Propagation Modeling and Analysis of Incidental Topics in Blogosphere

Li Zhao[1], Ruixi Yuan[1], Xiaohong Guan[1,2], and Mingyang Li[1]

[1] Center for Intelligent and Networked Systems and TNLIST Lab
Tsinghua University, Beijing, China 100084
[2] MEKLINNS Lab and SKLMS Lab, Department of Automation
Xian Jiaotong University, Xian, China 710049
{zhaoli4, li-my02}@mails.tsinghua.edu.cn,
{ryuan, xhguan}@mail.tsinghua.edu.cn

Abstract. Blog has become one of the most important media among the general public, and the propagation modeling of incidental topics in blogosphere is of great interest in social network studies. Most existing analysis methods are based on the infection models in epidemiology. However, many of these models are inconsistent with the widely observed power-law decay of the propagation velocity. In this paper, the propagation of incidental topics is described by a susceptible infection (SI) model based on the individual fitness. It is proved that the propagation velocity will asymptotically drop with power-law if the fitness density function satisfies certain conditions. Moreover, if the individual fitness is of uniform distribution, analytical solution of propagation velocity can be obtained based on our model. Model verifications are performed on the data from several widely discussed popular topics in Sina Blog and the results show that our model is consistent with the actual propagations.

Keywords: propagation modeling, blogosphere, SI model, fitness, power-law decay.

1 Introduction

Blog has become one of the important modern media among the general public, and the propagation modeling of incidental topics in blogosphere is of great interest in social network studies since a large number of netizens are involved in creating and spreading contents in blogosphere due to its accessible and timely nature. The Research Report of 2007 China Blog Market pointed out that the number of blog authors in China is about 47 million in December 2007 [1], 29 million more than in August 2006 [2]. With the explosion of online user-generated contents, there is a pressing need to understand the temporal propagation patterns of such contents in order to predict the trend of their propagation.

Most existing analysis methods are based on the infection models in epidemiology, where different underlying social network topologies serve to explain various spreading behaviors. Epidemic models on distinctive networks, such as exponential networks,

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, LNCS 5621, pp. 401–410, 2009.

scale-free networks and weighted scale-free networks, have been thoroughly studied [3, 4, 5, 6, 7]. The stationary properties and early temporal behaviors of disease propagations are well discussed [3, 4, 6]. Generally, the dynamical evolutions of the epidemic process on complex networks are studied by simulations [6, 7], as analytical solutions are not available in most cases.

The epidemic models built on network topologies have two shortcomings when applied in investigating the propagation process of incidental topics in blogosphere. First, these epidemic models need an underlying network which is difficult to obtain in the blogosphere. The most credible evidence of the spreading route is the reference link between posts. However, more than 70% people do not leave sources in their postings when they obviously get information from somewhere else [8]. Moreover, many blog authors know of the topics from traditional media rather than from inside the blogosphere. Second, the power-law decay of propagation velocity is widely observed [9]. However existing models can hardly reflect and analyze this phenomenon with finite population sizes.

In this paper, a susceptible/infective (SI) epidemic model based on individual fitness is developed to model the propagation process of a fundamental type of topics, the incidental topics, in blogosphere. In our model, the fitness is assumed to be the key parameter to get infected instead of the degree or strength used in the SI models to characterize the node's activeness on writing about the topic.

In Section 2, we formulate the dynamics of infection density by a set of differential equations to obtain the profile of the dynamic behavior of the infection process. In Section 3, it is proved that the propagation velocity will asymptotically drop with power-law if the fitness density function satisfies certain conditions, and analytical solution is obtained if the fitness is of uniform distribution. Comparisons of our model with the SI models on different underlying networks are made. In Section 4, model verifications are performed with the data from several popular topics. Simulation results show that the model is consistent with the actual propagations.

2 The Propagation Model of Incidental Topics

The incidental topics are usually externally induced by an incident in real world and a quick rise in the number of postings will occur shortly after the incident, leaving only one significant peak in the whole propagation process. It can serve as a fundamental for investigating more complicated types of topics such as those stimulated by a sequence of interrelated real world events. The incidental topics studied here do not include topics of sustained interest over a long period, e.g. comments on the new technologies, or topics driven inside the blogosphere.

In this section, the propagation population is defined in subsection 2.1, and the spreading mechanism is defined in subsection 2.2. Then the dynamic model of the propagation process is put forward in subsection 2.3.

2.1 The Propagation Population

Define $\Omega = \{V, W\}$ as the population involved the propagation of an incidental topic, where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of individuals interested in the topic, and $W = \{w_1, w_2, \ldots, w_N\}$ is the fitness set of the individuals in V. Both V and W are of size N.

Fitness is a measurement of the individual's intrinsic characteristics in the population, illustrating the individual's capability of learning about the topic and writing it on his blog. The fitness of each individual is assumed to be a positive random number, whose density function is denoted as $\rho(w)$.

Similar to the classical epidemic models, each individual is assumed to have two discrete states: Susceptible (S) and Infected (I). All individuals are susceptible at the beginning, and an individual is infected if he writes about the topic. Once infected, an individual will never go back to susceptible state. At time t, the there are $S(t)$ susceptible individuals and $I(t)$ infected individuals, and obviously $N = S(t) + I(t)$. The corresponding densities of susceptible and infected individuals are denoted as $s(t) = S(t)/N$ and $i(t) = I(t)/N$, respectively. Moreover, the propagation velocity $v(t)$ is defined as the differential of $i(t)$ and the increase rate in the number of infected individuals at t is $V(t) = Nv(t)$, i.e. the number of newly infected individuals in $(t, t+\Delta t]$ can be approximated as $I(t) - I(t-\Delta t) = Nv(t)\Delta t$ when Δt is small.

2.2 Spreading Mechanism

As discussed above, an individual gets infected based on its fitness. Different individuals may have different fitnesses in the population, and the distribution of the fitness represents the overall characteristics of the population. The infection rate is the probability of a susceptible individual gets infected in unit time. The infection rate should be is a monotonically increasing function of fitness. The following assumption on the infection rate is made in this paper.

Assumption: *Fitness based infection rate*. For an individual with fitness w, its infection rate $\lambda(w)$ is defined as

$$\lambda(w) = w^{\alpha}, \tag{1}$$

where $\alpha > 0$.

The assumption above means that in a small time interval Δt, the individual gets infected with probability $w^{\alpha}\Delta t + o(\Delta t)$ where $o(\Delta t)$ is ignored when Δt is small. In Section 3.2 we will shortly explain the reason why (1) is a simple and proper mapping function from fitness to infection rate.

2.3 Dynamic Propagation Model

Based on the assumption above, the propagation dynamics is built as follows:

$$\frac{di(w,t)}{dt} = (1 - i(w,t))\lambda(w). \tag{2}$$

where $i(w,t)$ is the infected density with given fitness w. Given initial condition $i(w,0) = 0$, the solution of (2) is

$$i(w,t) = 1 - e^{-\lambda(w)t}, \tag{3}$$

and the infected density $i(t)$ is

$$i(t) = \int_0^\infty \rho(w) i(w,t) dw. \tag{4}$$

Then the propagation velocity is obtained as

$$v(t) = \frac{di(t)}{dt} = \int_0^\infty \rho(w) e^{-\lambda(w)t} \lambda(w) dw. \tag{5}$$

The changing pattern of $v(t)$ is of interest itself and the details will be explored in the next section.

3 Analysis of Propagation Velocity

In subsection 3.1, we prove that if the fitness distribution satisfies some condition, the propagation velocity will drop with power-law form. As a simple and practical case, the analytical solution of $v(t)$ is obtained in subsection 3.2 when the fitnesses is of uniform distribution. The comparison between the proposed model and other classical SI models on several important networks are presented in subsection 3.3.

3.1 The Power-Law Decay of Propagation Velocity

It will be proved in this subsection that the propagation velocity drops with power-law when some condition is met and the exponent of the power-law decay is determined by the parameter α in infection rate definition (Equation (1)). Since the analytical solution of $v(t)$ is difficult to obtain under an arbitrary fitness distribution, the main idea of the proof is to apply the double side approximate method to investigate the property of $v(t)$ for large t.

Proposition 1: If the domain of the individual fitness density function $\rho(w)$ is right semi-continuous at 0 and $\rho(0) > 0$, the propagation velocity $v(t)$ in (5) asymptotically drops with power-law form and the exponent is $-(1+1/\alpha)$ when t goes to infinity.

Proof: Since $\rho(w)$ is right semi-continuous at 0, there exists $\varepsilon > 0$, s.t. $0 < \min(\rho(w)) < +\infty$ over $w \in [0, \varepsilon]$. Let

$$\theta_1 = \min_{0<w<\varepsilon} (\rho(w)) \tag{6}$$

Consider

$$v_1(t) = \int_0^\varepsilon \theta_1 e^{-w^\alpha t} w^\alpha dw, \tag{7}$$

and obviously $v_1(t) < v(t)$. Let $x = w^\alpha t$, then it can be proved that

$$v_1(t) = \frac{1}{\alpha t^{1+\frac{1}{\alpha}}} \theta_1 \gamma(1+\frac{1}{\alpha}, \varepsilon^\alpha t), \tag{8}$$

where $\gamma(a, x)$ is the lower incomplete gamma function. $\gamma(a, x)$ increases monotonously with x when $x > 0$ and will approach the steady value $\Gamma(a)$, thus

$$\lim_{t->\infty} \gamma(1+1/\alpha, \varepsilon^{\alpha} t) = \Gamma(1+1/\alpha). \tag{9}$$

Thereby, $v_1(t)$ will approach the line

$$v_1(t) \sim \frac{1}{\alpha t^{1+\frac{1}{\alpha}}} \theta_1 \Gamma(1+\frac{1}{\alpha}), \tag{10}$$

in log-log coordinate plane.

Let

$$\theta_2 = \max_{0<w<\varepsilon} (\rho(w)), \theta_3 = \max_{\varepsilon<w<\infty} (\rho(w)) \tag{11}$$

Consider

$$v_2(t) = \int_0^{\varepsilon} \theta_2 e^{-w^{\alpha} t} w^{\alpha} dw + \int_{\varepsilon}^{\infty} \theta_3 e^{-w^{\alpha} t} w^{\alpha} dw. \tag{12}$$

Obviously $v(t)<v_2(t)$. Using the same calculating technique we have

$$v_2(t) = \frac{1}{\alpha t^{1+\frac{1}{\alpha}}} (\theta_2 \gamma(1+\frac{1}{\alpha}, \varepsilon^{\alpha} t) + \theta_3 \Gamma(1+\frac{1}{\alpha}, \varepsilon^{\alpha} t)), \tag{13}$$

where $\Gamma(a, x)$ is the upper incomplete gamma function. $\Gamma(a, x)$ exponentially decreases with x and the second addend in (13) can be ignored when t is large. Thus $v_2(t)$ will approach the line

$$v_2(t) \sim \frac{1}{\alpha t^{1+\frac{1}{\alpha}}} \theta_2 \Gamma(1+\frac{1}{\alpha}), \tag{14}$$

in log-log coordinate plane.

Let $\rho_0 = \rho(0)$, we know $\lim_{\varepsilon \to 0} \theta_1(\varepsilon) = \lim_{\varepsilon \to 0} \theta_2(\varepsilon) = \rho_0$. When ε goes to infinitesimal, the bounds of $v(t)$, $v_1(t)$ and $v_2(t)$, will be getting close to each other, then the large-t tail of $v(t)$ can be given by

$$v(t) = \frac{\rho_0}{\alpha} \Gamma(1+\frac{1}{\alpha}) t^{-1-\frac{1}{\alpha}}. \tag{15}$$

Thus the proposition is proved. □

From the proposition we see that the exponent is determined only by α, and ρ_0 is related to the offset of the line in log-log graph.

3.2 The Solution and Analysis under Uniform Distribution

In this subsection, the analytical solution of $v(t)$ is obtained when the fitness is of uniform distribution.

Proposition 2: In the model defined in Section 2, if the individual fitnesses obey uniform distribution, i.e., $\rho(w)=I_{[0,b]}(w)$, $b>0$, the propagation velocity can be solved as

$$v(t) = \frac{1}{\alpha b} t^{-1-\frac{1}{\alpha}} \gamma(1+\frac{1}{\alpha}, b^{\alpha} t). \tag{16}$$

Proof: When $\rho(w)=I_{[0,b]}(w)$, (5) becomes

$$v(t) = \frac{1}{b} \int_0^b e^{-w^{\alpha} t} w^{\alpha} dw. \tag{17}$$

Let $x = w^{\alpha}t$, and the solution is

$$v(t) = \frac{1}{\alpha b} t^{-1-\frac{1}{\alpha}} \gamma(1+\frac{1}{\alpha}, b^{\alpha} t). \tag{18}$$

Thus the proposition is proved. □

When $\alpha = 0$, the infection rate degrades to a fixed number for all individuals in the population and the propagation velocity will drop exponentially. When $\alpha > 0$, with the increase of α, the density of infection rate near 0 will correspondingly increases. Especially, if $\alpha > 1$, the density of infection rate at 0 becomes infinite. This can be intuitively explained as follows. For a population of some topic, the larger the parameter α is, the larger percent of inactive individuals who prefer posting in late stage of the propagation exists. More inactive individuals (α is larger) will cause the propagation velocity drops slower (the exponent -1-1/α is larger). This is also reflected in the analytical solution of $v(t)$ in (18).

3.3 Comparison with Other Epidemic Models

In this subsection, we compare the proposed model with the classical model SI model on networks in terms of the power-law decay feature of propagation velocity. Three models are selected as the underlying networks in the classical SI models: exponential networks, scale-free networks and weighted networks.

For the SI model on exponential networks, the propagation velocity is ([6])

$$v(t) = \frac{di(t)}{dt} = \lambda \langle k \rangle i_0 (1-i_0) \frac{e^{-\lambda \langle k \rangle t}}{[(1-i_0)e^{-\lambda \langle k \rangle t} + i_0]^2}, \tag{19}$$

where λ is the spreading rate of the disease, and $\langle k \rangle$ is the average number of neighbors of each nodes in the network. From the solution of $v(t)$, it can be seen that the tail of $v(t)$ drops exponentially.

For the SI model on scale-free networks, the explicit expression of $v(t)$ can not be obtained. When t is large, the propagation velocity can be approximated by

$$v(t) \approx \int_1^{\infty} [\rho(k) \lambda k [1-i_k(t)](1-1/\langle k \rangle) dk. \tag{20}$$

where $i_k(t)$ is the average densities of infected vertices of degree k, λ is also the spreading rate, and $\Theta_k(t)$ is the density of infected neighbors of a vertex of degree. Using the same calculation technique in Section 3, $v(t)$ in (20) will drop exponentially since the lower limit of the integral is not zero but a positive real number.

For the SI model on weighted scale-free networks in [7], the tail of $v(t)$ is explored in an intuitive way for lack of dynamic equations. Actually, given a statical underlying network, if all infection rates of individuals in the population are larger than a positive value, the velocity will finally drop with exponentially. The smallest infection rate in [7] is $(w_0/w_M)^{\alpha}$, where α is a positive constant, w_M is the largest value of weights of edges in the network and w_0 is a fixed value assigned for the weight of new edges. When the network grows, w_M will goes to an infinite value and the smallest infection rate goes to zero, thus the propagation velocity on the weighted network will drop with power-law. However, for a finite network, the w_M is a finite value and the smallest infection rate may be far from 0, and the propagation velocity will definitely drop exponentially finally.

It can be seen from the comparison above that the SI models on the three types of networks either cannot emulate the power-law decay or have to assume the size of network is infinite.

4 Parameter Estimation

In this section, the techniques which we use to get the real data are introduced in subsection 4.1. Then the preliminary analysis of the real data is presented in subsection 4.2. In the last subsection, we put forward the parameter estimation method by the real data base on the analysis in 3.2.

4.1 Comparison with Other Epidemic Models

We crawled the blog posts from one of the most popular blog service providers, Sina Blog (http://blog.sina.com.cn). Four hot incidental topics, three entertainment events (Event 1~3) and a political event (Event 4) are collected for case studies. A post is considered to be related to a topic if it contains the key words of the topic. A crawler script is written to get all the posts in Sina Blog. The crawler first sent request to Baidu Search Engine for all the posts which contained the key words of the topic, then collected the result pages and isolated the address of the posts. By adding the string 'site:blog.sina.com.cn' to the keys, the results are limited in Sina Blog Space. Then another thread of the crawler downloads the posts by the obtained addresses and gets the detailed information of the posts such as the publish time and blog author's ID.

In this section, we focus on the data of the first 30 days after "peak time" of propagation velocity. The period of 30 days is a proper interval for propagation modeling. The peak time appeared in the same day as the incident for three topics in the third day for one topic. The total post numbers of the four topics in the 30 days after their peaks are 3304, 2866, 550 and 11450 respectively. By spot-checking, the accuracy of the collected posts are 93%, 99%, 95% and 99%. We then handled the data by removing duplicate posts which were published by the same authors, leaving 2945, 2687, 516 and 10295 posts of each topic.

4.2 Preliminary Analysis

Consider a day as the time unit, and the propagation velocity in real system $V_r(t)$, t = 1, 2, …, can be calculated by the following difference equation

$$V_r(t) = I_r(t) - I_r(t-1), \tag{21}$$

where $I_r(t)$ is the total infected individuals, i.e., the number of all posts related to the topic before t. We plot $V_r(t)$ for the four topics in log-log plane, and we can easily find the feature of power-law decay of $V_r(t)$ as the curves drop along some lines in the log-log plane.

4.3 Parameter Estimation

We use the solution of the model in 3.2 to fit the data. The number of infected individuals $I_r(t)$ at each day can be calculated from the real data, but the density is unknown since the total number cannot be obtained from the first 30 days data. Thus we use the differential of the number of infected individuals $V(t)$ to fit the real data. As defined in 2.1 $V(t)$=$Nv(t)$. Here, we use the instantaneously value of the model $V(t)$ to approximate the averaged value $V_r(t)$ calculated from real data over (t-1, t].

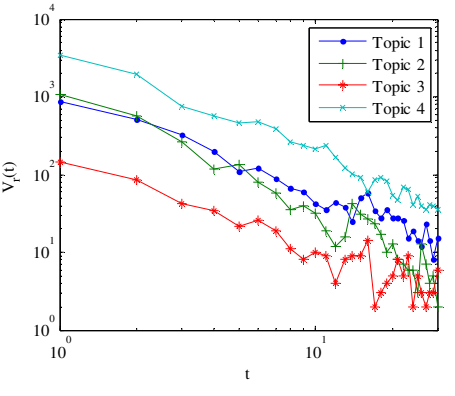

Fig. 1. Propagation velocity of incidental Topics 1~4

Thus the unknown parameters are α in infection rate, b in fitness distribution and the size of the network N. The purpose of fitting procedure is to find a triple of parameters which minimizes the mean square error $J(\boldsymbol{\beta})$ between the observed and expected daily infected individuals over a period of T days, namely

$$J(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T}(V(\boldsymbol{\beta}, t) - V_r(t))^2. \tag{22}$$

where $\boldsymbol{\beta}$=[b, α, N], $V(\boldsymbol{\beta}, t)$ is the output of our model given the parameter $\boldsymbol{\beta}$, $V_r(t)$ is the number of individuals infected in t^{th} days calculated from the real data. Using the nonlinear least-squares regression toolbox in Matlab, we find the optimal parameters for the four topics as follows

Table 1. Optimal parameters

Topic	$\hat{\alpha}$	$\hat{b}$	$\hat{N}$	$J(\hat{\boldsymbol{\beta}})/(\sum_{t=1}^{T} V_r(t))$
1	1.96	1.10	3.75×10^{3}	2.00×10^{5}
2	0.98	1.35	3.81×10^{3}	3.36×10^{5}
3	3.63	1.15	8.43×10^{2}	4.05×10^{5}
4	2.16	1.24	1.41×10^{4}	5.39×10^{5}

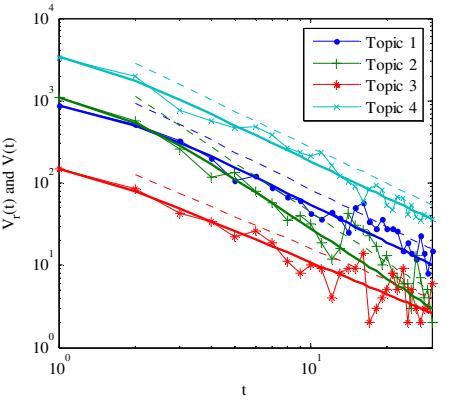

Fig. 2. Comparison of $V(t)$(thick lines), $V_r(t)$ (thin lines with marks) and the lines which exactly drop with power-law(dotted lines)

In Fig. 2, the solution of $V(t)$ with optimal parameters and $V_r(t)$ calculated by the real data are plotted. The lines (dotted lines) which exactly decay with power-law with corresponding exponent -1-1/$\hat{a}$ for each topic are also plotted. Most surprisingly, the propagation velocity drops along the lines from the 2nd day to the 30th day after the peak time of propagation velocity, and this indicates that the power-law decay behavior is an important feature in the propagation of an incidental topic in blogosphere.

As analyzed in section 3, especially from Equation (18), the three parameters in $\boldsymbol{\beta}$ have different effects in the fitting the curve: α is associated with the exponent of the $V(t)$ in log-log plane; b is related to when the curve starts power-law decay, the larger b is, the earlier $V(t)$ starts power-law decay ; N mainly determines the offset of $V(t)$, the curve drifts up while N increasing. The fact that the exponents are the same as the solution in (18) indicates the power-law decay emerges very early in the propagation process, and its study really makes sense.

5 Summary

In summary, we studied the temporal propagation behaviors of the incidental topics in blogoshpere. The individual fitness is introduced to represent the individual's inherent characteristics on the reaction to the topic. The main result is that the power-law

decay of propagation velocity is associated with the existence of the individuals with very low fitness. The power-law exponent of propagation velocity is related to the parameter in the infection rate definition. The analytical solution of $v(t)$ is obtained with the uniform fitness distribution. The solution facilitates the process of determining the parameters. By analyzing the relationship between the fitness distribution and the infection rate distribution, we obtain the intuitive understanding on the power-law decay. By comparing with the classic SI epidemic models, we find our model can model the power-law decay well and precisely fits the actual data with the optimal parameters. It is also observed that the power-law decay appear very early in the propagation process.

In the next step, the propagation characteristics will be investigated to predict the trend based on the actual data obtained in the early propagation stage. Moreover, the concept of node fitness can be integrated with networks when the topics can only spread among related people.

References

1. China Internet Network Information Center, Research Report of 2007 China Blog Market, Statistical Report (December 2007) (in Chinese), http://www.cnnic.net.cn/html/Dir/2007/12/26/4948.htm
2. China Internet Network Information Center, Research Report of 2006 China Blog, Statistical Report (October 2006) (in Chinese), http://www.cnnic.net.cn/html/Dir/2006/09/25/4176.htm
3. Pastor-Satorras, R., Vespignani, A.: Epidemic dynamics and endemic states in complex networks. Physical Review E 63 (2001)
4. Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Phys. Rev. Lett. 86 (2001)
5. Pastor-Satomas, R., Vespignani, A.: Epidemic dynamics in finite scale-free networks. Physical Review E 65 (2002)
6. Barthélemy, M., Barrat, A., Pastoras-Satorras, R., Vespignani, A.: Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. Phys. Rev. Lett. 92 (2004)
7. Yan, G., Zhou, T., Wang, J., Fu, Z.Q., Wang, B.H.: Epidemic spread in weighted scale-free networks. Chin. Phys. Lett. 22, 510 (2005)
8. Adar, E., Zhang, L., Adamic, L.A., Lukose, R.M.: Implicit structure and the dynamics of blogspace. In: Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference (2004)
9. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs. In: SIAM International Conference on Data Mining, SDM 2007 (2007)