TRACING TOPIC DISCUSSIONS WITH THE EVENT-DRIVEN SIR MODEL FOR ONLINE FORUMS

Jiyoung Woo
Center for Information Security Technologies (CIST)
Korea University
Seoul, Korea
jy0503.woo@gmail.com

Sung Ho Ha[*]
School of Business Administration
Kyungpook National University
Daegu, Korea
hsh@mail.knu.ac.kr

Hsinchun Chen
Eller College of Management
University of Arizona
Arizona, USA
hchen@eller.arizona.edu

ABSTRACT

Members of online communities on social media (e.g., Web forums) disseminate and receive information by interacting with one another, a situation that provides a good opportunity to examine information diffusion on social media. This study models topic-level opinion diffusion in Web forums through an epidemic model, namely the Susceptible, Infective, and Recovered (SIR) model, which has been used to analyze disease outbreaks and knowledge diffusion. In addition, this study proposes an event-driven SIR model that reflects the effect of a given event on opinion diffusion. The proposed model incorporates the effects of news postings on social media in terms of the growth in numbers of potential authors, the increase in the infection rate for a given topic, and the acceleration of the transition from potential authors to active ones. This study evaluates the proposed model based on a large longitudinal data set from a Web forum of a major company. The analysis results show that the event-driven SIR model performs well in terms of the estimation of the number of authors who discuss a given topic, and outperforms the baseline SIR model in terms of model fit and the forecasting of major hot topics that reflect outbreaks of author participation. This study has some practical business implications: Web forums are places where corporate brands and reputations are formed, as evidenced by forum posts; and Web forums are advertising platforms for contacting audiences and tracking users in order to hone marketing messages and encourage discussions and reviews on a firm's products and services.

Keywords: Social media; Information diffusion; Online forum; SIR model

1. Introduction

The Internet has supplied new forms of interactivity such as blogs, online forums, and chat rooms [Robbin & Buente 2008]. This interactivity allows users to disseminate information by posting blog entries and discussing their opinions on Web forums. The dissemination process has a considerable influence on society, with numerous citizens and consumers actively expressing their opinions and preferences. Thus, information diffusion in the Web domain has become a major research topic, and studies reveal that it can influence public opinion and society both economically and politically [Flew 2005; Habermas 2006; Robbin & Buente 2008; Woo et al. 2011; Heverin & Zach 2012].

Since the arrival of the social media era, it is necessary for political, economic, and marketing purposes to understand more effectively the mechanism and properties of information diffusion through this new publication method [Wirtz et al. 2013]. For example, firms' strategic decisions can be made by examining a greater number of in-depth discussions and their dynamics from the perspectives of participants. Further, firms can leverage the information

[*] Corresponding author

diffusion process on the Web as a form of viral marketing for their marketing strategies. A careful examination of information diffusion also enables firms to determine how communication with consumers through Web 2.0 channels works and to predict how events such as product launches, promotions, campaigns, and related-news influence consumers.

The most powerful features of this diffusion process are its contagiousness and speed. However, as social media becomes more prevalent [Lai & To 2015], it is easier to propagate not only different views but also misleading or false information. Surprisingly, views found on the Web are sometimes accepted and transmitted to others without much critical examination [Wu & Liu 2008]. This highly contagious property accelerates the information diffusion process. Among the various forms of social media on the Internet, Web forums have been paid special attention as a popular platform for the formation of public opinion, and are considered to be a crucial source for information diffusion whereby users with common interests express and discuss their opinions and influence others. In this regard, interactions occur when an author posts a thread and other authors reply to that thread, thus enabling the author to infect others' opinions.

Many researchers have examined the diffusion mechanism and predicted diffusion behaviors in the fields of epidemiology and sociology. For example, the diffusion process encompasses diseases in biology; computer viruses in networks; and information, knowledge, rumor, social behavior, and innovation in society. Web forums, where participants disseminate as well as receive information online, and form self-contained communities, provide promising opportunities for modeling information diffusion based on the epidemic model. However, despite the attractiveness of Web forums, massive amounts of data and the casual writing styles of participants can make it difficult to analyze opinions on them. This indicates that substantial efforts and advances are needed to collect and analyze highly complex user interactions on Web forums.

This study considers Web forums in order to understand more effectively their information diffusion mechanism and examines whether the formation of opinions about topics on Web forums is similar to the epidemic process. For this, the study considers a large Web forum managed by a major company by using the Susceptible, Infective, and Recovered (SIR) model, a popular epidemic model consisting of mathematical differential equations based on rules of interactions among classes of participants. In addition, the study proposes a novel event-driven SIR model that encompasses the effect of an external event such as news on the number of postings for a given topic.

This study overcomes the several limitations that are evident in prior research. First, few studies have examined information diffusion on Web forums. Social media studies have been limited to blogs and have focused on how a network structure influences information flow. Moreover, few have considered diffusion at the topic level. Second, most studies addressing online information diffusion have focused only on interactions among individuals, ignoring external factors such as the effect of news media on the diffusion process. Third, most studies have conducted numerical simulations using synthetic data because of the difficulty in obtaining a large data set from the Web. Finally, studies applying the SIR model to information diffusion have considered the overall model fit as an evaluation measure and have seldom validated the model in terms of forecasting accuracy. The present study narrows these research limits by addressing the following questions:

- Can opinion diffusion for a topic on Web forums be modeled using a population-based epidemic model?
- Can the modeling accuracy of topic diffusion on Web forums be improved by adding the effect of news postings to the epidemic model?

This study has several practical business implications. Web forums are "social" in nature because they support communication among individuals and share opinions on the products and services of firms within user networks. Thus, the most visible business use of Web forums is as a branding and marketing tool. Web forums are where corporate brands and reputations are formed, as evidenced by forum posts. Web forums are used as advertising platforms to display advertisements for products and services of a company, encouraging discussions and reviews. The corporate can contact users on Web forums and track them in order to hone marketing messages to attract them. This study reveals the branding and marketing knowledge from the Web forums and gives business insights and strategies to the company.

2. **Literature Review**

Thoughts and opinions have a contagious property [Lynch 1996]. Patterns of the spread of epidemics and social contagion processes are similar; and thus, it is natural to address social contagion by using the same theoretical principles that are applied to epidemics. Researchers have developed various epidemic models to describe the spread of diseases, and some models designed for disease contagion have been applied to social contagion through minor modifications to explain various social phenomena. For example, innovation diffusion models have been designed to explain product/innovation adoption and, recently, information diffusion in social networks.

Information dissemination happens through human interactions; namely, social contagion encompassing diffusion processes through interactions among people. This includes the spread of technological innovations, word-of-mouth effects in marketing, and the spread of news and opinions. Kleinberg [2008] states that rumors, political messages, and links to Web pages can be examples of information that can spread from one individual to another in a contagious manner. The purpose of the diffusion model is to understand how the spread of new diseases/ideas/products works, to predict their success or failure in the early stages, and to shape the underlying process to increase or reduce the likelihood of diffusion. To begin with, we explain the widely adopted theories that govern the basis of modeling mechanism of diffusion process.

2.1. Diffusion Theory and Model

The most popular reference theories of diffusion can be classified into epidemic diffusion theories and innovation diffusion theories. Theories of epidemic diffusion define diffusion as the spread of memes by infection [Blackmore 1999]. Epidemic diffusion models are built based on the epidemic diffusion theory and consist of non-linear differential equations based on the population structure, assuming that people have a constant contact rate, are infected by a disease that has a unique transmission rate, and recover at a certain rate. Theories of innovation diffusion assume that diffusion occurs as a result of people acting as innovators and imitators. Such theories develop a logistic-equation method of cumulative adopters consisting of innovation and imitation terms [Mansfield 1961; Schmittlein & Mahajan 1982].

The other perspective on the classification of diffusion theories is whether diffusion occurs in a homogeneous population or in a heterogeneous network. A population model divides a population into classes reflecting the status of individuals in the population and assumes a homogeneous mixture of individuals that experience random contact. A network-based model reflects the network structure underlying the population and the network properties of individuals in diffusion modeling. Researchers have been encouraged by recent findings from real-world networks, including social networks and their topological features [Barabasi & Albert 1999], and have considered network structures under a diffusion process instead of random contact in homogeneous populations. Table 1 summarizes relevant theories of diffusion and selected references of diffusion models.

Table 1: Diffusion theories and relevant diffusion models

		Diffusion model
Epidemic diffusion theory	Population level	Deterministic susceptible, infective, and recovered (SIR), and susceptible, infective, and susceptible (SIS) models [Kermack & Mckendrick 1927]
	Network level	Complex network-based SIR and SIS models [Newman 2002]
Innovation diffusion theory	Population level	Rogers' model [1962], Bass' model [1969]
	Network level	Threshold model [Granovetter 1987], Independent cascade model [Goldenberg et al. 2001]

2.2. Modeling Methods

Based on how to generate diffusion models, they can be implemented as either equation- or agent-based model. Equation-based models (EBMs) operate based on global rules defined by equations and can be applied to all compartments. The assumption underlying EBMs is that the population is homogeneous and governed by holistic rules. Stochastic models use the concept of independent and identical distribution, and describe the diffusion process based on holistic rules [Bobashev et al. 2007]. EBMs depict the diffusion process at the macro level but cannot track individuals' status.

In agent-based models (ABMs), rules are defined at the individual level, which facilitates the capture of agents' local interactions and adaptive behaviors. The diffusion process in ABMs is tracked by aggregating the status of individuals. ABMs can better describe the diffusion process when it is more complex, includes a large number of equations, and is less tractable [Bobashev et al. 2007]. ABMs have a disadvantage in that they are computationally extensive and hard to analyze because of parametric complexity.

2.3. Evaluation Methods

Diffusion models have to be evaluated whether they are fit for the data that are used for modeling. There are three evaluation methods for diffusion models: empirical analyses, numerical simulations, and mathematical proofs.

Empirical analyses validate a diffusion model by using real data. A model is parameterized based on real data, and the estimated diffusion process is compared to the real diffusion process.

When real data are not available, the numerical simulation method using synthetic data can be an alternative. It, then, focuses on demonstrating the superiority of a proposed model over existing models under considered scenarios. Mathematical proofs concentrate on deriving statistical characteristics and evaluating models by testing their feasibility and equilibrium status based on mathematical statements and logic.

Physicians and mathematicians have adopted numerical simulations and mathematical proofs to validate the theoretical value of a model. However, most diffusion studies of products, innovations, and social behaviors have evaluated their models by using real data and focusing on the practical purpose of diffusion models. Many studies of online information diffusion, particularly in the context of social media, have provided empirical support by obtaining large data sets and representing the often complex and unstructured social interaction among people.

2.4. Application Domains

Information diffusion studies have been observed from a variety of application domains, and they can be classified into several areas, including scientific research, consumer and financial markets, and social media.

2.4.1. Scientific Research

Pioneering contributions to information diffusion modeling based on epidemic models have been made using research on the spread of scientific ideas. Goffman and Newill [1964] suggest that the transmission and development of scientific information within a population can be examined in terms of infectious disease. Bettencourt et al. [2006, 2008] examine how a physical theory is adopted by physicists using the SIR model and suggest the competency model to describe the diffusion process of two competing theories. Scientific collaboration networks inherent in the growth dynamics of research have been analyzed by Newman [2001]. Further, Cintr'on-Arias [2006] uses a genetic algorithm for the parameter estimation of the SIR model through an empirical study in the physics literature.

To minimize disease or virus diffusion in a population or network, researchers have investigated immunization strategies reflecting features of diseases or viruses, individuals, and population or network structures. Network-based epidemic models have been applied to the propagation of computer viruses in the late 1990s and early 2000s [Huberman & Admic 2004; Piqueira 2005; Yang et al. 2008]. Chen and Carley [2004] investigate the effects of countermeasure propagation on the prevalence of computer viruses, an approach that differs from other immunization strategies such as random immunization, target immunization, and kill-signal strategies. Dagon et al. [2006] devise a malware epidemic model reflecting the time zone and location effect on malware-spread dynamics. The epidemic models have also been tested for their fit to rumor propagation through email [Dietz 1967; Bampo et al. 2008; Kawachi 2008]. Kleinberg [2008] states that a rumor, a political message, and a link to a Web page are all examples of information that can spread from person to person, contagiously, in an epidemic style.

2.4.2. Consumer and Financial Markets

From a marketing perspective, online word-of-mouth communication has become a new topic in diffusion modeling. Goldenberg et al. [2001] conduct a study on the process of word-of-mouth on the networks. Domingos and Richardson [2001] examine customers' influence based on social networks with collaborative filtering and construct a model by using the Markov random field. Richardson and Domingos [2002] propose a probabilistic model to mine network data from online customers' reviews. Song et al. [2007] suggest a rate-based information flow model using the Markov chain and apply it to recommendation systems. Bampo et al. [2008] apply the SIR model to identify email networks and optimize e-mail marketing campaigns, and measure the efficiency of such campaigns. Leskovec et al. [2007] analyze the dynamics of viral marketing.

Explorations and algorithms have been developed to maximize diffusion [Sultan et al. 1990]. Kempe et al. [2003], and Kimura and Saito [2006] propose approximation algorithms that can enable diffusion to prevail in the network. Kleinberg [2007] provides a theoretical basis for identifying a set of influential nodes. Exploration approaches aim to determine the distribution of each cascade shape to trigger the most prevalent cascade and discover influential nodes based on network properties. Through the selection of *k* nodes with maximum influence, Song et al. [2007] identify opinion leaders in the blogosphere.

Epidemic models have also been used more recently to model the diffusion of financial information: Fan [1985] proposes the ideodynamics model that embeds people's contact and content characteristics, and Fan and Cook [2003] model consumer sentiment regarding the economy with the sentiment of mass media content. Shtatland and Shtatland [2008] approximate the SIR model into an autoregressive first-order model for financial outbreaks. Shive [2010] modifies the SIR model to predict the buying and selling of a stock by adding situational determinants such as the total amount of trade, return on investment, and the income level to social interactions.

2.4.3. Social Media

Because of the growth of email systems, Web sites, and social media, epidemic models have been considered for information diffusion on the Web, resulting in attempts to apply diffusion models to email messages, blogs, and forums

[Newman et al. 2002; Fu et al. 2006; Kubo et al. 2007; Bampo et al. 2008]. Studies of social networks have examined the characteristics of information diffusion by taking into account topological properties of social or technological networks. Such studies have focused on realistic networks reflecting social relationships and investigated how the diffusion process varies according to network properties. Haggith et al. [2003] examine how an idea is diffused in a network according to the features of the idea and the network structure. Pastor-Satorras and Vespignani [2001], Wu et al. [2004], and Wu and Huberman [2008] investigate the general characteristics of information diffusion in various networks by using the epidemic model. Keeling and Eames [2005] review a network-based epidemic model.

Blog communication as a proxy for information diffusion in online networks has attracted considerable attention from researchers. The intensive literature on social networks has led to studies of information diffusion through blogs based on the assumption that a blog implies a real-world social network. Users reflect their social relationships in blogs and build online social networks through them. Many studies of information diffusion in the blogosphere have employed the independent cascade model (ICM) and showed that the diffusion process takes place through contact among neighbors in the social network. Thus, the major research topics include inferring the probability of diffusion among nodes, maximizing influence, finding the distribution of each cascade shape, discovering influential nodes, and suggesting network generation models.

Gruhl et al. [2004] define the characteristics of topics that diffuse through blogs and propose a method for estimating the transmission probability for ICMs. Saito et al. [2008] use expectation maximization to estimate the transmission probability for ICMs. Java et al. [2006] conduct an experiment to maximize the number of infective individuals by using targeting methods such as PageRank, hits, and indegree centrality. Leskovec et al. [2007] analyze the pattern of cascades by using a large blog data set and suggest a cascade generation model under the Susceptible, Infective, and Susceptible (SIS) framework based on the fixed transmission probability. Kubo et al. [2007] draw an analogy among the disease propagation model, the SIR model, and data postings on Web forums. Zhou et al. [2008] predict the tendencies of topic discussions in online social networks by using a dynamic probability model that embeds individuals' interests, behaviors, and time lapses.

More recently, information diffusion studies have led to an examination of online social network applications. Sun et al. [2009] perform an empirical investigation of diffusion through a large social network site such as Facebook. They devise a regression model to identify factors affecting the diffusion chains. Cha et al. [2009] trace information dissemination in the Flickr social network. They find that even popular photos spread slowly and do not spread widely throughout the network, given its structure. Lerman and Ghosh [2010] measure the dynamics and distribution of fan votes in Twitter. Sakaki et al. [2010] monitor tweets on Twitter and detect a target event such as earthquakes. Romero et al. [2011] identify the differences in the mechanics of information diffusion across topics on Twitter. Table 2 selects previous studies dealing with information diffusion, which shows the information about diffusion vector, diffusion model, testbed, empirical analysis (Yes/No), and external factor (Yes/No).

2.5. Effects of News Media on Information Diffusion

A small number of studies have considered the effects of news media on information diffusion on the Web [Fan & Cook 2003]. At first, studies investigated how mass media influence the formation of public opinion from the perspective of sociology. Katz [1957] introduces a two-step flow of communication showing that influence flows first from mass media to opinion leaders and then from opinion leaders to society as a whole. Greenberg and Bradley [1964] suggest that news media and personal communication have a considerable influence on news diffusion, proposing a mixed model of diffusion by combining an element reflecting logistic and internal growth with another derived from external and exponential growth.

Valente [1993] suggests a mixed model combining adoption by interpersonal communication and awareness by news media. Since the arrival of social media, some studies have revealed that information disseminated through news media can heavily influence opinion formation and discussions on social media. Toole et al. [2012] model the adoption of Twitter by using the SIS model, which is modified to reflect the presence of geographic and media influences. Myers et al. [2012] show that information can reach a node via the links of the social network in Twitter or through the influence of external sources on Twitter. The authors say that approximately 71% of the information in Twitter results from network diffusion, and the remaining 29% is due to external events from outside the network.

However, several research gaps are identified in the literature. First, most studies of information diffusion through social media have been limited to blogs and have typically focused on how a network structure influences information flow. Few studies have examined information diffusion on Web forums and have considered diffusion at topic level. Second, most studies addressing information diffusion on the Web have focused on interactions among individuals, ignoring the effect of external factors such as news media on the diffusion process. Third, most studies have conducted numerical simulations using synthetic data because of the difficulty in obtaining a large data set from the Web. Finally, studies applying the SIR model to information diffusion have considered an overall model fit as the evaluation measure. They have seldom validated the model in terms of forecasting accuracy. The present study narrows these research gaps by applying a SIR model, a popular epidemic model consisting of mathematical differential equations, to a large Web forum and by then proposing a new event-driven SIR model that encompasses an external event such as news.

Table 2: Previous studies on information diffusion

Research	Diffusion vector	Diffusion model	Testbed	Empirical analysis (Yes/No)	External factor (Yes/No)
Kubo et al. [2007]	Discussion	Deterministic SIR	Forum	Yes	No
Zhou et al. [2008]	Discussion	Probability model	Forum	Yes	No
Pastor-Satorras and Vespignani [2001]	Computer virus	Network-based SIS	Email	No	No
Huberman and Admic [2004]	Computer virus	Network-based SIR	Email	No	No
Bampo et al. [2008]	Campaign email	Network-based SIR	Email	Yes	No
Gruhl et al. [2004]	Blog connection	Network-based SIR	Blog	Yes	No
Fu et al. [2006]	Blog connection	Network-based SIR	Blog	No	No
Leskovec et al. [2007]	Blog connection	Network-based SIS	Blog	Yes	No
Saito et al. [2008]	Blog connection	Network-based SIR	Blog	Yes	No
Bettencourt et al. [2006, 2008]	Physical theory	Deterministic SIR	N/A	Yes	No
Cintr'on-Arias [2006]	Physical theory	Deterministic SIR	N/A	Yes	No
Shive [2010]	Discussion	Deterministic SIR	Financial market	Yes	Yes
Shtatland and Shtatland [2008]	Discussion	Deterministic SIR	Financial market	No	No
Fan and Cook [2003]	Consumer sentiment	Deterministic SIR (variant)	Public opinion	Yes	Yes
Yang et al. [2008]	Computer virus	Network-based SIR, SIS	Portable device	No	No
Piqueira [2005]	Computer virus	Deterministic SIR	Computer network	No	No
Dagon et al. [2006]	Computer virus	Deterministic SIS	Computer network	Yes	No
Kleinberg [2007]	Wireless worm	Network-based SIR	Wireless network	No	No

3. Research Design

This section presents the baseline SIR and the proposed event-driven SIR models. The feasibility of applying the SIR models to topic diffusion on Web forums is discussed, and then the baseline SIR model is extended by adding influence factors, including external events. The fitting procedure for building the models and the forecasting procedure for validating them are also detailed.

3.1. Baseline SIR Model for Web Forums

In order to address the research questions, the SIR model is used as a modeling tool to represent the process of topic diffusion through continuous discussions about a topic on Web forums. The posting action causes instant contagion as well as responses to/from other users on the Web forum. The epidemic model that builds equations based on contact and instant infection between infective and susceptible individuals can be applied to depict the topic diffusion process on the Web.

Topic diffusion on Web forums can be explained within the epidemic discipline as follows: By participating in the discussion on a topic, an author may infect others with his or her ideas. The initial author posts a thread on a topic, and forum users with a certain level of interest in the topic may read the thread and start a discussion by posting comments on it. Some commenters and readers can post other threads on the topic, thereby infecting others through their posts. In the SIR model, users with a certain level of interest in a given topic conform to the class of susceptible users, and those authors who post threads or comments conform to the class of infective users. After a certain period of time, some authors stop participating in discussions and thus recover. Through the interactive discussion process, a topic diffuses from one author to another on the Web forum.

Because infection occurs through contact between susceptible and infective individuals, an increase in the number of infective authors is determined by effective contact between possible and existing authors. When *S*, *I*, and *R* indicate the numbers of susceptible, infective, and recovered individuals respectively, the possibility of contact between susceptible and infective individuals during a unit of time can be expressed as $S \times I$. The number of susceptible authors who become infected can be expressed as $\alpha \times S \times I$, in which α is the infection rate. The decrease in the number of infective authors is determined by $\beta \times I$, in which β is the recovery rate. The total change in the number of infective authors is expressed as $(\alpha \times S \times I) - (\beta \times I)$, which is the difference between the inflow of infective authors from the class of susceptible authors and the outflow of infective authors to the class of recovered authors. Table 3 describes the elements of the SIR model from the perspective of epidemics and topic diffusion in Web forums.

Table 3: The analogy between epidemics and topic diffusion in Web forums

Elements of SIR model	Epidemics	Topic diffusion in Web forums
What flows	Disease	Topic (keywords)
Susceptible	People who can have contact with an infective person and will possibly become infected	Possible authors (including commenters) who have latent interests in a topic
Infective	People who have a disease and will possibly infect others	Current authors whose posts recruit other authors
Recovered	People who recover from a disease and lose the power to infect others	Past authors who lose their influence with others
Infection rate: α	The probability of contact between an infective person and susceptible person	The probability of reading titles or posts on the topic
Recovery rate: β	The probability that the infective person recovers	The probability that authors lose their influence with others

Moreover, on Web forums, the cumulative number of users increases over time; then turnover and decline in an epidemic curve occur because fewer susceptible individuals are left to infect [Brauer 2008]. An emerging topic, however, can recruit existing users by detracting attention from old topics. Such an increase in the number of authors becomes a source of susceptible individuals to stimulate a second epidemic. Indeed, the time-series pattern of author participation shows the recurrence of peaks. Based on these observations, insights from logistic growth are included [Fan and Cook 2003].

The growth of the total population is determined by interactions between the recruiting pool and the total population. Because the number of people that a given topic can recruit is limited, the upper limit needs to be considered. This upper limit is referred to as the carrying capacity. The recruiting pool is determined by the gap between the carrying capacity and the total population. Because incoming users are initially susceptible, the growth term is included in the differential equation for the class of susceptible individuals. The population grows at a rate of reproduction proportional to both the existing population and the quantity of available resources. Consequently, the mathematical equations for the SIR model are formulated as follows:

$$\frac{dS}{dt} = \mu N(1-\frac{N}{K}) - \alpha SI, \qquad (1a)$$

$$\frac{dI}{dt} = \alpha SI - \beta I, \qquad (1b)$$

$$\frac{dR}{dt} = \beta I, \qquad (1c)$$

In Eq. (1a)-(1c), *S*, *I*, and *R* are individuals in three states such as susceptible, infective, and recovered respectively. In addition, *N* represents the total number of individuals at time *t*, namely, $N = S + I + R$; μ is the population growth rate; *K* is the carrying capacity; α is the infection rate; and β is the recovery rate.

3.2. Event-driven SIR Model

Fourt and Woodlock [1960] assume that the diffusion process is driven mainly by mass media communication or external influences. Bass [1969] assumes that innovators who adopt new products only because of mass media or external influences are present in the innovation diffusion process. These models that consider environmental factors in the diffusion process incorporate them as systematic terms so that the effect of an environmental factor is aggregated in the model because of the regularity of an event.

The proposed event-driven SIR model considers the randomness of an event and includes the effect of an occasional event as a random term in the equation. Some discussion topics come from news media, and the spread of a topic on Web forums is affected by news media. Studies have verified that news media have a considerable influence on public opinion. When there is an event related to a topic, news media tend to report the event immediately. An event is defined as a collection of documents within a time window, or is represented by a centroid aggregated from the content of related documents [Chen & Chen 2009].

In the context of Web forums, an event can be defined as excessive news postings. The importance of an event is proportional to the number of news postings about it. This suggests that the impact of an event on topic diffusion can be determined based on the number of news postings on the topic. When the number of postings on a topic exceeds the average number of postings for that topic, an event is assumed to have occurred at that time. This event increases people's interest in the topic because of their exposure to news media and can accelerate the infection rate; namely, the infection rate for posts increases as a result of the increase in the level of authors' interest. The event attracts potential authors on a topic by frequent postings and thus increases the number of susceptible individuals.

The event-driven SIR model incorporates the acceleration of the infection rate and the expansion of the class of susceptible individuals and that of infective ones. Event-influence terms are added to the differential equations for *S*, *I*, and the infection rate. With the incorporation of the effect of events into topic diffusion, the following model is derived:

$$\frac{dS}{dt} = (\mu + \delta_1 e)N(1 - \frac{N}{K}) - (\alpha + \delta_2 e)SI, \tag{2a}$$

$$\frac{dI}{dt} = (\alpha + \delta_2 e)SI - \beta I + \delta_3 eI, \tag{2b}$$

$$\frac{dR}{dt} = \beta I, \tag{2c}$$

These differential equations combine the baseline SIR model with the event effect. In Eq. (2a)-(2c), *S*, *I*, and *R* are individuals in three states such as susceptible, infective, and recovered respectively. In addition, α is the infection rate and β is the recovery rate, as in the case of Eq. (1a)-(1c). Further, *e* is a new variable included in the event-driven SIR model and is a dummy for the occurrence of a given event such that it is 1 if an event occurs at time *t* and 0 otherwise.

δ_1 represents a coefficient of the event effect on the growth of susceptible individuals. The event increases the population growth rate by as much as δ_1, as shown by $(\mu + \delta_1 e)$ in the equation (2a). δ_2 is a coefficient of the event effect on the acceleration of the infection rate. The event accelerates the infection rate by as much as δ_2, as shown by $(\alpha + \delta_2 e)$ in the equation (2b). δ_3 denotes a coefficient of the event effect on the growth of infective individuals. The infective individuals increase by as much as the event effect, δ_3, depending on the number of infective individuals. This effect term is expressed as $\delta_3 eI$ in the same equation (2b).

3.3. Model Fitting and Validating Procedures

The procedure for building (fitting) and evaluating the diffusion models follows the steps in Figure 1. First, the model fitting step uses the complete data set for diffusion to build both the baseline SIR and event-driven SIR models. Simulated annealing is employed as an optimization algorithm for parameter estimation. Simulated annealing is a well-known and effective optimization algorithm that attempts to avoid becoming trapped in local minima [Mendes & Kell 1998].

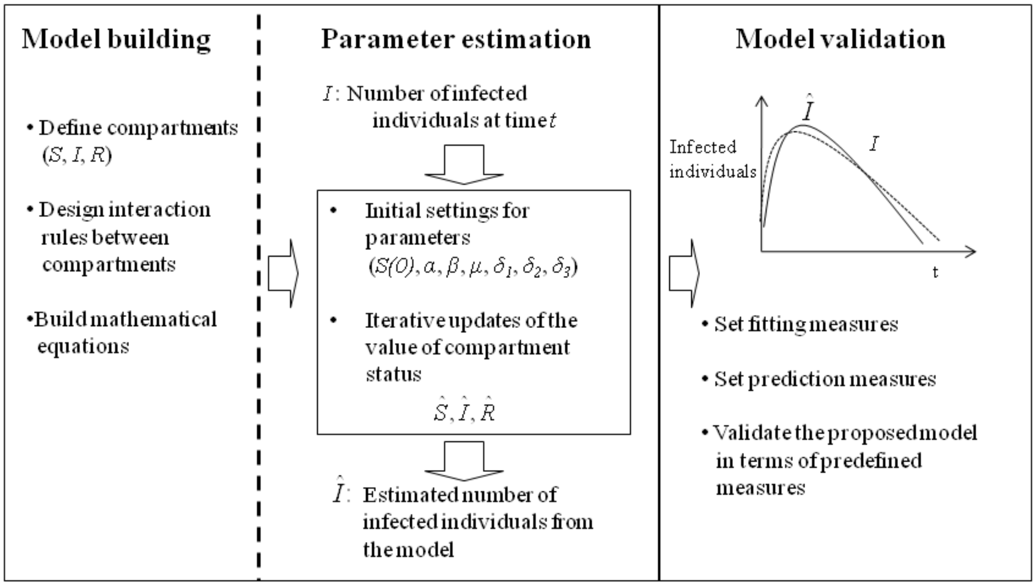
Model building
• Define compartments (S, I, R)
• Design interaction rules between compartments
•Build mathematical equations
Parameter estimation
I: Number of infected individuals at time t
• Initial settings for parameters (S(0), α, β, μ, δ1, δ2, δ3)
• Iterative updates of the value of compartment status
Ŝ, Î, R̂
Î: Estimated number of infected individuals from the model
Model validation
Î
I
Infected individuals
t
• Set fitting measures
• Set prediction measures
• Validate the proposed model in terms of predefined measures

Figure 1: Model construction and validation procedures for the SIR models

The initial conditions *(S(0), I(0), R(0))*, the parameters *(α, β, μ, δ$_1$, δ$_2$, δ$_3$)*, and the carrying capacity (*K*) are fed to the estimation process with the observed variable *I(t)*. Simulated annealing stops after successive iterations when non-linear least squares are achieved (i.e., residuals between the estimate and the observation are minimized) [Srinivasan & Mason 1986], as shown in Eq. (3):

$$\underset{\theta \in F}{\arg\min} \; J(\theta) = \sum_{i=1}^{n} (I_i - \hat{I}(t_i, \theta))^2 \tag{3}$$

Eq. (3) is the objective function for the parameter estimation. I_i is the observed value at time *i*, $\hat{I}_i$ is the estimated value from the model, *F* represents a feasible set of parameters, and *θ* represents a set of parameters to be estimated. After the two models are built, they are validated and compared in terms of forecasting performance. Model validation makes use of the same procedure with model fitting. The validation process uses the optimal parameter estimates and a data set available at the forecasting point.

Both the fitting and forecasting performance is evaluated on the basis of the goodness-of-fit, which is assessed in terms of the mean-squared error (MSE) and R-squared (R^2) values. They are computed as the following equations (4a)-(4b):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (I_i - \hat{I}(t_i, \hat{\theta}))^2 , \tag{4a}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (I_i - \hat{I}(t_i, \hat{\theta}))^2}{\sum_{i=1}^{n} (I_i - \bar{I})^2} , \tag{4b}$$

where I_i is the number of infective individuals at time *i*, $\hat{I}_i$ is the estimated number of infective individuals at time *i*, $\bar{I}$ is the average value of I_i, *n* is the number of samples, and $\hat{\theta}$ is the estimated parameter set.

4. Experimental Results

4.1. Research Test Bed

As one of the largest firms in the world, Wal-Mart has drawn attention from investors, analysts, and consumers on various platforms, particularly on various social media sites [Chen 2010]. Experiments were conducted using the Wal-Mart message board in Yahoo! Finance and the Wal-Mart-related news in the Wall Street Journal. The former discussion forum contains a longitudinal data set covering a 10-year period with various stakeholders actively expressing their opinions on various topics. The latter news outlet supplies nine years of news concerning Wal-Mart and is less likely to produce biased news reports because of a more balanced view of the firm [Schumaker & Chen 2009].

The first experiment applied the baseline SIR model to the Wal-Mart message board, and the following experiment built the event-driven SIR model based on the Wal-Mart message board and the Wall Street Journal. Table 4 shows the basic statistics for the forum and the news Web site. On the Web forum, the number of messages includes the number of threads and replies to these threads, and the number of users includes the authors of threads and posts.

Table 4: Description of the data set

Data source	Duration	# of threads	# of messages	# of users
Wal-Mart message board in Yahoo! Finance	Jan. 1999 - Jun. 2008	139,062	441,954	25,500
Wal-Mart-related news in Wall Street Journal	Aug. 1999 - Mar. 2007	N/A	4,081	657

4.2. Data Extraction

On the Wal-Mart message board in Yahoo! Finance, employees, investors, and customers discuss various topics. In order to collect data from the Web forum, a spider and a parser were developed. The spider crawled within the Web forum, which consists of Web pages linked to each other by hyperlinks. The spider extracted contents from a Web page; then, the parser detected the tags that contained required data fields such as the thread ID, message ID, author ID, posting time, and message. The parsed results were stored in a database.

Key topics were obtained from the Web forum by extracting mutual information, which measures the degree of relevance between a given topic and a word in the posts [Lin & He 2009; Xiaojun & Jianguo 2010]. The mutual information was calculated as follows:

$$I(w;t) = \sum_{t}\sum_{w} p(w,t)\ln\left(\frac{p(w,t)}{p(w)\,p(t)}\right) \tag{5}$$

In Eq. (5), regarding a word *w* and a given topic *t*, *p(w, t)* is the joint probability of *w* and *t*, *p(w)* is the marginal probability of *w*, and *p(t)* is the marginal probability of *t*. Mutual information is a measure of the dependence expressed in the joint distribution of *w* and *t* relative to the joint distribution of *w* and *t* under the assumption of independence. Thus, *I(w; t)* is equal to zero if and only if *w* and *t* are independent random variables [Matsuo & Ishizuka 2003].

For all words in posts, mutual information values were aggregated. Topics with high scores were then obtained. Those with intermittent bursts were excluded from the list of key topics because of their low frequency. Chatter topics with ongoing patterns were also excluded from the analysis because topics without epidemic patterns were not considered to be contagious or to cause contagion among users. Four topics showed major outbreaks based on their time-series patterns, as illustrated in Figure 2. The topics included a healthcare and insurance topic with the keywords “healthcare” and “health insurance,” a topic related to minimum wage and pay with the keywords “minimum wage” and “minimum pay,” a stock-price-related topic with the keywords “stock index” and “stock price,” and a topic related to product price with the keyword “low price.”

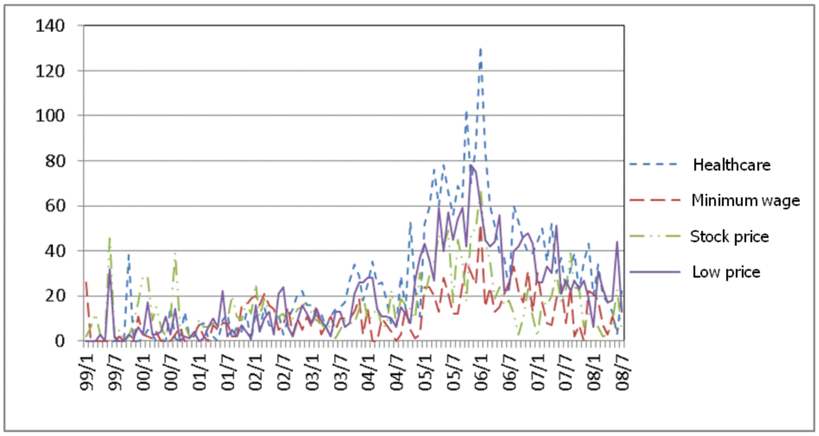

Figure 2: Time-series patterns of four selected Wal-Mart topics

The number of distinct authors (including commenters) was counted monthly for these four topics. The threads, including the pre-defined keywords, were accumulated, and the authors of the threads were counted. At this point, aggregate data based on the three-month moving average were used to smooth the time-series fluctuations. For the event-driven SIR model, the event indicator variable was calculated on a monthly basis from the number of news postings on the four topics. Table 5 shows the basic statistics for the four topics after the extraction.

Table 5: Description of the four topics

Topic	# of threads	# of messages	# of users
Healthcare and insurance	1,803	5,649	1,419
Minimum wage and pay	470	1,419	629
Stock price	581	1,763	912
Low price	953	2,858	1,142

4.3. Model Fitting Results

Model fitting has been conducted for one topic at a time. The data set that belongs to each topic was divided into training (50%) and testing data sets (50%). To begin with, the parameter set was fitted over the training data set (a five-year-long data set) for the healthcare and insurance topic. A simulated annealing algorithm, a method for parameter estimation within finite iterations, was used to minimize the objective function (the sum of the difference between estimated values implied by the estimator and actual values) [Brooks & Morgan 1995]. For each iteration of the simulated annealing algorithm, neighbor points within a certain distance were selected and their objective function values were checked. The acceptance of a candidate, which was one of the new neighbors of the current point, was defined as the objective function values of a candidate, the current point, and the time-varying parameters. A candidate point could be accepted when it raised the objective function value with a certain probability, and the state was moved to the candidate. Because of the randomness of the threshold value, the algorithm avoided entrapment in local minima and found global solutions.

Optimal parameter values were estimated at 120 for *S(0)*, 0.007 for α, 0.605 for β, 0.085 for μ, and 1,500 for *K*. This means that the number of forum users who may be interested in the healthcare and insurance topic and who may possibly become authors is approximately 120; the number of infective individuals who become infected through contact between an infective individual and susceptible individuals is 0.007 per month; and the number of infective individuals who recover is 0.605 per month. The carrying capacity for this topic is estimated to be approximately 1,500, with the population growing at a rate of 0.085 in proportion to the existing population.

By solving the differential equations for the baseline SIR model using the optimal parameter set, the numbers of susceptible, infective, and recovered individuals could be obtained for successive time periods, as illustrated in Figure 3. This shows that the SIR model reproduces the general shape of the time series with a good R-squared value of 0.52 and an MSE value of 289 for the healthcare and insurance topic. The SIR model generates smooth estimation values but sometimes cannot capture fluctuations in real data.

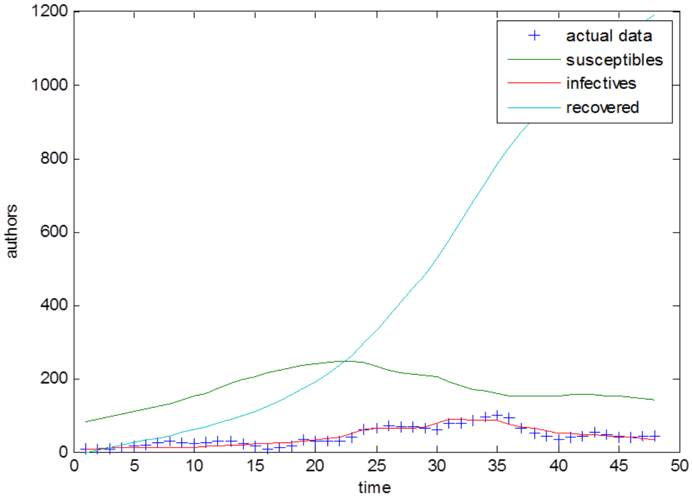

Figure 3: A curve of the baseline SIR model for the healthcare and insurance topic

Thus, the parameters of the event-driven SIR model were estimated using the same data set as the baseline SIR model. The optimal set $\hat{\theta} = (K, \hat{S}(0), \hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$ = (1,490, 82, 0.003, 0.590, 0.076, 0.054, 0.001, 0.003) was obtained after iterations of the simulated annealing algorithm. $\hat{S}(0)$, $\hat{\alpha}$, and $\hat{\beta}$ are interpreted in the same manner as in the baseline SIR model. The growth rate of susceptible individuals, $\hat{\delta}_1$, indicates that an event causing the number of postings on the news forum to exceed the average number of postings on a specific topic increases the number of existing susceptible individuals at a rate of 0.054 per susceptible individual.

An event accelerates the transmission rate by as much as 0.001, which is the estimated infection acceleration rate $\hat{\delta}_2$. In addition, the growth rate of infective individuals, $\hat{\delta}_3$, implies that an event increases the number of existing infective individuals at a rate of 0.003. Figure 4 shows the estimates for *S*, *I*, and *R* in the event-driven SIR model. When infective individuals show dramatic participation because of an event, the event-driven SIR model provides a better fit than the baseline SIR model, with a smaller MSE and a higher R-squared value.

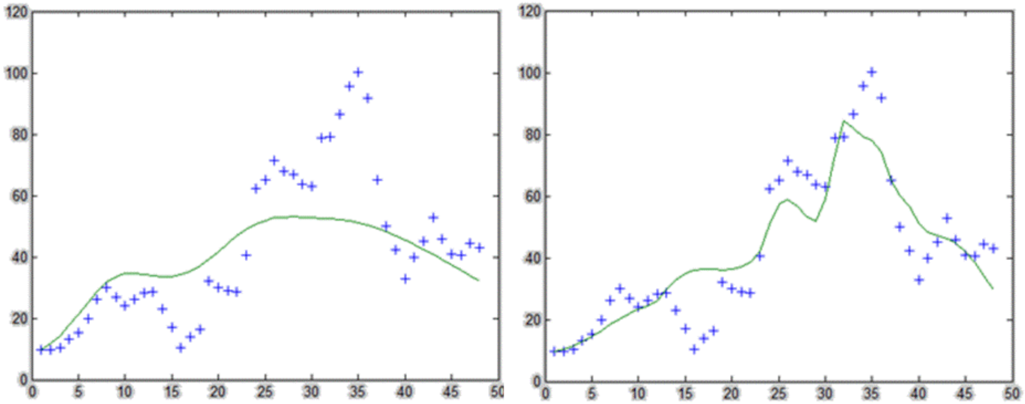

Figure 4: Curves for the healthcare and insurance topic: the baseline SIR model (left) and the event-driven SIR model (right)

Tables 6 and 7 summarize the results of fitting and parameter estimation for the baseline SIR and the event-driven SIR models respectively for the four topics.

Table 6: Parameter estimation and fitting results for the baseline SIR model

Topic	*K*	*S(0)*	*α*	*β*	*μ*	*MSE*	*R*^{2}
1	1,500	120	0.007	0.605	0.085	289	0.52
2	687	49	0.007	0.465	0.058	28	0.64
3	1,000	65	0.013	0.810	0.096	80	0.57
4	1,213	65	0.011	0.633	0.089	106	0.66

Note: Topic 1 is healthcare and insurance, Topic 2 is minimum wage and pay, Topic 3 is stock price, and Topic 4 is low price.

Table 7: Parameter estimation and fitting results for the event-driven SIR model

Topic	*K*	*S(0)*	*α*	*β*	*μ*	δ_1	δ_2	δ_3	*MSE*	*R*^{2}
1	1,490	82	0.003	0.590	0.076	0.054	0.001	0.003	90	0.85
2	700	50	0.018	0.662	0.046	0.073	0.016	0.000	26	0.66
3	992	79	0.002	0.661	0.072	0.019	0.001	0.005	25	0.86
4	1,208	71	0.004	0.562	0.046	0.075	0.001	0.004	35	0.89

Note: Topic 1 is healthcare and insurance, Topic 2 is minimum wage and pay, Topic 3 is stock price, and Topic 4 is low price.

In terms of the estimation results for the parameters, *α*, *β*, and *μ* are generally smaller in the event-driven SIR model than in the baseline SIR model. Because the event-driven SIR model considers the effects of external events on the original population, it reflects the diffusion effects through the interaction among classes less than the baseline

SIR model. Although the results suggest that the two models provide a good fit to the data in terms of their MSE and R^2 values, the results are also consistent with the event-driven SIR model outperforming the baseline SIR model in terms of its fit for the four major topics.

4.4. Model Validation

A diffusion model aims to depict the successive increase in the number of adopters and to predict the continued development of a diffusion process already in progress [Mahajan et al. 1990]. To argue that the two SIR models are appropriate for describing the underlying mechanism of information diffusion in the Web forum, the two models have to be compared with a time-series model, because it statistically identifies regular patterns from time-series data and generalizes a baseline model to predict future values based on the previous values. Thus, the autoregressive integrated moving average (ARIMA) models were adopted to compare them with the two SIR models.

A popular ARIMA model is referred to as *ARIMA*(*p*, *d*, *q*) where parameters *p*, *d*, and *q* are non-negative integers that represent the autoregressive, integrated, and moving average parts of the model, respectively [Box & Jenkins 1970]. Given time-series data, X_t, an ARIMA(p, d, q) model is mathematically expressed as follows:

$$Y_t = (1-L)^d X_t, \tag{6a}$$

$$(1-\sum_{i=1}^{p}\alpha_i L^i)X_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t, \tag{6b}$$

where Y_t are forecasts, *L* is the lag operator, α_i are the parameters of the autoregressive part, θ_i are the parameters of the moving average part, and ε_t are error terms.

Table 8 shows the four ARIMA models built for each topic and their goodness-of-fit measures on the basis of MSE and R^2 values. The MSE values are much higher than those for the baseline SIR and the event-driven SIR models, and the R^2 values are much lower than those for the counterpart models.

Table 8: Construction of ARIMA models for each topic

Topic	ARIMA Model	*MSE*	R^2
Healthcare and insurance	ARIMA(4,1,3) with intercept	671	0.29
Minimum wage and pay	ARIMA(3,1,3) without intercept	83	0.25
Stock price	ARIMA(2,1,3) without intercept	171	0.12
Low price	ARIMA(3,1,3) without intercept	272	0.23

Measuring the time-variant forecasting performance is a way to validate the diffusion model [Carbone & Longini 1977; Xie et al. 1997; Cintr’on-Arias 2006]. Table 9 shows that the overall goodness-of-fit of time-variant forecasting is lower than that of diffusion model construction (implying higher MSE and lower R^2 values). The event-driven SIR model outperforms the baseline SIR model in time-variant forecasting. The R^2 values for the event-driven SIR model are consistently higher than those for the baseline SIR model. The improvement in the MSE values for the event-driven SIR model over those for the baseline SIR model range from 61.5% (stock price), 48.7% (low price), and 39.8% (healthcare and insurance) to 11.4% (minimum wage and pay) in time-variant forecasting.

Table 9: Comparison of the two SIR models in time-variant forecasting

Topic	Baseline SIR model		Event-driven SIR model	
	MSE	R^2	*MSE*	R^2
Healthcare and insurance	323.6	0.23	194.7	0.54
Minimum wage and pay	41.1	0.30	36.4	0.38
Stock price	119.2	0.34	45.9	0.75
Low price	125.8	0.36	64.5	0.67

Figure 5 illustrates real values and forecasted values based on the baseline SIR and the event-driven SIR models for the four major topics.

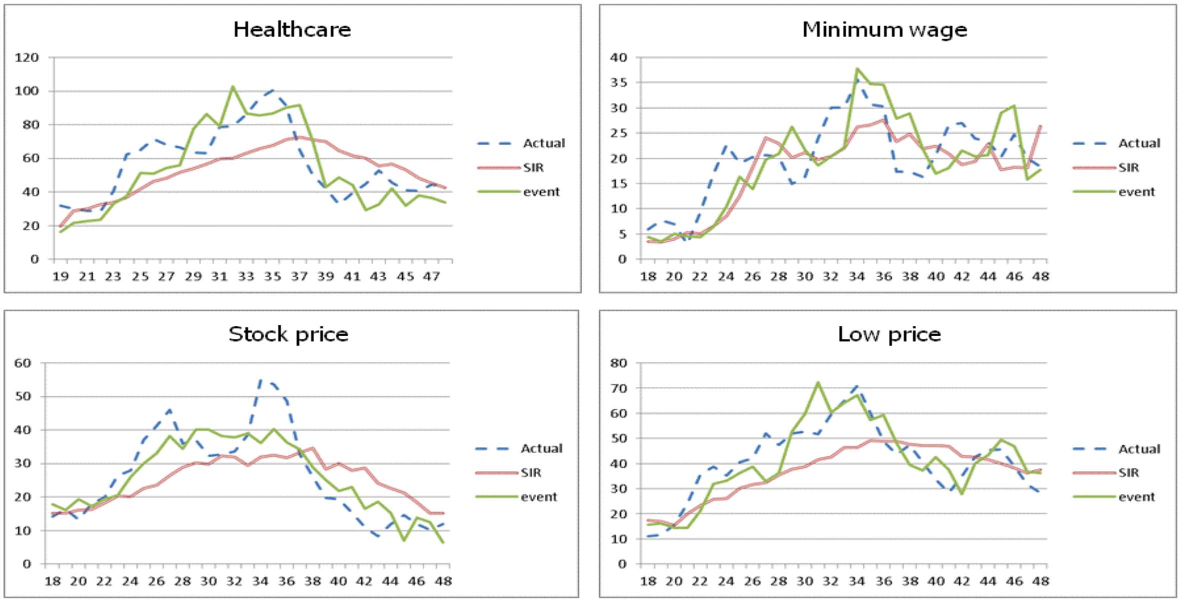

Figure 5: Forecasting values for the baseline SIR and the event-driven SIR models

Specifically, forecasting can aim to estimate the timing (when authors involved in a topic reach the maximum point) and magnitude (how many authors are involved at the maximum point) of the peak at an early stage of diffusion [Bass 1969; Kamakura & Balasubramanian 1987]. To achieve this, based on Wright et al. [1997], the period up to the peak is divided in half. The data from the first half of the pre-peak period are then used for training and forecasting, and the data from the second half are used for the validation of forecasting. The amount of available data varies incrementally until reaching the peak, and the parameters are estimated over the data available at each period.

Tables 10 and 11 display the forecasting errors with respect to timing and magnitude of the peak. Table 10 shows topics 1 and 2, and Table 11 shows topics 3 and 4. In the case of the first topic, healthcare and insurance, the time when the number of authors reach the maximum is 20; thus, the starting point of forecasting is 10 (= 20/2). The baseline SIR model predicts the peak number of authors who post on the topic with an error range of (-36%, 12%) and the peak time with an error range of (-7, 1). However, the event-driven SIR model predicts the peak number of authors with an error range of (-32%, 15%) and the peak time with an error range of (-6, 0). The event-driven SIR model generally predicts the timing (T_max) and magnitude (M_max) of the peak with a smaller error range than the baseline SIR model.

Table 10: Forecasting of the timing and magnitude of the peak for topics 1 and 2

Estimation point	Topic 1: healthcare and insurance				Topic 2: minimum wage and pay			
	Baseline SIR model		Event-driven SIR model		Baseline SIR model		Event-driven SIR model	
	M-diff	T-diff	M-diff	T-diff	M-diff	T-diff	M-diff	T-diff
10	-0.09	-2	**0.15**	**0**	-0.10	2	-0.20	**0**
11	**0.12**	**1**	-0.18	-4	-0.08	3	-0.16	0
12	-0.18	-4	-0.22	-5	**0.00**	**4**	-0.16	1
13	-0.09	-2	-0.29	**-6**	-0.21	-2	-0.15	1
14	**-0.36**	**-7**	**-0.32**	-6	-0.25	**-4**	**-0.11**	0
15	-0.32	-6	-0.27	-5	**-0.27**	-4	-0.19	1
16	-0.30	-6	-0.26	-5	-0.22	-2	**-0.26**	**-4**
17	-0.27	-5	-0.22	-5	-0.20	-2	-0.21	0
18	-0.23	-4	-0.14	0	-0.14	2	-0.17	1
19	-0.25	-4	-0.11	0				

T_max = 20, M_max = 100.33 T_max = 19, M_max = 242.33

Table 11: Forecasting of the timing and magnitude of the peak for topics 3 and 4

Estimation point	Topic 3: stock price				Topic 4: low price			
	Baseline SIR model		Event-driven SIR model		Baseline SIR model		Event-driven SIR model	
	M-diff	T-diff	M-diff	T-diff	M-diff	T-diff	M-diff	T-diff
10	**0.01**	**14**	0.10	**13**	**-0.39**	**-4**	**-0.19**	**-1**
11	-0.25	8	0.16	12	-0.31	-2	-0.23	-1
12	**-0.36**	-1	-0.01	11	-0.28	-1	**-0.43**	**-3**
13	-0.17	4	-0.16	1	-0.35	-3	-0.28	-1
14	-0.12	3	**0.22**	11	-0.34	-3	-0.32	-1
15	-0.09	5	-0.03	5	-0.32	-2	-0.31	-2
16	-0.18	2	-0.15	1	-0.34	-3	-0.31	-2
17	-0.26	0	-0.17	1	-0.30	**1**	-0.31	-1
18	-0.32	-2	-0.30	**0**	**-0.26**	-1	-0.29	-1
19	-0.33	**-3**	**-0.31**	0				

T_max = 21, M_max = 55 T_max = 19, M_max = 71

5. Conclusions

Studies have verified the applicability of the SIR model, which is widely used to analyze disease outbreaks and knowledge diffusion, to scientific theories, rumors, word-of-mouth communication, and computer viruses, among others. Based on such research, this study examined the feasibility of applying the SIR model to topic diffusion on Web forums that have the following properties: lack of social networking among authors and instant responses to others' opinions without cumulative effect. In terms of topical discussions on Web forums, this study defined the elements of the SIR model and built interaction rules among three classes of individuals: those who can be infected, those who are currently infected, and those who recover.

Hot topics occurred mainly by interactions among participants in the Web forums. Further, discussions on a given topic generated a steady and significant exponential rise, followed by a decline. This model can estimate how many authors have some latent interest in each topic in the initial stages of the diffusion process and how many are eventually involved in this process. In addition, it can compare the infection rate with the recovery rate for a given topic and identify highly infectious topics from the infection rate and long-lasting topics from the recovery rate.

A deterministic mathematical model such as the baseline SIR model simplifies the diffusion process and is useful for obtaining system-level measurements and for testing hypotheses based on them. According to the fitting and validation results, the baseline SIR model performs well in modeling topic diffusion on Web forums. However, it does not provide a better explanation of the diffusion pattern with fluctuations caused by an event. To address this limitation of the baseline SIR model, this study devised a new event-driven SIR model that incorporates the effect of events (i.e., news postings) on the topic diffusion process. From the application of the new model to a large Web forum, the experiment's results reveal that the parameter estimation procedure with simulated annealing converges to optimal values within finite iterations, and the event-driven SIR model outperforms the baseline SIR model in terms of fitting and forecasting performance.

Even though this study employed the Wal-Mart discussion forum, the proposed model and experimental design for testing are applicable to any Web forums that satisfy the following conditions: The topics discussed in the Web forums must be retrieved from the news media in order to incorporate the event effect, and the Web forums should allow the automatic crawler to capture the forum posts. Further, the proposed model is efficient when it deals with a specific application domain such as one that focuses on a relevant corporate and events. In summary, this study makes three important contributions by extending the literature on information diffusion to a new domain, namely Web forums; by examining the possibility of applying the epidemic model to topic diffusion on Web forums; and by devising an event-driven SIR model that adds external factors to improve modeling accuracy.

This study also reveals some branding and marketing knowledge from the Wal-Mart forum and gives business insights and strategies to the company. The most visible business use of Web forums is as a branding and marketing tool. As evidenced by forum posts, Web forums are where corporate brands and reputations are formed. In this sense, Web forums can become an extension of corporate customer relationship management and extend existing market research programs. Beyond branding, Web forums can also be used as advertising platforms to contact and attract audiences; a corporate can display advertisements for its products and services on Web forums, thereby encouraging discussions and reviews.

Future research can employ other modeling techniques. A genetic algorithm may improve the efficiency of parameter estimation by allowing the initial exploration to be omitted. The Kalman filter may be used to trace the diffusion curve over longer periods and trace periodic epidemics. Even though building a mathematical model is useful in incorporating forum characteristics into the diffusion model, other factors that may influence diffusion dynamics may also be included in the model. For instance, the sentiments of posts may determine the infectivity of a topic; a model can then be designed using an infection rate that varies according to the total sentiment score for posts. In addition, an epidemic model that takes account of sentiments reflected in news content may improve the accuracy of modeling sentiment diffusion on Web forums.

Another possible factor is the decay mechanism underlying Web forums. In this regard, it is common sense that when there is fresh news, some prior topics disappear. The same phenomenon may occur on Web forums such that when a new and influencing topic emerges, participants may stop discussing a prior topic. Indeed, the competing relationship between a new topic and an existing one can be incorporated into the event-driven SIR model. Future research may also provide experiments using other types of social media (e.g., Facebook or Twitter) in a systematic manner.

REFERENCES

Bampo, M., M.T. Ewing, D.R. Mather, D. Stewart, and M. Wallace, “The effects of the social structure of digital networks on viral marketing performance,” *Information Systems Research*, Vol. 19, No. 3:273-290, 2008.

Barabasi, A.L. and R. Albert, “Emergence of scaling in random networks,” *Science*, Vol. 286, 509-512, 1999.

Bass, F., “A new product growth model for consumer durables,” *Management Science*, Vol. 15, No. 5:215-227, 1969.

Bettencourt, L.M.A., A. Cintr’on-arias, D.I. Kaiser, and C. Castillo-chavez, “The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models,” *Physica A*, Vol. 364, 513-536, 2006.

Bettencourt, L.M.A., D.I. Kaiser, J. Kaur, C. Castillo-chavez, and D.E. Wojick, “Population modeling of the emergence and development of scientific fields,” *Scientometrics*, Vol. 75, No. 3:495-518, 2008.

Blackmore, S., The meme machine, Oxford: Oxford University Press, 1999.

Bobashev, G., J. Epstein, D. Goedecke, and F. Yu, “A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches,” In: *Proceedings of the Winter Simulation Conference*, Washington, DC, 1532-1537, 2007.

Box, G. and G. Jenkins, Time series analysis: Forecasting and control, CA, San Francisco: Holden, 1970.

Brauer, F., “Compartmental models in epidemiology,” In: Brauer, F., P. van Den Driessche, and J. Wu (Eds.), Mathematical Epidemiology, Springer, 19-79, 2008.

Brooks, S. and B. Morgan, “Optimization using simulated annealing,” *The Statistician*, Vol. 44, No. 2:241-257, 1995.

Carbone, R. and R. Longini, “A feedback model for automated real estate assessment,” *Management Science*, Vol. 24, No. 3: 241-248, 1977.

Cha, M., A. Mislove, and K.P. Gummadi, “A measurement-driven analysis of information propagation in the Flickr social network,” In: *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, 721-730, 2009.

Chen, H., “Business and market intelligence 2.0 Part 2,” *IEEE Intelligent Systems*, Vol. 25, No. 2:74-82, 2010.

Chen, L.C. and K.M. Carley, “The impact of countermeasure propagation on the prevalence of computer viruses,” *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 34, No. 2:823-833, 2004.

Chen, C.C. and M.C. Chen, “An adaptive threshold framework for event detection using HMM-based life profiles,” *ACM Transactions on Information Systems*, Vol. 27, No. 2:1-35, 2009.

Cintr’on-arias, A., Modeling and parameter estimation of contact processes, Dissertation of Cornell University, 2006.

Dagon, D., C. Zou, and W. Lee, “Modeling botnet propagation using time zones,” In: *Proceedings of the 13th Network and Distributed System Security (NDSS) Symposium*, San Diego, CA, 2006.

Dietz, K., “Epidemics and rumours: A survey,” *Journal of the Royal Statistical Society*. *Series A (General)*, Vol. 130, No. 4:505-528, 1967.

Domingos, P. and M. Richardson, “Mining the network value of customers,” In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 57-66, 2001.

Fan, D.P., “Ideodynamics - The kinetics of the evolution of ideas,” *Journal of Mathematical Sociology*, Vol. 11, No. 1:1-24.

Fan, D.P. and R.D. Cook, “A differential equation model for predicting public opinions and behaviors from persuasive information: Application to the index of consumer sentiment,” *Journal of Mathematical Sociology*, Vol. 27, No. 1:29-51, 2003.

Flew, T., New media: An introduction, Melbourne: Oxford University Press, 2005.

Fourt, L.A. and J.W. Woodlock, “Early prediction of market success for new grocery products,” *The Journal of Marketing*, Vol. 25, No. 2:31-38, 1960.

Fu, F., L. Liu, and L. Wang, “Information propagation in a novel hierarchical network,” arXiv:math/0605293v1, 2006.

Goffman, W. and V.A. Newill, “Generalization of epidemic theory: An application to the transmission of ideas,” *Nature*, Vol. 204, 225-228, 1964.

Goldenberg, J., B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,” *Marketing Letters*, Vol. 12, No. 3:211-223, 2001.

Granovetter, M., “Threshold models of collective behavior,” *American Journal of Sociology*, Vol. 83, No. 6:1420-1443, 1987.

Greenberg, G. and S. Bradley, “Diffusion of news of the Kennedy assassination,” *The Public Opinion Quarterly*, Vol. 28, No. 2:225-232, 1964.

Gruhl, D., R. Guha, D. Liben-nowell, and A. Tomkins, “Information diffusion through blogspace,” In: *Proceedings of the 13th International Conference on World Wide Web*, New York, 491-501, 2004.

Habermas, J., “Political communication in media society: does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research,” *Communication Theory*, Vol. 16, No. 4:411-426, 2006.

Haggith, M., R. Prabhu, C.J, Colfer, B. Ritchie, A. Thomson, and H. Mudavanhu, “Infectious ideas: Modeling the diffusion of ideas across social networks small-scale forest economics,” *Management and Policy*, Vol. 2, No. 2:225-239, 2003.

Heverin, T. and L. Zach, “Use of microblogging for collective sense-making during violent crises: A study of three campus shootings,” *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 1:34-47, 2012.

Huberman, B.A. and L.A. Admic, “Information dynamics in the networked world,” In: *Proceedings of the 23rd Annual Conference of the Center for Nonlinear Studies at Los Alamos National Laboratory*, Santa Fe, NM, 371-398, 2004.

Java, A., P. Kolari, T. Finin, and T. Oates, “Modeling the spread of influence on the blogosphere,” URL:<http://ebiquity.umbc.edu/paper/html/id/300/Modeling-the-Spread-of-Influence-on-the-Blogosphere>, 2006.

Kamakura, W.A. and S.K. Balasubramanian, “Long-term forecasting with innovation diffusion models: The impact of replacement purchases,” *Journal of Forecasting*, Vol. 6, 1-19, 1987.

Katz, E., “The two-step flow of communication: An up-to-date report on a hypothesis,” *The Public Opinion Quarterly*, Vol. 21, No. 1:61-78, 1957.

Kawachi, K., “Deterministic models for rumor transmission,” *Nonlinear Analysis: Real World Applications*, Vol. 9, No. 5:1989-2028, 2008.

Keeling, M.J. and K. Eames, “Network and epidemic models,” *Journal of the Royal Society Interface*, Vol. 2, No. 4:295-307, 2005.

Kempe, D., J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 137-146, 2003.

Kermack, W.O. and A.G. Mckendrick, “A contribution to the mathematical theory of epidemics,” In: *Proceedings of the Royal Society*, London, 700-721, 1927.

Kimura, M. and K. Saito, “Tractable models for information diffusion in social networks,” In: *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data*, Singapore, 259-271, 2006.

Kleinberg, J., “Cascading behavior in networks: algorithmic and economic issues,” In: Nisan, N., T. Roughgarden, E. Tardos, and V.V. Vazirani (Eds.), Algorithmic Game Theory, UK, Cambridge: Cambridge University Press, 613-632, 2007.

Kleinberg, J., “The convergence of social and technological networks,” *Communications of the ACM*, Vol. 51, No. 11:66-72, 2008.

Kubo, M., K. Naruse, H. Sato, and T. Matubara, “The possibility of an epidemic meme analogy for web community population analysis,” In: Yin, H., P. Tino, E. Corchado, W. Byrne, and X. Yao (Eds.), Intelligent Data Engineering and Automated Learning, Springer, 1073-1080, 2007.

Lai, L.S.L. and W.M. To, “Content analysis of social media: A grounded theory approach,” *Journal of Electronic Commerce Research*, Vol. 16, No. 2:138-152, 2015.

Lerman, K. and R. Ghosh, “Information contagion: An empirical study of the spread of news on Digg and Twitter social networks,” In: *Proceedings of International AAAI Conference on Weblogs and Social Media*, Washington, DC, 90-97, 2010.

Leskovec, J., L.A. Admic, and B.A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web*, Vol. 1, No. 1: doi:10.1145/1232722.1232727, 2007.

Lin, C. and Y. He, "Joint sentiment/topic model for sentiment analysis," In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, 375-384, 2009.

Lynch, A., Thought contagion: How belief spreads through society, NJ, New York: Basic Books, 1996.

Mahajan, V., E. Muller, and F.M. Bass, "New product diffusion models in marketing: A review and directions for research," *The Journal of Marketing*, Vol. 54, No. 1:1-26, 1990.

Mansfield, E., "Technical change and the rate of imitation," *Econometrica*, Vol. 29, No. 4:741-766, 1961.

Matsuo, Y. and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 1:157-169, 2004.

Mendes, P. and D. Kell, "Non-linear optimization of biochemical pathways: Applications to metabolic engineering and parameter estimation," *Bioinformatics*, Vol. 14, No. 10:869-883, 1998.

Myers, S.A., C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, Beijing, China, 33-41, 2012.

Newman, M.E.J., "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 98, No. 2:404-409, 2001.

Newman, M.E.J., "The spread of epidemic disease on networks," *Physical Review E*, Vol. 66, No. 1:016128, 2002.

Newman, M.E.J., S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Physical Review E*, Vol. 66, No. 3:035101-1-4, 2002.

Pastor-satorras, R. and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, Vol. 86, No. 14:3200-3203, 2001.

Piqueira, J.R.C., "Epidemiological models applied to viruses in computer networks," *Journal of Computer Science*, Vol. 1, No. 1:31-34, 2005.

Richardson, M. and P. Domingos, "Mining knowledge-sharing sites for viral marketing," In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, 61-70, 2002.

Robbin, A. and W. Buente, "Internet information and communication behavior during a political moment: The Iraq war, March 2003," *Journal of the American Society for Information Science and Technology*, Vol. 59, No. 14: 2210-2223, 2008.

Rogers, E.M., The diffusion of innovations, NJ, New York: Free Press of Glencoe, 1962.

Romero, D.M., B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," In: *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, 695-704, 2011.

Saito, K., R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," In: *Proceedings of 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, Zagreb, Croatia, 67-75, 2008.

Sakaki, T., 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*, Raleigh, NC, 851-860, 2010.

Schmittlein, D.C. and V. Mahajan, "Maximum likelihood estimation for an innovation diffusion model of new product acceptance," *Marketing Science*, Vol. 1, No. 1:57-78, 1982.

Schumaker, R.P. and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Transactions on Information Systems*, Vol. 27, No. 2:1-19, 2009.

Shive, S., "An epidemic model of investor behavior," *Journal of Financial and Quantitative Analysis*, Vol. 45, No. 1:169-198, 2010.

Shtatland, E.S. and T. Shtatland, "Early detection of epidemic outbreaks and financial bubbles using autoregressive models with structural changes," In: *Proceedings of the Northeast SAS Users Group Annual Conference*, Pittsburgh, 2008.

Song, X., Y. Chi, K. Hino, and B.L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," In: *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, 191-200, 2007.

Srinivasan, V. and C.H. Mason, "Non-linear least squares estimation of new product diffusion models," *Marketing Science*, Vol. 5, No. 2:169-178, 1986.

Sultan, F., J.U. Farley, and D.R. Lehmann, "A meta-analysis of applications of diffusion models," *Journal of Marketing Research*, Vol. 27, No. 1:70-77, 1990.

Sun, E., I. Rosenn, C.A. Marlow, and T.M. Lento, “Gesundheit! modeling contagion through Facebook news feed,” In: *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, San Jose, CA, 2009.

Toole, J., M. Cha, and M.C. Gonzalez, “Modeling the adoption of innovations in the presence of geographic and media influences,” *PLoS One*, Vol. 7, No. 1: doi:10.1371/journal.pone.0029528, 2012.

Valente, T.W., “Diffusion of innovations and policy decision-making,” *Journal of Communication*, Vol. 43, No. 1:30-45, 1993.

Wirtz, B.W., R. Piehler, and S. Ullrich, “Determinants of social media website attractiveness,” *Journal of Electronic Commerce Research*, Vol. 14, No. 1:11-33, 2013.

Woo, J., J. Son, and H. Chen, “An SIR model for violent topic diffusion in social media,” In: *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, 15-19, 2011.

Wright, M., C, Upritchard, and T. Lewis, “A validation of the bass new product diffusion model in New Zealand,” *Marketing Bulletin*, Vol. 8, 15-29, 1997.

Wu, F. and B.A. Huberman, “Social structure and opinion formation,” URL:<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.10&rep=rep1&type=pdf>, 2008.

Wu, F., B.A. Huberman, L.A. Adamic, and J. Tyler, “Information flow in social groups,” *Physica A*, Vol. 337, No. 1-2:327-335, 2004.

Wu, X. and Z. Liu, “How community structure influences epidemic spread in social networks,” *Physica A*, Vol. 387, No. 2-3:623-630, 2008.

Xiaojun, W. and X. Jianguo, “Exploiting neighborhood knowledge for single document summarization and key phrase extraction,” *ACM Transactions on Information Systems*, Vol. 28, No. 2:1-34, 2010.

Xie, J., X.M. Song, M. Sirbu, and Q. Wang, “Kalman filter estimation of new product diffusion models,” *Journal of Marketing Research*, Vol. 34, No. 3:378-393, 1997.

Yang, S., H. Jin, X. Liao, and S. Liu, “Modeling modern social-network-based epidemics: A case study of rose,” In: Rong, C., M.G. Jaatun, F.E. Sandnes, L.T. Yang, and J. Ma (Eds.), Autonomic and Trusted Computing, Springer, 302-315, 2008.

Zhou, Y., X. Guan, Z. Zhang, and B. Zhang, “Predicting the tendency of topic discussion on the online social networks using a dynamic probability model,” In: *Proceedings of the Hypertext Workshop on Collaboration and Collective Intelligence*, Pittsburgh, 7-11, 2008.