Dirichlet allocation (LDA) relies on observed messages (or documents) to infer their hidden topic structure. As a probabilistic model, the generative process behind LDA assuming the words in each message are generated by a two-stage process: 1) randomly choose a distribution over topics, and 2) for each word in the message randomly choose a topic from the distribution in step 1 and then randomly choose a word from a distribution over a vocabulary. As a result, we obtain a collection of messages sharing the same set of topics,

but each message discusses those topics in different proportion [5]. In this report, we make use of hashtags to identify disturbances from messages; however, we would like to emphasize that the identifications and the signatures of disturbances are ultimately based on their individual distribution over a common vocabulary. We do expect more than one hashtags could coexist in a message, and also one or more disturbances could coexist in a message no matter whether it carries hashtags or not.

Despite of the similarity of what LDA can offer with our disturbance discovery problem, this is not the solution. Our focus is on known disturbances; and therefore, we can establish our model based on a partially supervised learning method by using hashtags as labels in messages. The learned topic model should be able to be applied to the whole collection of messages $\mathfrak{m}_i \in \mathcal{M}$ to identify similar disturbances embedded in messages no matter whether they have the related hashtags on them. Labeled LDA (L-LDA) is a perfect match to our requirements, which incorporates supervision by capturing preferred topics [35, 36].

L-LDA assumes that there is a set of labels Λ, and each of them are characterized by a multinomial distribution function $\boldsymbol{\psi}_j$ for $\lambda_j \in \Lambda$ over all words in a common vocabulary, where $j \in \{1 \dots |\Lambda|\}$. Each message $\mathfrak{m}_i$ uses only $\Lambda_i \subseteq \Lambda$ and $\mathfrak{m}_i$ could have a preference on some labels over others as represented by a multinomial distribution $\boldsymbol{\theta}_i$ over Λ_i. Each word w_k in $\mathfrak{m}_i$ is drawn from a word distribution $\boldsymbol{\psi}_z$ associated with the message's label $\lambda_z \in \Lambda_i$, where $k \in \{1 \dots N\}$. The word is drawn in proportion both to how much the message prefers the label $\theta_{i,z}$ and to how much that label prefers the word $\psi_{z,w}$. By substituting Λ by disturbances $\mathcal{D}$, the generative process is summarized in Algorithm 1.

Algorithm 1 Generative process in L-LDA

for each disturbance $\mathfrak{d}_j \in \mathcal{D}$ **do**
    Draw a multinomial distribution $\boldsymbol{\psi}_j$ from symmetric Dirichlet prior η
end for
for each message $\mathfrak{m}_i \in \mathcal{M}$ **do**
    Build a disturbance set $\mathcal{D}_i \subseteq \mathcal{D}$ describing the message from a deterministic prior Φ
    Select a multinomial distribution $\boldsymbol{\theta}_i$ over the disturbance $\mathcal{D}_i$ from symmetric Dirichlet prior α
    for each word position k in $\mathfrak{m}_i$ **do**
        Draw a label $\lambda_z \in \Lambda_i$ from label multinomial $\boldsymbol{\theta}_i$
        Draw a word w_k from word multinomial $\boldsymbol{\psi}_z$
    end for
end for

In this report, we adopt L-LDA to obtain disturbance signatures from our datasets so that we can identify all the related messages, e.g., the number of new cases time-series of disturbance *ZodiacFacts* (distribution defined by using tweets associate with hashtag #ZodiacFacts) is shown in Figure 3.1. In this example, our acceptance criteria is that the involvement of tweets in disturbance *ZodiacFacts* is larger or equal to 98%. Please note that #ZodiacFacts is first used on 29 March 2010 but disturbance *ZodiacFacts* exists since 1 January 2010 (the beginning of the dataset). The next step is to establish a time-series

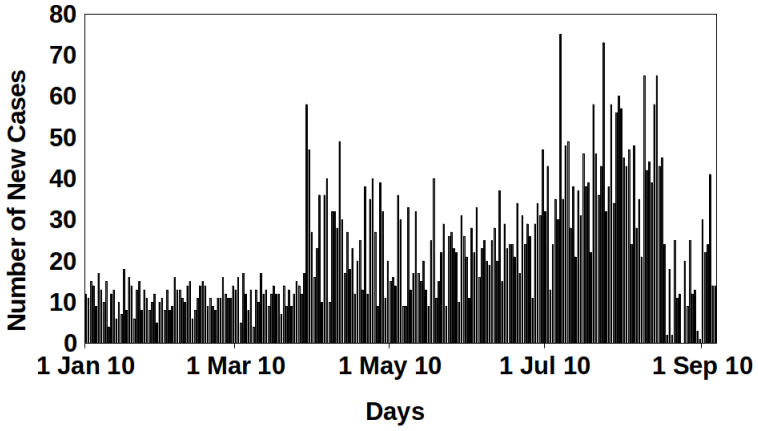


Figure 3.1: Propagation pattern of new cases of disturbance *ZodiacFacts* (only the first 8 months data are shown)

modeling framework and we begin with explaining SIR model.

3.2 SIR Model

SIR (susceptible-infectious-recovered) model [11, 23, 24, 25] divides a closed population set $\mathcal{N}$ into three mutually exclusive compartments according to their status: a) Susceptible set $\mathcal{S}$, b) Infectious set $\mathcal{I}$, and c) Recovered set $\mathcal{R}$, where $\mathcal{N}^{(t)} = \mathcal{S}^{(t)} \cup \mathcal{I}^{(t)} \cup \mathcal{R}^{(t)}$ in each time-step $t \in \{1 \dots T\}$. Hence, $|\mathcal{S}^{(t)}| + |\mathcal{I}^{(t)}| + |\mathcal{R}^{(t)}| = |\mathcal{N}^{(t)}|$, the size of the population. The following assumptions are made to describe the diffusion of a disease $\mathfrak{d}_j$ over $\mathcal{N}$:

- no demographics (births, deaths or migration) in $\mathcal{N}$ such that $|\mathcal{N}^{(t)}| = N$ for all t,
- all individuals in $\mathcal{N}$ are born at $t = 0$,
- all $\mathfrak{s} \in \mathcal{S}$ have no immunity to $\mathfrak{d}_j$, they move into $\mathcal{I}$ once being infected,
- every $\mathfrak{n} \in \mathcal{N}$ is in contact with all $\mathfrak{n}' \in \{\mathcal{N} \setminus \mathfrak{n}\}$ ("complete mixing assumption"),
- every $\mathfrak{i} \in \mathcal{I}$ spreads $\mathfrak{d}_j$ to $\mathfrak{s} \in \mathcal{S}$ with a constant rate β (transmission rate) and remains in $\mathcal{I}$ over an infectious period before moving into $\mathcal{R}$ with a constant rate γ (recovery rate), and
- all $\mathfrak{r} \in \mathcal{R}$ is immune from $\mathfrak{d}_d$ for life.

Let $S(t) = |\mathcal{S}^{(t)}|$, $I(t) = |\mathcal{I}^{(t)}|$ and $R(t) = |\mathcal{R}^{(t)}|$, $t \in \{1 \dots T\}$. Based on SIR model, the susceptible-infectious-recovered dynamics of the propagation of

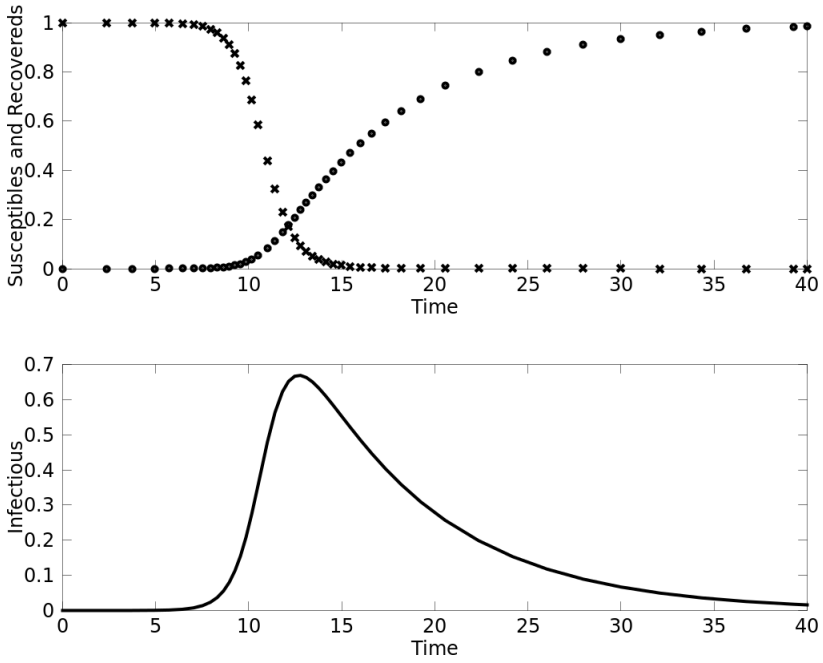


Figure 3.2: Epidemic outbreak dynamics ($\beta = 1.4$ and $\gamma = 0.14$) generated by SIR model ($\times$ – susceptibles $S(t)/N$; $\bullet$ – recovereds $R(t)/N$; line – infectious $I(t)/N$)

$\mathfrak{d}_j$ over $\mathcal{N}$ can be described by the following differential equations:

$$\frac{dS}{dt} = -\beta \frac{S}{N} I, \tag{3.1}$$

$$\frac{dI}{dt} = \beta \frac{S}{N} I - \gamma I, \tag{3.2}$$

$$\frac{dR}{dt} = \gamma \frac{I}{N}. \tag{3.3}$$

Kermack and MacKendrick [23, 24, 25] obtain an epidemic threshold result that basic-reproduction-number $R_0 = (\beta/\gamma)\, S(0)/N$ must exceed unity when $S(0)/N \approx 1$ in order for an epidemic outbreak to occur, e.g., the estimated R_0 of whooping cough in UK is 16-18 [2], smallpox is 3.5-6 [15] etc. In another words, an epidemic outbreak can only happen when $S(0)/N$ is larger than $1/R_0$. An example of epidemic outbreak dynamics with $\beta = 1.4$ and $\gamma = 0.14$ generated by SIR model is given in Figure 3.2.

R_0 has long been used as a key metric for the comparison of the propagation of diseases in a community, which is commonly defined as the average number of secondary infections that occur when one infectious is introduced into a completely susceptible host population [17]. Despite of the seemingly unrealistic assumptions made in the model, e.g., no demographics, complete mixing, and homogeneities (constant probabilities), SIR model has been shown to fit well with epidemics data [20, 31]. According to the assumptions made in SIR, it best fits the cases with short modeling time as well as large and dense population. The mathematical method to model the characteristics of spreading

of diseases can be enhanced to model an outbreak in digital social media, like Twitter, due to the following reasons:

- The population size of social media is large and it should be relatively static within the time frame of a social outbreak.
- Since most of the social outbreak topics are new, nearly everyone should be susceptible (no immunity).
- For a social media like Twitter, tweets are freely accessible; and therefore, a complete mixing assumption should not be totally unrealistic.
- For a large population, a homogeneous transmission rate and recovery rates are not a bad assumption for high level measurements.
- In today's fast-pacing digital society, only new topics can catch people attention. People tend to immune from older topics.

However, we find that the SIR model does not fit well with the Twitter hasttags data, e.g., #ZodiacFacts as shown in Figure 1.1. In addition, the investigation of Lehmann *et al.* [26] suggest that epidemic spreading of hashtags in Twitter plays a minor role in hashtag popularity and recommend that it is mostly driven by exogenous factors. It leads us to investigate whether an emerging outbreak style model should be used to model the propagation of emerging disturbances in Twitter.

3.3 Emerging Outbreak Model

An emerging disturbance outbreak may result from two sources alternating over time:

1. Endogenous – influence between peers such as follower/followee relationships and retweets, and
2. Exogenous – publicity campaigns out of Twitter or mass media communication.

A simple SIR model only considers endogenous propagation; and therefore, would not be able to model emerging outbreak completely. For emerging outbreaks, multiple introductions from exogenous sources may contribute to a number of observed cases. Similar concern has also been considered in epidemic outbreak modeling [4]. We enhance their case progression model approach and exclude death events based on the assumption of no demographics to establish the relationship of change in cases between discretized consecutive time-steps for endogenous propagation.

We define $\Sigma(t)$ as the total number of cases of a disturbance up to time t, such that

$$\frac{d\Sigma(t)}{dt} = \beta S/NI. \tag{3.4}$$

The number of new cases over a short period of time τ is equal to $\Sigma(t+\tau)-\Sigma(t) = \Delta\Sigma(t+\tau)$. Besides, by integrating Equation (3.2) between t and $t+\tau$, we obtain

$$I(t+\tau) = I(t)\exp\left[\gamma\int_t^{t+\tau}(R_0 S(t')/N - 1)\,dt'\right] \tag{3.5}$$

$$\approx I(t)\exp\left[\tau\gamma\left(R_t - 1\right)\right] \tag{3.6}$$

$$\approx I(t)b(R_t), \tag{3.7}$$

where $b(R_t) = \exp\left[\tau\gamma\left(R_t - 1\right)\right]$, by assuming that $S(t)/N$ remains constant over the time period of $[t, t+\tau]$ where $R_0 = \beta/\gamma$ and $R_t = (S(t)/N) \times R_0$. The approximation is valid provided that $I(t)/N$ is small ($\tau\gamma I(t)/NR_0 \ll 0$) [4]. By assuming that $S(t)/N$ is piecewise constant over $[t, t+\tau]$ but varies between intervals, we discretize the differential equation for the change in total number of cases between t and $t+\tau$ as

$$\frac{\Sigma(t+\tau) - \Sigma(t)}{\tau} = \beta\frac{S(t+\tau)}{N}I(t+\tau) \tag{3.8}$$

$$\approx \beta\frac{S(t)}{N}I(t)b(R_t) \tag{3.9}$$

by using Equation (3.7). Similarly, the change in total number of cases between $t-\tau$ to t is equal to

$$\frac{\Sigma(t) - \Sigma(t-\tau)}{\tau} = \beta\frac{S(t)}{N}I(t). \tag{3.10}$$

As a result, we obtain

$$\Delta\Sigma(t+\tau) = b(R_t)\Delta\Sigma(t) \tag{3.11}$$

by substituting Equation (3.10) into Equation (3.9).

According to Equation (3.11), a close to linear relationship is expected between $\Delta\Sigma(t+\tau)$ and $\Delta\Sigma(t)$ with a slope of $b(R_t)$. Evidence of this near linear relationship can be obtained diagrammatically by plotting a τ time-step shifted time-series against the original time-series, namely a lag-τ plot. A lag-1 plot ($\tau = 1$) of the new cases time-series of disturbance *ZodiacFacts* is generated as shown in Figure 3.3. Lag-1 plots ($\tau = 1$) of the time-series of A(H3) (Figure 3.4(a)) and H5N1 (3.4(b)) are produced and they confirm our expectation – A(H3) exhibits a close to linear relationship as R_t changes smoothly over the time-series. However, H5N1 does not; it may due to that the model does not cater for exogenous effects. We can observe that lag-1 plot of the time-series of hashtag #ZodiacFacts exhibits a pattern closer to what appears to be emerging disease H5N1.

We then cater for the impact of exogenous impacts by extending Equation (3.2) with an additional term $dB(t)/dt$, representing the external influence to the same disturbance caused by exogenous factors during a time period of dt:

$$\frac{dI}{dt} = \left(\beta\frac{S(t)}{N}I(t) - \gamma I(t)\right) + \frac{dB(t)}{dt}, and \tag{3.12}$$

$$\frac{d\Sigma}{dt} = \beta\frac{S(t)}{N}I(t) + \frac{dB(t)}{dt}. \tag{3.13}$$

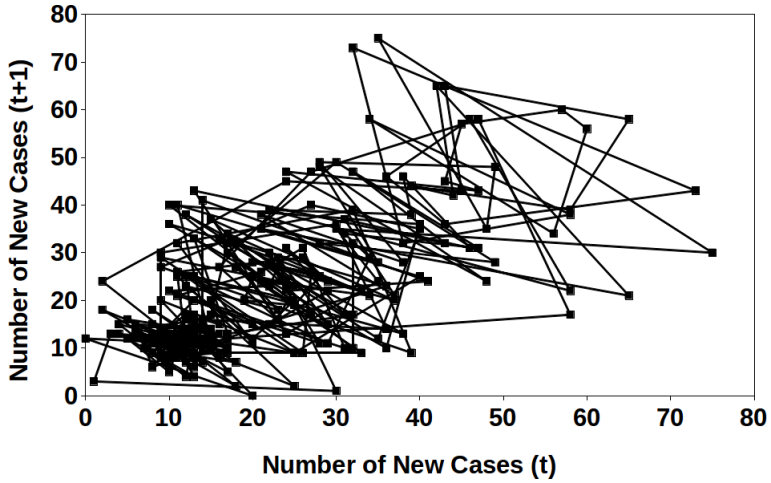


Figure 3.3: Lag-1 plot ($\Delta\Sigma(t+1)$ vs $\Delta\Sigma(t)$) of disturbance *ZodiacFacts* time-series as shown in Figure 3.1, where the lines between data points represent consecutive time connections

Please note that $dB(t)/dt$ is the net external effect which could be driven by a complex process from multiple exogenous forces. We take the assumption that $dB(t)/dt$ is constant in this initial investigation.

By integrating Equation (3.12) between t and $t+\tau$ (similar to how we obtain Equations (3.5), (3.6), and (3.7)), we obtain

$$I(t+\tau) = b(R_t)\left[I(t) + \int_t^{t+\tau} \exp\left(\int_t^{t1} -\gamma\left(R_0\frac{S(t_2)}{N} - 1\right)dt_2\right)\frac{dB(t_1)}{dt_1}dt_1\right] \tag{3.14}$$

$$\equiv b(R_t)\left[I(t) + \Psi(t,\tau,B)\right] \tag{3.15}$$

where $\Psi(t,\tau,B)$ represents the integral term in Equation (3.14). We come up with

$$\Delta\Sigma(t+\tau) \;=\; \Delta B(t+\tau) + b(R_t)\left[\Delta\Sigma(t) - \Delta B(t) + \tau\gamma R_t\Psi(t,\tau,B)\right] \tag{3.16}$$

for the case covering both endogenous and exogenous factors similar to how we obtain Equation (3.11) for the case where only endogenous influence is considered.

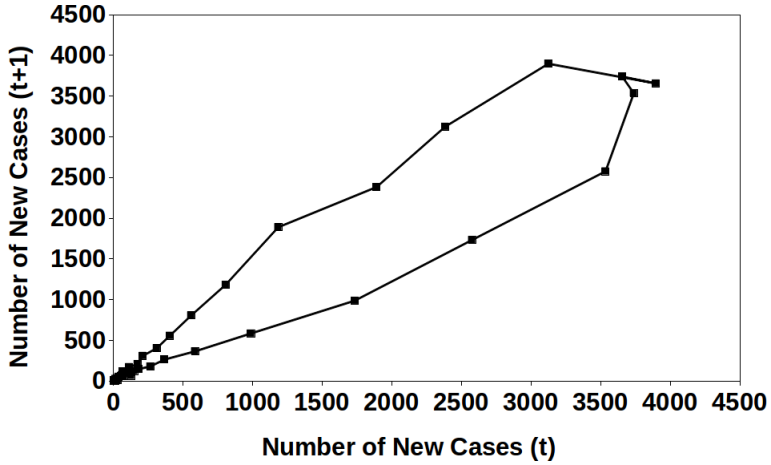
Finally, let $\Delta\Sigma'(t) = \Delta\Sigma(t) - \Delta B(t)$, we can rewrite Equation (3.16) to

$$\Delta\Sigma'(t+\tau) = b(R_t)\left[\Delta\Sigma'(t) + \tau\gamma R_t\Psi(t,\tau,B)\right]. \tag{3.17}$$

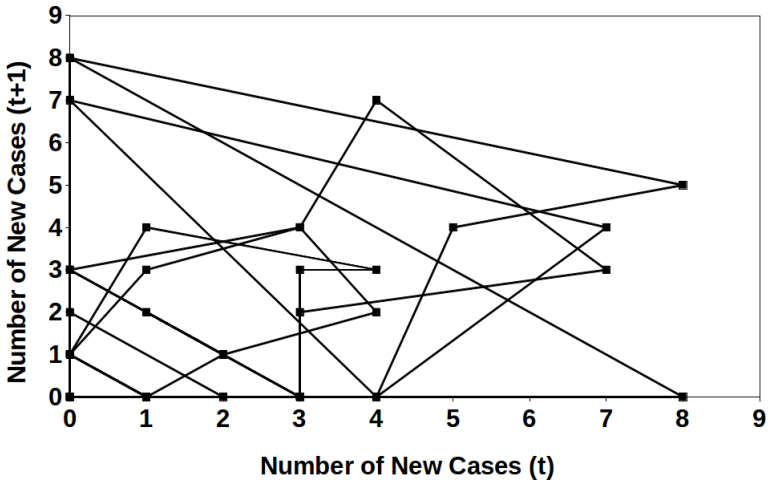
It is worthwhile to highlight the resemblance between Equations (3.11) and (3.17); however, a near linear lag-1 relationship does not hold in the latter due to the integral term even with the $\Delta B(t)$ term deducted from $\Delta\Sigma(t)$.

4 Online Emerging Disturbances Mining

In this section, we incorporate disturbance discovery and SIR time-series emerging outbreak model into an online mining framework.



(a) A(H3) time-series as shown in Figure 1.2(a) [8]



(b) H5N1 time-series as shown in Figure 1.2(b) [4]

Figure 3.4: Lag-1 plot ($\Delta\Sigma(t+1)$ vs $\Delta\Sigma(t)$), where the lines between data points represent consecutive time connections

Presume that $\Delta\Sigma'(t+\tau)$ is a discrete random variable generated from a probability distribution where only the average number of new cases λ is known, and is given by Equation (3.17). According to the principle of maximum entropy, if the information we know about a distribution is only the class it belongs to but nothing else, the distribution with the highest entropy should be chosen as its default distribution. It is because maximizing entropy not only minimizes the amount of prior information built into the distribution [21], many physical systems tend to move towards maximal entropy configurations over time [13].

For a distribution with only its mean is available (average number of new cases λ in our case), the maximum entropy distribution is Poisson [16]; hence, we assign $\Delta\Sigma(t+\tau) \sim Pois(\lambda)$. Assuming that the exogenous effect is more or less constant between t and $t+\tau$, we can calculate the integral to first order as $\Psi(t,\tau,B) = \tau dB/dt$ and approximate it by its discrete approximation $\Delta B(t)$ (the number of exogenous influence per unit time). As a result, Equation (3.17) can be rewritten to:

$$\Delta\Sigma'(t+\tau) = b(R_t)\left[\Delta\Sigma'(t) + \tau\gamma R_t \Delta B(t)\right] = \lambda_t. \tag{4.1}$$

We follow the dynamic model form of Bayesian melding approach [10, 33] for parameter estimation by applying Bayes' theorem to model each time-step:

$$p(R,\gamma|\lambda_{t+1} \leftarrow \lambda_t) = \frac{p(\lambda_{t+1} \leftarrow \lambda_t|R,\gamma)p(R,\gamma)}{p(\lambda_{t+1} \leftarrow \lambda_t)}. \tag{4.2}$$

The effective-reproduction-number R and γ can be estimated by successively applying Equation (4.2) with the posterior distribution for R and γ at time-step t as the prior at time-step $t+1$, where Markov Chain Monte Carlo (MCMC) is used to explore the parameter space with Gelman-Rubins' R as convergence diagnostic. R is expected to change over time because of the progressive reduction of $S(t)/N$ and the availability of more information from new observations.

To commence the estimation, we adopt an unbiased uniform distribution function for R and γ with their minimum and maximum values between 0 and 6, and 0 and 1, respectively, i.e., $p(R,\gamma) \sim \textsc{Uniform}((0,6),(0,1))$, as the initial prior. The selection of the range for R is based on the findings from our naïve approach to R_0 estimation in Section 5.2. The outstanding parameter for the successive Bayes' rule application is now remains with the selection of an appropriate ΔB so that we can derive $\Delta\Sigma'(t)$ from $\Delta\Sigma(t)$. In our experiments, deviance information criterion (DIC) [41] is used to guide our selection of ΔB.

Although it is possible to estimate the average frequency of exogenous inference per unit time by mining the periodic patterns from a dataset, we elect a simple approach assuming a constant rate of influence to focus on the subject in our initial investigation and leave it as a future extension. In our experiments in Section 5.3, we choose $\tau = 1$ day and calculate the number of new cases at time t by counting the number of users first use a hashtag to form a time-series.

To conduct the online emerging disturbances mining process, we firstly separate hashtags from messages in a dataset (steps 1 and 2 as shown in Algorithm 2). We then execute L-LDA modeling on the dataset to obtain disturbance signatures (probability distributions over a common vocabulary). The L-LDA model can then be applied to the whole dataset to conduct LDA inference to obtain a message-topic-distribution table, where each message is assigned with their involvement in the identified disturbances from the L-LDA model.

By electing an acceptance level of involvement in topics, e.g., at least 98%, we can extract all the messages related to each disturbance $\mathfrak{d}_j \in \mathcal{D}^{(1:t)}$. After calculating the number of new cases at time t by counting the number of users first involved in a disturbance $\mathfrak{d}_j$, we obtain a time-series $\Delta\Sigma(t)$, and a time-series $\Delta\Sigma'(t)$ of disturbance $\mathfrak{d}_j$ by selecting a suitable $\Delta B(t)$. Bayesian parameter estimation can then be applied to obtain $\hat{R}$ for each disturbance; whereas alerts can be issued for cases where $\hat{R} \geq 1$. The mining and alert process at each time-step t is detailed in Algorithm 2.

Algorithm 2 Online emerging disturbance mining

1: Extract hashtags $\mathcal{H}^{(t)}$ from $\mathcal{M}^{(t)}$
2: $Model^{(1:t)}$ = L-LDA-Model($Model^{(1:t-1)}, \mathcal{H}^{(t)}, F(\mathcal{M}^{(t)})$), where $F(\cdot)$ is a hashtag removal filter
3: $Inference^{(1:t)}$ = LDA-Inference($Model^{(1:t)}, F(\mathcal{M}^{(1:t)})$)
4: **for** each $\mathfrak{d}_j \in Disturbance(Inference^{(1:t)})$ **do**
5: Obtain time-series λ_t from $Msg(Inference^{(1:t)}, \mathfrak{d}_j)$ by calculating the number of new cases adjusted by elected $\Delta B(t)$
6: Calculate $\hat{R}$ and $\hat{\gamma}$ of $\mathfrak{d}_j$ by using Bayesian parameter estimation method
7: Generate alerts on disturbance $\mathfrak{d}_j$ if $\hat{R}_{\mathfrak{d}_j} \geq 1$
8: **end for**

5 Experiments and Discussion

The following experiments are conducted by using a HP EliteBook 8470p with Intel® Core^{TM}i5-3320M CPU 2.60 GHz × 4 and 8GB RAM with Ubuntu 12.04.1 LTS, Python 2.7.3, GNU Octave 3.2.4 and OpenJDK Java 7 Runtime.

A summary of our Twitter data sources is detailed in Table 5.1. Each tweet in the Twitter small corpus only includes a time-stamp at which it was posted in additional to its content and user name; however, each tweet of the Twitter large corpus includes textual content, user name, the time at which it was posted, whether or not it was in reply to another tweet, and additional metadata. The hashtags and attages are identified from the message content by the following regular expressions `#[a-zA-Z0-9_]+` and `@[a-zA-Z0-9_]+` respectively.

The characteristics of our Twitter large corpus is further explained by comparing with the dataset used in [26] where they are roughly one year apart. Our observations are:

- The dataset used in [26] comprises of 130 million tweets posted between November 20, 2008 and May 27, 2009 from about 6.1 million users. The number of tweets with hashtags is about 4.3 million, which is equivalent to 3.3% of the total number of tweets, and the average number of tweets per user is about 20.34.
- The messages in our Twitter large corpus were retrieved in 2010 (one year later). The number of tweets with hashtags had increased nearly fourfold to 11% out of the total number of tweets, and the average number of tweets per user was about 705.

	Twitter small corpus	Twitter large corpus
Source	Infolab, Texas A&M University	Courtesy of CSIRO
Number of tweets	305,310	$\approx$ 79 million
Number of users (Avg. tweets per user)	1,000 (305)	$\approx$ 112 thousand (705)
Tweets with hashtags (% out of total)	5,684 (1.86%)	$\approx$ 9 million (11%)
Time period	9 Mar 2007 to 31 Jan 2009	1 Jan 2010 to 20 Oct 2010

Table 5.1: Data sources

- Although the number of users covered in Twitter large corpus is not as large as the dataset used in [26], its average number of tweets per user is more than 34 times of the latter.

Because of the complete mixing assumption in SIR model, we expect a higher per user contribution to the dataset, i.e., more active users, should provide more accurate results in emerging disturbance mining.

5.1 Disturbance Discovery

In this experiment, we conduct disturbance discovery by using open source Stanford Topic Modeling Toolbox version 0.4.0[1]. We apply Labeled LDA (L-LDA) topic modeling on messages tagged with #ZodiacFacts (23,990 in total) from the Twitter large corpus through 2,000 iterations with the following filters:

- hastag filter - hastag are removed from the content of the messages and are saved in a separate filed as labels for L-LDA modeling,
- term minimum document count filter – terms appeared in less than 4 messages are removed,
- term dynamic stop list filter – 30 most common terms are removed, and
- document minimum length filter – messages with less than 3 terms are removed.

LDA inference is then made on the Twitter large corpus by applying the model. The messages with at least 99% involvement in the disturbance of #ZodiacFacts are identified, which is 164 in total. By excluding the messages already have the same hashtag #ZodiacFacts, 44 tweets are found in the month of July 2010 (in the middle of the time-series).

Out of the 44 tweets, the number of messages which carry the same disturbance is 21 (a precision of 48% within this subset); counted by manual inspection. They are marked with a '$*$' at the beginning in Table **??** where samples of the identified messages are displayed. Please note that some of the

[1] `http://nlp.stanford.edu/software/tmt/tmt-0.4/`

additional messages carry hashtags such as #ZodiacFact (missing a 's'), #hbu, #BestOfSigns etc., which we can reasonably confirm the usefulness of our disturbance mining approach. On the other hand, it is important to highlight that most of the remaining messages carry a sense of self-profiling based on our observations. It is arguable whether they should be included in the same disturbance.

5.2 A Naïve Approach to R_0 Estimation

In this experiment, we adopt a naïve approach to roughly estimate R_0 of a hashtag based on the SIR model described in Section 3.2. We estimate γ by the reciprocal of average "infectious" period (no. of days) and estimate β by the average number of "infection" per day. The "infectious" period of a hashtag is defined as the time period between the first message that a user posts with a hashtag and the last message that he/she posts with the same hashtag. R_0 is then calculated as β/γ by definition. The results of the selected hashtags from the Twitter small corpus are summarized as follows:

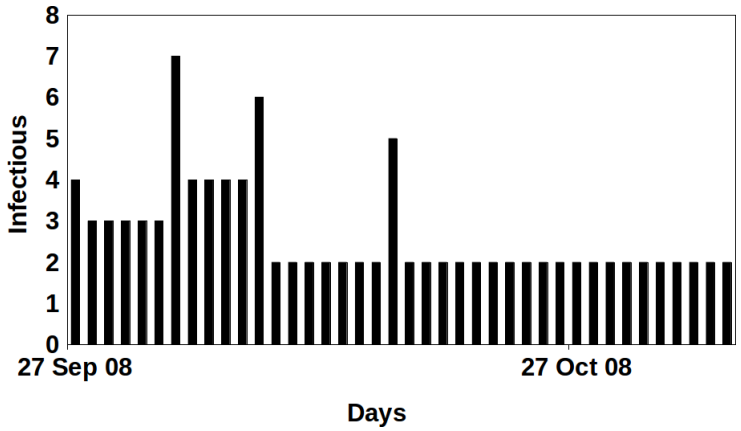
Hashtag	Estimated R_0	Infectious disease with similar estimated R_0
#debate08	1.38	FIV (cats) [40]
#current	2.68	Influenze (humans) [31]
#8217	3.16	SARS (humans) [43]
#votereport	4.0	Smallpox (humans) [15]
⋮		
#TwitVote	6.25	Rubella (humans) [1]

Based on the SIR model, the expected epidemic outbreak dynamic parameters of hashtags #current is $\beta = 0.3250$ and $\gamma = 0.1215$ and of hashtag #8217 is $\beta = 0.0237$ and $\gamma = 0.0075$. The actual infectious counts are summarized in Figure 5.1 and one would have difficulties to identify any clear pattern similarity between them and SIR model, e.g., Figure 3.2. This confirms that a basic SIR model may not be appropriate to model hashtag outbreak. However, it would be interesting to compare the roughly estimated R_0 of hashtags with the R_0 of known infectious disease to get an idea of their contagiousness.

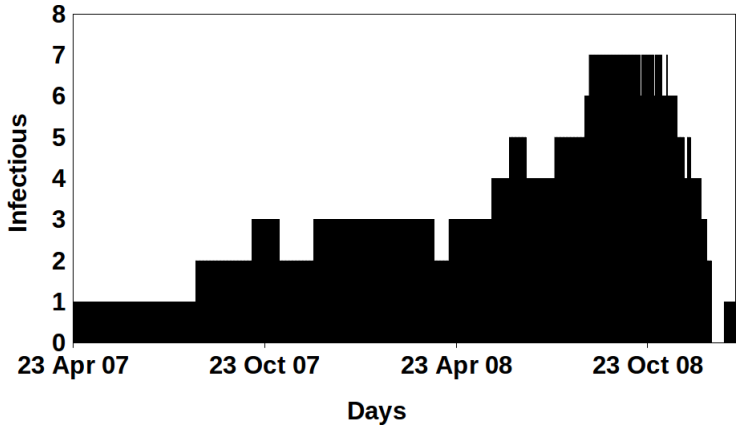
5.3 Emerging Outbreaks R Estimation

In this experiment, we verify online emerging disturbances mining framework outlined in Section 4 by using an open source Bayesian-inference Python package[2] to estimate R out of a time-series λ_t. We firstly execute the parameter estimation on new cases time-series derived from messages attached with hashtag #ZodiacFacts by using $\Delta B(t) = 0$. The results of estimation at the end of the time-series are shown in Figures 5.2 as baseline, where Figure 5.2(a) illustrates the fitted model and Figure 5.2(b) shows the $\hat{R}$ taken from each time-step. We obtain maximum-likelihood estimation (MLE) of $\hat{R} = 0.60 < 1$ and $\hat{\gamma} = 0.79$ at the end of the time-series.

[2] `http://code.google.com/p/bayesian-inference/`

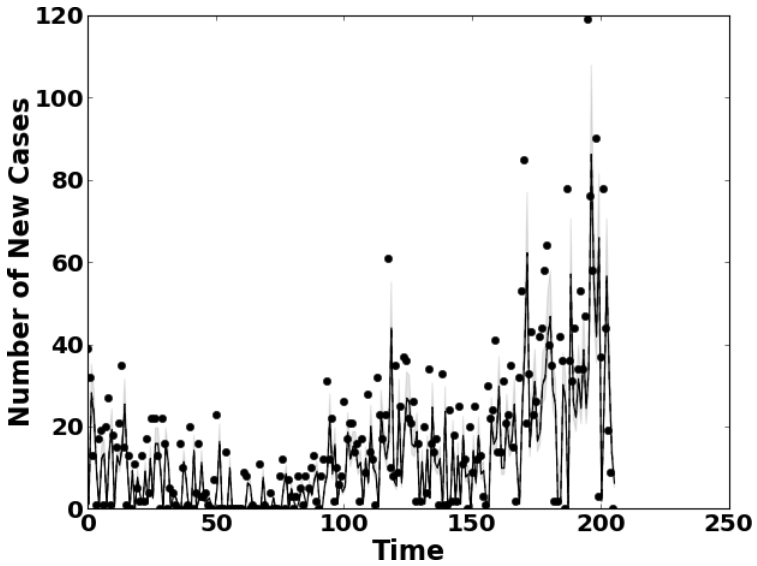


(a) Hashtag #current

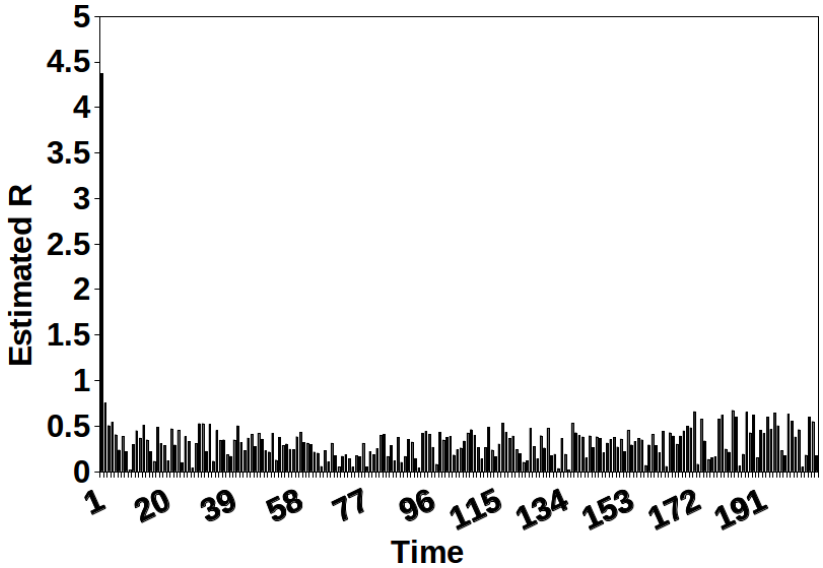


(b) Hashtag #8217

Figure 5.1: Hashtag infectious counts

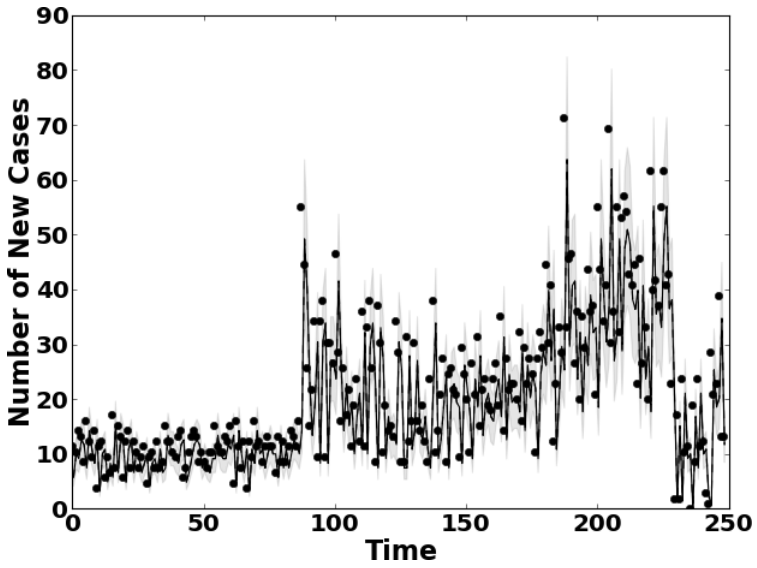


(a) Fitted model (• - observed value; line - fitted value)

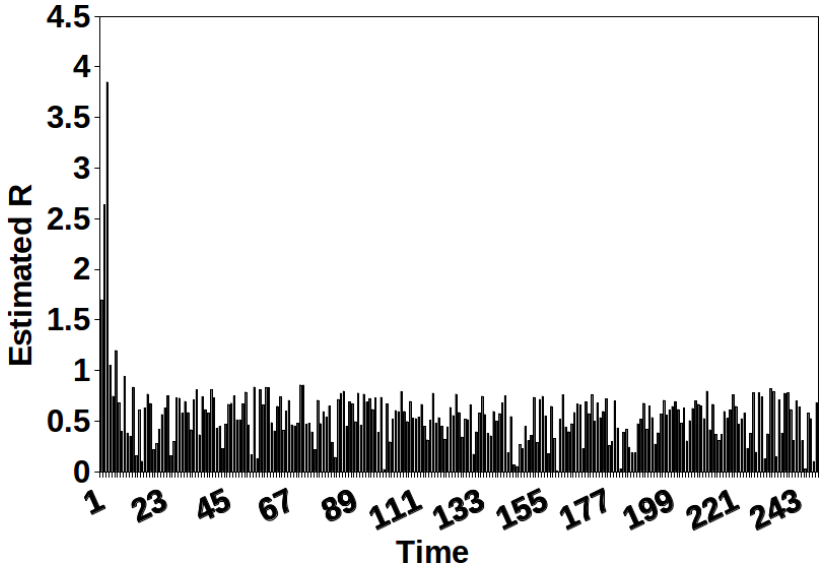


(b) Daily $\hat{R}$

Figure 5.2: Bayesian parameter estimation of the new cases of hashtag #ZodiacFacts with $\Delta B(t) = 0$



(a) Fitted model (• - observed value; line - fitted value)



(b) Daily $\hat{R}$

Figure 5.3: Bayesian parameter estimation of the new cases of disturbance *ZodiacFacts* with $\Delta B(t)$ equal to 10% of $\Delta\Sigma(t)$

We then increase the proportion of $\Delta B(t)$ out of $\Delta\Sigma(t)$ progressively with deviance information criterion (DIC) [41] collected after estimation at each increment. The result is that minimum DIC is consistently obtained from $\Delta B(t)/\Delta\Sigma(t)$ close to 0.05 (5%); and hence, we make a conclusion that the exogenous impact to hashtag #ZodiaFacts is about 5%. After rerunning the estimation with $\Delta B(t)/\Delta\Sigma(t) = 0.05$, we obtain MLE $\hat{R} = 0.16$ and $\hat{\gamma} = 0.39$, and $\hat{R}$ is under parity over the whole time-series. As a result, we can interpret that there is no epidemic occurs in this case and the exogenous effect is only about 5%.

Furthermore, the disturbance of *ZodiacFacts*, i.e., disturbance identified by the messages carry the hashtag #ZodiacFacts is identified by using disturbance discovery method described in 3.1. The messages with at least 98% involvement in the disturbance of ZodiacFacts are identified, where the total number of messages is now 43% more than just considering the message with the hashtag. Based on DIC, we find that $\Delta B(t)$ is equal to 10% of $\Delta\Sigma(t)$, which is 5% more than the case above if we consider hashtag only representing a higher exogenous involvement in driving the disturbance. The results are summarized in Figure 5.3. It is important to note that hashtag #ZodiacFacts is first used on 29 March 2010 but disturbance *ZodiacFacts* exists since 1 January 2010 (the beginning of the dataset). By comparing with Figures 5.2, there was a small epidemic happening at the beginning of the time-series and then the disturbance *ZodiacFacts* was sustaining by about 10% of exogenous force. If one was monitoring disturbances in Twitter, he or she should have predicted that *ZodiacFacts* was a topic gathering momentum at the time before hashtag #ZodiacFacts became popular.

6 Conclusions

This report proposes an online emerging outbreak monitoring framework. First, we propose a notion of disturbances in Twitter, defined as probability distributions over a common vocabulary, to include related tweets which are with or without hashtag to overcome the problem of low hashtag adoption rate. By analysing the disturbance data by a SIR model in time-series form, we compare the diffusion patterns of disturbances with seasonal diseases and emerging epidemic, and conclude that the disturbance time-series exhibit temporal patterns which are closer to the latter. Second, we enhance an emerging epidemic model which considers both endogenous and exogenous factors. Finally, based on the enhanced model, we propose an online emerging outbreak monitoring framework on Twitter by using Bayesian parameter estimation approach for dynamic systems. Experimental results indicate that our new approach is consistent with expectations and can form a baseline for our future work. We foresee that the presented online monitoring framework will be extremely valuable to public relations and marketing industries [39], political campaigns [27] etc., due to its intrinsic momentum prediction capabilities. Further work on $\Delta B(t)$ estimation is required. We envisage that we can obtain the average frequency of exogenous inference per unit time by mining the periodic patterns from the dataset. Besides, as some of the hashtags may be slightly different even though still referring to the same disturbance, it is essential that we can consolidate the corresponding disturbance signatures, so that we can evaluate its true impact to a society.

Bibliography

[1] R. Anderson and R. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford science publications. OUP Oxford, 1992.

[2] R. M. Anderson and R. M. May. Directly transmitted infectious diseases: control by vaccination. *Science*, 215(4536):1053–1060, 1982.

[3] M. Baker, A. McNicholas, N. Garrett, N. Jones, J. Stewart, V. Koberstein, and D. Lennon. Household crowding a major risk factor for epidemic meningococcal disease in auckland children. *The Pediatric infectious disease journal*, 19(10):983–990, 2000.

[4] L. M. Bettencourt and R. M. Ribeiro. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5):e2185, 2008.

[5] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[7] M. Boily, C. Lowndes, and M. Alary. The impact of HIV epidemic phases on the effectiveness of core group interventions: insights from mathematical models. *Sexually transmitted infections*, 78(suppl 1):i78–i90, 2002.

[8] Centers for Disease Control and Prevention, USA. Influenza viruses isolated by WHO/NREVSS collaborating laboratories 2012 - 2013 season. `http://www.cdc.gov/flu/weekly/weeklyarchives2012-2013/data/whoAllregt35.htm/`, 2013. [Online; accessed 10-September-2013].

[9] C. Chew and G. Eysenbach. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PloS one*, 5(11):e14118, 2010.

[10] F. C. Coelho, C. T. Codeço, and M. G. M. Gomes. A Bayesian framework for parameter estimation in dynamical models. *PloS one*, 6(5):e19616, 2011.

[11] K. Dietz. Epidemics and rumours: a survey. *Journal of the Royal Statistical Society. Series A (General)*, pages 505–528, 1967.

[12] J. N. Eisenberg, M. A. Brookhart, G. Rice, M. Brown, and J. M. Colford Jr. Disease transmission models for public health decision making: analysis of epidemic and endemic conditions caused by waterborne pathogens. *Environmental Health Perspectives*, 110(8):783, 2002.

[13] E. Fagiolini and C. Gruber. Entropy-based method for optimal temporal and spatial resolution of gravity field variations. In A. Abbasi and N. Giesen, editors, *EGU General Assembly Conference Abstracts*, volume 14 of *EGU General Assembly Conference Abstracts*, page 8916, Apr. 2012.

[14] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[15] R. Gani and S. Leach. Transmission potential of smallpox in contemporary populations. *Nature*, 414(6865):748–751, 2001.

[16] P. Harremoës. Binomial and poisson distributions as maximum entropy distributions. *Information Theory, IEEE Transactions on*, 47(5):2039–2041, 2001.

[17] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[18] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[19] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[20] T. House. Modelling epidemics on networks. *Contemporary Physics*, 53(3):213–225, 2012.

[21] E. T. Jaynes. Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):227–241, 1968.

[22] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[23] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. part I. In *Proceedings of the Royal society of London. Series A*, volume 115, pages 700–721, 1927.

[24] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. II. the problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

[25] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. III. further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A*, 141(843):94–122, 1933.

[26] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.

[27] C. A. Lin. Communicator in chief: How Barak Obama used new media technology to win the White House. *Journal of Broadcasting & Electronic Media*, 55(2):271–272, 2011.

[28] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, et al. Transmission dynamics and control of severe acute respiratory syndrome. *science*, 300(5627):1966–1970, 2003.

[29] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, page 3, 2005.

[30] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 889–892, New York, NY, USA, 2013. ACM.

[31] C. E. Mills, J. M. Robins, and M. Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

[32] D. Pittet, B. Allegranzi, H. Sax, S. Dharan, C. L. Pessoa-Silva, L. Donaldson, and J. M. Boyce. Evidence-based model for hand transmission during patient care and the role of improved practices. *The Lancet infectious diseases*, 6(10):641–652, 2006.

[33] D. Poole and A. E. Raftery. Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000.

[34] K. Puniyani, J. Eisenstein, S. Cohen, and E. P. Xing. Social links from latent topics in microblogs. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 19–20. Association for Computational Linguistics, 2010.

[35] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[36] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[37] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465. ACM, 2011.

[38] P. Rangachari. Evidence-based medicine: old french wine with a new canadian label? *Journal of the Royal Society of Medicine*, 90(5):280, 1997.

[39] D. M. Scott. *The new rules of marketing and PR: how to use social media, blogs, news releases, online video, and viral marketing to reach buyers directly*. Wiley. com, 2009.

[40] G. Smith. Models of *Mycobacterium bovis* in wildlife and cattle. *Tuberculosis*, 81(1):51–64, 2001.

[41] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[42] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[43] J. Wallinga and P. Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516, 2004.

[44] 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.

[45] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.