Modeling Time-Sensitive Information Diffusion in Online Social Networks

Xin Xu   Xin Chen   Do Young Eun

Abstract—**After a piece of information is released in Online Social Networks (OSNs), will it spread to the entire network or reach only a small population of users? In a time window of interest, how many users will forward or comment on this information? Limited effort has been made at this point to develop an effective model to address these issues, as the time-sensitive nature of information spreading and the complexity of network structure make it a very challenging task. In this paper, we propose a continuous-time model for information diffusion with time-varying diffusion (infection) rate to address these issues, and provide an interface between our proposed model and the well-studied SI model with constant diffusion rate. We prove that there exists an elegant time-rescaling relationship between these two cases, such that any available result on the standard SI model can readily carry over to our time-varying case. We then show how the shape of the time-dependent infection rate will influence the temporal evolution of the size of infection and the time until the information reaches a given node on a graph. This also explains why some information stops spreading before reaching the entire population. Simulation results on Digg graph validate our findings.**

I. Introduction

Recently, Online Social Networks (OSNs) such as Facebook, Twitter, Digg and Microblog have exploded in popularity and drawn much attention from the research community. They offer a unique information sharing mechanism, which allows users to forward information like news articles, public opinions, videos, photos, etc. to their friends, and thus possibly to a wider audience. The convenient interaction and personalized feature of this mechanism makes the form of public information dissemination undergo a significant structural transformation. Under such circumstances, understanding/modeling the dynamics of information diffusion over OSNs has become an important research problem. The applications of modeling this process include locating the most influential users for commercial purpose, finding the source of malicious information and evaluating the social influence of some political and social events, among others.

The research on information diffusion originates from the study on the well-defined epidemic spreading or rumor spreading [14]. However, many recent measurement studies have revealed several unique characteristics of information diffusion in OSNs. For example, in [15], Lerman and Ghosh discuss the effect of topological structure on information spreading, and show that a specific message usually reaches less than 0.1% of the entire population in reality. In [13], the authors quantitatively evaluate different kinds of information spreading online by analyzing the ways hashtags spread, and find significant variations over different topics. [19] outlines a number of empirical findings on competition among memes in Twitter, and points out their massive heterogeneity in popularity and persistence. These findings tell us that the pattern and dynamic of information diffusion over OSNs, albeit bearing some similarity with the epidemic spreading, are much more complex and thus cannot be described under the same mechanism.

In terms of modeling information diffusion over OSNs, most of the existing works have relied on Independent Information Cascade [7] and Linear Threshold models [8]. Further studies in [5], [9], [18] have also extended these models. However, almost all of these assume time-invariant spreading speed and mainly concern if the statistical properties obtained from the model, with parameters appropriately adjusted, would match the empirical observation. But in reality, information diffusion on OSNs is usually affected by multiple factors and highly time-sensitive [4], [17]. Taking these factors into consideration for tractable analysis, however, is a nontrivial task. The state-of-art literature, even for the time-homogeneous case, have to rely on some approximations (e.g. mean-field approach) to estimate its performance except for few special cases, or focus on the final prevalence of epidemics [6], [12].

Related Work: Among the limited recent efforts, [16] adopts the classical branching process to simplify the directed Twitter graph and incorporates a killing process in the end so as to explain the phenomenon that information over OSNs only reaches a small population. While the model therein, to some extent, is able to mimic the way information stops spreading, this is mainly due to careful parameter tuning for the exogenous random time at which the spreading stops, independently of all other factors. Further, the assumption of sudden termination of the information diffusion process over all nodes is not consistent with the reality. [17] proposes a Diffusive Logistic (DL) model to take into account both temporal and spatial factors in the model of information diffusion over OSNs. But again this model is mainly built upon mean field type of approximation as in classical epidemic modeling [14], and thus cannot capture the effect of general network topology on the information diffusion process.

Our Contributions: In this paper, we take a step towards establishing a new time-varying information diffusion model

The authors are with Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911, Email: {xxu5, xchen, dyeun}@ncsu.edu

This work was supported in part by National Science Foundation under Grant No. CNS-1217341 and CNS-1423151.

to fill this gap. We explore the relationship between our time-varying information diffusion model and the well-known SI model with unit diffusion rate ("standard model"), and theoretically show that there exists a simple time-rescaling mapping between these two processes. This finding, to some degree, helps separate the time-varying nature from the influence of complex network topology on the time-varying process, and implies that any available result for the standard process can be easily carried over to our time-varying case with appropriate time transformation. We further discuss the impact of the shape of infection rate β_t on the evolution of the size of infection and time till a given node gets infected. Our simulation results demonstrate the time sensitive nature of information diffusion over Digg graph by reproducing similar information diffusion dynamics using appropriately rescaled time-varying infection rate. We also observe that the shape of β_t for different stories are similar in nature, all displaying piecewise power-law decaying patterns $\alpha t^{-\gamma}$ with $\gamma > 1$ and $\alpha > 0$ for large t. With these modeling for β_t, our theory well matches with the trace in the sense that the information does not go pandemic.

To the best of our knowledge, this is the first work to establish a theoretical framework under which the impact of the shape of infection rate on the information diffusion dynamics is discussed. In addition, our finding on the time transformation relationship from the standard case readily provides a convenient shortcut to analyze the time-dependent information diffusion dynamics by harnessing any available results on the standard SI model on a general graph.

II. System model

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a connected network with a set of finite nodes $\mathcal{N} = \{1, 2, \dots, n\}$ and a set of links $\mathcal{E}$. In this paper, we assume that the evolution of the network structure is much slower compared with the speed of information spreading, and thus can be neglected. If node j lists node i as a friend, then i's interface allows node j to access the messages that node i posts or forwards, as well as other activities associate with i, but *not vice versa*. Then there is a directed link pointing from i to j such that information can flow from i to j, denoted by $(i, j) \in \mathcal{E}$. Note that this friend relationship is asymmetric. We assume that $\mathcal{G}$ has no self-loops and no multiple links between any two nodes.

For a given topic or a piece of message/information in the network, we say a node i is infected if i either initiates this message or forwards this message from its infected neighbors/friends; otherwise, it is considered as uninfected. We then model the diffusion of this information over $\mathcal{G}$ using a process $\mathbf{S}(t) = (S_1(t), S_2(t), \cdots, S_n(t)) \in \{0, 1\}^n$, where $S_i(t) = 1$ if i has been infected by time t and $S_i(t) = 0$ otherwise [1], [6]. Let $|\mathbf{S}(t)| = \sum_{j\in\mathcal{N}} S_j(t)$ be the size of the infected node set (or simply the number of infected nodes) at time t, and $\mathbf{S}(0)$ be the initial set of source nodes. To keep the notation simple, we will also use $\mathbf{S}(t)$ to represent the set of infected notes at time t, i.e., $\{i \in \mathcal{N} \mid S_i(t) = 1\}$, whenever no confusion arises. We allow that the diffusion starts from a single user ($|\mathbf{S}(0)| = 1$) or a connected initial component ($|\mathbf{S}(0)| > 1$). Clearly, all the infected nodes remain connected at any time $t > 0$. Let $N(\mathbf{S}(t)) = \{j \in \mathcal{N} \setminus \mathbf{S}(t) \mid \exists (i, j) \in \mathcal{E}, i \in \mathbf{S}(t)\}$ be the set of 'neighbors' of the infected nodes at time t, and $\partial(\mathbf{S}(t), j) = \{(i, j) \in \mathcal{E} \mid i \in \mathbf{S}(t), j \in N(\mathbf{S}(t))\}$ be the set of edges originating from $\mathbf{S}(t)$ to the neighboring node j.

If $S_i(t) = 1$, then all nodes who list i as a friend are exposed to her message, and are willing to forward the message with *time-varying* rate $\beta_t \geq 0$ because of the influence of node i. Here, β_t captures people's changing enthusiasm to forward the message depending on how old the message is. For a fresh news/message (small t), a user may be more willing to share it with her friends (followers) on her personal page, while she loses her interest in doing so for not-so-fresh message (e.g., smaller β_t for large t). In this setting, at time t, a node j will be infected with rate β_t multiplied by the number of its infected friends. That is,

$$\lim_{\Delta\to 0} \frac{1}{\Delta} \mathbb{E}\{S_j(t+\Delta) - S_j(t) \mid \mathbf{S}(t)\} = \beta_t |\partial(\mathbf{S}(t), j)| \quad (1)$$

if $j \in N(\mathbf{S}(t))$, and zero otherwise, where $|\partial(\mathbf{S}(t), j)|$ is the number of edges from the set $\mathbf{S}(t)$ to the neighboring node j. This model can be considered as the well-known Susceptible-Infected (SI) model on a graph, but with time-dependent infection rate. Here, we use SI model instead of SIR or SIS [3], [6] since our interest is on the temporal dynamics of $\mathbf{S}(t)$ for a given message over time t. Thus, we do not take into account message removal from user's personal page or a user being reinfected by the same message.

Our construction above makes $\{\mathbf{S}(t)\}_{t\geq 0} \in \Omega$ a time-inhomogeneous continuous-time Markov Chain, where $\Omega = \{s_1, s_2, \dots, s_{|\Omega|}\} \subset \{0, 1\}^n$ consists of 2^n possible states recording whether or not each node is infected. We define

$$\mathbb{P}(\mathbf{S}(t') = s' | \mathbf{S}(t) = s) \triangleq p_{s,s'}(t, t'), \ \forall s, s' \in \Omega,$$

for $t > t'$, as the transition probability from state s to s', and $\mathbf{P}(t, t') \triangleq [p_{s,s'}(t, t')]_{s,s'\in\Omega} \in \mathbb{R}^{|\Omega|\times|\Omega|}$ as its transition probability matrix. Let

$$R_{s,s'}(t) = \lim_{\Delta\to 0} \frac{1}{\Delta} p_{s,s'}(t, t+\Delta) \quad (2)$$

be the time-dependent transition rate from state s to s' at time $t > 0$. Note that this transition is possible only when the state s' has one more infected node than s, i.e., $s' = s \cup j$ for some $j \in N(\mathbf{S}(t))$, in which case $R_{s,s'}(t)$ is given by the expression in (1). We call all such possible next states s' as 'follow-up' states from s. Clearly, $R_{s,s'}(t) = 0$ if s' is not one of the follow-up states from s. Then, the infinitesimal generator matrix $\mathbf{Q}(t) = [Q_{s,s'}(t)]_{s,s'\in\Omega}$ is given by $Q_{s,s'}(t) = R_{s,s'}(t)$ for $s \neq s'$, with its diagonal entries $Q_{s,s}(t) = -\sum_{s'\in\Omega, s'\neq s} Q_{s,s'}(t)$. Let $\pi_s(t) \triangleq \mathbb{P}\{\mathbf{S}(t) = s\}$, then the row vector $\boldsymbol{\pi}(t) \triangleq (\pi_{s_1}(t), \pi_{s_2}(t), \cdots, \pi_{s_{|\Omega|}}(t))$ records the probability distribution of all possible states at time t. See Figure 1 for illustration.

Specifically, if $\beta_t = \beta$ is a constant, then our model degenerates to the SI model with constant infection rate on

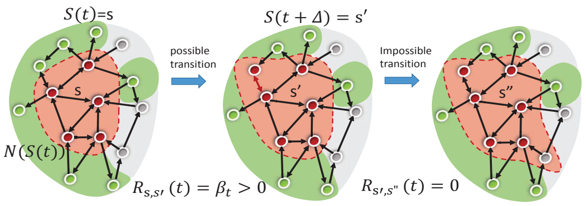


Fig. 1. Information diffusion over $\mathcal{G}$: red nodes are infected; green nodes are in $N(\mathbf{S}(t))$; gray nodes are outside of $\mathbf{S}(t)\cup N(\mathbf{S}(t))$. Here, s' is a possible follow-up state from s, but $s'\to s''$ is not a possible transition.

a finite graph [6], [12]. This time-homogeneous diffusion process has been extensively studied in the literature. We call the process with $\beta_t=1$ for all t as "standard process" in this paper and reserve the notations $\tilde{\mathbf{S}}(t)$, $\tilde{\mathbf{P}}(t,t')$, $\tilde{\mathbf{Q}}$, and $\tilde{\boldsymbol{\pi}}(t)$ to denote its corresponding infected set, transition probability matrix, infinitesimal generator, and the probability distribution, respectively.

III. Dynamics of Information Diffusion

A. From standard SI model to time-varying information diffusion over OSNs

In our continuous-time diffusion model proposed in Section II, the infection rate β_t captures the time-sensitive nature of the information diffusion process over OSNs, which reflects users$'$ time-dependent interest on a given message. Even for the standard process with $\beta_t=1$, $\forall t\ge 0$, the temporal stochastic dynamics of the process $\mathbf{S}(t)$ is intricately dependent on the local neighborhood structure of the currently infected set at time t and on the whole graph structure in the end. Characterizing the distribution of $\mathbf{S}(t)$ (or $\tilde{\mathbf{S}}(t)$ for $\beta_t=1$) over time t on a general graph representing the OSN largely remains elusive and clearly beyond the scope of this paper.[*] Instead, our focus is on the modeling of the time-varying infection rate β_t for information diffusion over OSNs, and integrating it into a general interfacing framework in which we can 'import' available results from the case with $\beta_t=1$ in the standard literature and translate them into a more realistic setting with time-varying infection rate on a general graph.

Our main result in this section is as follows.

Theorem 1: Let $m(t)=\int_0^t \beta_s ds$. Consider the standard process $\tilde{\mathbf{S}}(t)$ with $\beta_t=1$ (i.e, $m(t)=t$) for all t. If $\mathbf{S}(0)=\tilde{\mathbf{S}}(0)$, then $\{\mathbf{S}(t)\}_{t\ge 0}\overset{d}{=}\{\tilde{\mathbf{S}}(m(t))\}_{t\ge 0}$. In particular, for any increasing sequence $0<t_1<t_2<\cdots<t_r$, $(\mathbf{S}(t_1),\mathbf{S}(t_2),\dots,\mathbf{S}(t_r))$ have the same joint distribution as $\left(\tilde{\mathbf{S}}(m(t_1)),\tilde{\mathbf{S}}(m(t_2)),\dots,\tilde{\mathbf{S}}(m(t_r))\right)$ □

Proof. First, we observe that, from (1), for any $t>0$ and from any state $s\in\Omega$ to any other possible follow-up state $s'\in\Omega$, the time-dependent transition rate in (2) becomes $R_{s,s'}(t)=\beta_t\tilde{R}_{s,s'}$, where $\tilde{R}_{s,s'}$ is the transition rate for the standard process with $\beta_t=1$. Since replacing β_t with any other function in (1) does not alter s' being one of follow-up states or not, it follows that $R_{s,s'}(t)=\beta_t\tilde{R}_{s,s'}$ for all $s\neq s'\in\Omega$, and therefore

$$\mathbf{Q}(t)=\beta_t\tilde{\mathbf{Q}},\ \forall t\ge 0. \tag{3}$$

From Kolmogorov's forward differential equation for time-inhomogeneous Markov chain [10], for any $t<t'$, we have

$$\frac{d}{dt'}\mathbf{P}(t,t')=\mathbf{P}(t,t')\mathbf{Q}(t')=\beta_{t'}\mathbf{P}(t,t')\tilde{\mathbf{Q}}, \tag{4}$$

where the second equality is from (3). The only solution to this differential equation is

$$\mathbf{P}(t,t')=\exp\left(\tilde{\mathbf{Q}}\int_t^{t'}\beta_s ds\right)=\exp(\tilde{\mathbf{Q}}(m(t')-m(t))). \tag{5}$$

(Note that if $\beta_t=1$, we would have $\tilde{\mathbf{P}}(t,t')=e^{\tilde{\mathbf{Q}}(t'-t)}$ as in [2].) In particular, (s,s')-th entry in $\mathbf{P}(t,t')$ is given by

$$\begin{aligned} p_{s,s'}(t,t') &= \sum_{k=0}^{\infty}\frac{1}{k!}\left[\tilde{\mathbf{Q}}\left(m(t')-m(t)\right)\right]^k_{s,s'} \\ &= \tilde{p}_{s,s'}(m(t),m(t')), \end{aligned} \tag{6}$$

where the first equality follows from (5).

Now, for any arbitrarily given time sequence $0=t_0<t_1<t_2<\cdots<t_r$, from (6) and $\mathbf{S}(0)=\tilde{\mathbf{S}}(0)$, and by making use of the Markov property, we have

$$\begin{aligned} &\mathbb{P}\{\mathbf{S}(t_1)=s_1,\mathbf{S}(t_2)=s_2,\cdots,\mathbf{S}(t_r)=s_r\mid \mathbf{S}(0)=s_0\} \\ &=\prod_{k=1}^{r}p_{s_{k-1},s_k}(t_{k-1},t_k)=\prod_{k=1}^{r}\tilde{p}_{s_{k-1},s_k}\left(m(t_{k-1}),m(t_k)\right) \\ &=\mathbb{P}\left\{\tilde{\mathbf{S}}(m(t_1))=s_1,\cdots,\tilde{\mathbf{S}}(m(t_r))=s_r\mid \tilde{\mathbf{S}}(0)=s_0\right\}. \end{aligned}$$

Thus, the entire joint distribution of the process $\mathbf{S}(t)$ is the same as that of $\tilde{\mathbf{S}}(m(t))$. This completes the proof. ■

Theorem 1 says, if starting from the same initial set of infected nodes, the whole process $\mathbf{S}(t)$ becomes indistinguishable in distribution from the standard process with its time-axis 'rescaled' by $m(t)$. This implies that the properties of the standard SI model on a finite graph can be inherited by the time-varying case through rescaling the time axis.

To further demonstrate the utility of this transformation, we consider the sequence of newly infected nodes over time and the time instants of such infection. Specifically, let

$$T_k\triangleq\inf\{t>0\mid |\mathbf{S}(t)|-|\mathbf{S}(0)|=k\},\ \ k=0,1,2,\dots, \tag{7}$$

be the time instants by which k additional nodes have been infected on a given graph $\mathcal{G}$ and let $I_k\in\mathcal{N}$ be this k-th (additionally) infected node. Let τ_j be the time at which node $j\in\mathcal{N}$ becomes infected. Clearly, $T_0=0$ and $\tau_j=0$ for $j\in\mathbf{S}(0)$. We define $\tilde{T}_k,\tilde{I}_k$, and $\tilde{\tau}_j$ similarly for the standard process $\tilde{\mathbf{S}}(t)$ starting from the same initially infected nodes $\mathbf{S}(0)=\tilde{\mathbf{S}}(0)$. Then, we have the following.

Proposition 1: If $\mathbf{S}(0)=\tilde{\mathbf{S}}(0)$, then $\mathbb{P}\{T_k>t\}=\mathbb{P}\{\tilde{T}_k>m(t)\}$ and $\mathbb{P}\{\tau_j>t\}=\mathbb{P}\{\tilde{\tau}_j>m(t)\}$ for any $k\ge 1, j\in\mathcal{N}$. □

Proof. The first assertion follows directly from Theorem 1 and by noting that

$$\mathbb{P}\{T_k\le t\}=\mathbb{P}\{|\mathbf{S}(t)|-|\mathbf{S}(0)|\ge k\}$$

[*] Note that the Markov chain $\tilde{\mathbf{S}}(t)$ is transient, starting from $\tilde{\mathbf{S}}(0)$ with absorbing state in which every node is infected.

$$= \mathbb{P}\{|\tilde{\mathbf{S}}(m(t))| - |\tilde{\mathbf{S}}(0)| \geq k\} = \mathbb{P}\{\tilde{T}_k \leq m(t)\}.$$

Similarly, the second assertion follows by noting $\mathbb{P}\{\tau_j > t\} = \mathbb{P}\{j \notin \mathbf{S}(u), \forall u \in [0,t]\}$ and by applying Theorem 1. ■

From Theorem 1, it is straightforward to note that the sequence of newly infected nodes over time in our time-varying model $(I_1, I_2, \ldots)$ have the same joint distribution as that of the standard process $(\tilde{I}_1, \tilde{I}_2, \ldots)$. Here, for a given set of infected nodes $\mathbf{S}(t)$ at time t, the probability of a neighboring node $j \in N(\mathbf{S}(t))$ being the next infected node is proportional to the number of edges emanating from $\mathbf{S}(t)$ to the node j, i.e., it is given by $|\partial(\mathbf{S}(t), j)| / \sum_{l \in N(\mathbf{S}(t))} |\partial(\mathbf{S}(t), l)|$. This suggests that the embedded Markov chain (observed at T_k) evolves in the same way (in distribution) for both time-varying β_t and the standard process case, while the embedding time sequences $\{T_k\}$, and $\{\tau_j\}$ are again 'rescaled' as in Proposition 1.

B. Impact of the shape of β_t on information diffusion

In this subsection, we show how the shape of the time-varying infection rate β_t (or equivalently $m(t) = \int_0^t \beta_s ds$) impacts the temporal dynamics of the information diffusion, such as the expected size of the epidemics over time ($\mathbb{E}\{|\mathbf{S}(t)|\}$). In particular, we investigate under what condition on the shape of β_t the information stops diffusing without going pandemic, and study the relationship between specific forms of the decreasing infection rate β_t and time till infection (τ_j) of a node j in a general graph. We then discuss the versatility of our framework and compare with other existing literature. We assume $\mathbf{S}(0) = \tilde{\mathbf{S}}(0)$ throughout this section to avoid unnecessary repetition.

First, from Theorem 1, the expected size of all the infected nodes at time t can be written as

$$\mathbb{E}\{|\mathbf{S}(t)|\} = \mathbb{E}\left\{\left|\tilde{\mathbf{S}}(m(t))\right|\right\}. \tag{8}$$

This relationship proves to be very useful later on when we empirically measure β_t from a real data set showing how many nodes are infected over time. (That is, we have $\mathbb{E}\{|\mathbf{S}(t)|\}$ from real trace.) Once we identify the source node (or source set), we can run the standard SI process on the same graph starting from the same source(s) to obtain $\mathbb{E}\{\tilde{\mathbf{S}}(u)\}$. We then only have to find out the corresponding time index t such that $\mathbb{E}\{\mathbf{S}(t)\} = \mathbb{E}\{\tilde{\mathbf{S}}(u)\}$ to get $u = m(t)$ (or $t = m^{-1}(u)$) to recover the time-varying infection rate β_t.

Now, observe that

$$|\mathbf{S}(t)| = \sum_{j \in \mathcal{N}} 1_{\{\tau_j \leq t\}} = n - \sum_{j \in \mathcal{N}} 1_{\{\tau_j > t\}}.$$

By taking expectation and setting $t \to \infty$, the expected final size of the epidemic becomes

$$\begin{aligned} \mathbb{E}\{|\mathbf{S}(\infty)|\} &= n - \sum_{j \in \mathcal{N}} \mathbb{P}\{\tau_j > \infty\} \\ &= n - \sum_{j \in \mathcal{N}} \mathbb{P}\{\tilde{\tau}_j > m(\infty)\} = \mathbb{E}\{|\tilde{\mathbf{S}}(m(\infty))|\}. \end{aligned} \tag{9}$$

This tells us that, if the infection rate β_t decays slowly such that $m(\infty) = \int_0^\infty \beta_s ds = \infty$, then eventually every node will get infected, i.e., the infection goes pandemic. On the other hand, if β_t decays quickly enough so that $m(\infty) = \int_0^\infty \beta_s ds = M < \infty$, then the final size of the infection is less than n, i.e., $\mathbb{E}\{|\mathbf{S}(\infty)|\} < n$, since $\mathbb{P}\{\tilde{\tau}_j > M\} > 0$.[†] In this case, the final size of the infection under time-varying β_t would correspond to the expected size of the infection of the standard SI process stopped at M. In words, if the public interest on a specific topic/message is fading fast enough, the message will stop spreading over the network before it is adopted by all users. This scenario supports the phenomenon that the spreading of a specific topic usually reaches only a small population in reality, instead of going pandemic [15].

We now turn our attention to characterizing τ_j, the time till node $j \notin \mathbf{S}(0)$ gets infected, depending on whether or not β_t is integrable. To proceed, let $|\partial(\mathbf{S})| = |\partial(\mathbf{S}, \mathcal{N} \setminus \mathbf{S})|$ be the number of edges originating from $\mathbf{S}$ to its outside. Our results are then summarized as follows.

Proposition 2: Fix any arbitrary node $j \in \mathcal{N} \setminus \mathbf{S}(0)$. If $m(\infty) < \infty$, then $\mathbb{E}\{\tau_j\} = \infty$. If $\beta_t \sim \alpha t^{-\gamma}$ for $0 < \gamma < 1$ and $\alpha > 0$, then $\mathbb{E}\{\tau_j\} < \infty$. If $\beta_t \sim \alpha/t$, then $\mathbb{E}\{\tau_j\} < \infty$ for $\alpha > 1$, while $\mathbb{E}\{\tau_j\} = \infty$ for $0 < \alpha \leq 1/|\partial(\mathbf{S}(0))|$. □

Proof. Let h be the length of the shortest path from the source(s) $\mathbf{S}(0)$ to the node j. Our key observation here is to note that τ_j can be upper-bounded by considering the infection spreading from $\mathbf{S}(0)$ to node j only along this shortest path of length h, and lower bounded by T_1 – the time until one of the neighbors in $N(\mathbf{S}(0))$ first gets infected. Consider the standard process with $\beta_t = 1$ infection rate. Let $N^\lambda(0,t)$ be the Poisson process with constant rate λ and $Y^\lambda(k)$ be its k-th arrival time instant, i.e., $Y^\lambda(k) \overset{d}{=} \sum_{i=1}^k X_i^\lambda$ where X_i^λ, $i = 1, 2, \ldots$ are *i.i.d.* exponential random variables with rate λ. Then, the preceding arguments yield, for any $u \geq 0$,

$$\mathbb{P}\{\tilde{T}_1 \geq u\} \leq \mathbb{P}\{\tilde{\tau}_j > u\} \leq \mathbb{P}\{Y^1(h) \geq u\}. \tag{10}$$

From Proposition 1, we have

$$\mathbb{E}\{\tau_j\} = \int_0^\infty \mathbb{P}\{\tau_j > t\} dt = \int_0^\infty \mathbb{P}\{\tilde{\tau}_j > m(t)\} dt, \tag{11}$$

and from (10) we can write

$$\int_0^\infty \mathbb{P}\{\tilde{T}_1 \geq m(t)\} dt \leq \mathbb{E}\{\tau_j\} \leq \int_0^\infty \mathbb{P}\{Y^1(h) \geq m(t)\} dt. \tag{12}$$

Under the standard process, $\tilde{\tau}_j$ for $j \in N(\mathbf{S}(0))$ are independent and exponentially distributed with rate $|\partial(\mathbf{S}(0), j)|$. Thus, $\tilde{T}_1 = \min_{j \in N(\mathbf{S}(0))} \{\tilde{\tau}_j\}$ is also exponentially distributed with rate $\sum_{j \in N(\mathbf{S}(0))} |\partial(\mathbf{S}(0), j)| = |\partial(\mathbf{S}(0))|$. From (11) and (12) and by noting $\mathbb{P}\{Y^\lambda(h) \geq u\} = \mathbb{P}\{N^\lambda(0,u) \leq h\}$, we get

$$\int_0^\infty e^{-|\partial(\mathbf{S}(0))| m(t)} dt \leq \mathbb{E}\{\tau_j\} \leq \sum_{i=0}^h \frac{1}{i!} \int_0^\infty e^{-m(t)} (m(t))^i dt. \tag{13}$$

[†] Even for a node j right next to the source, the density of τ_j has exponential tail, giving rise to non-zero probability of $\tau_j > M$ for any finite M. See also the proof of Proposition 2.

Now, if $m(\infty) < \infty$ (and note that $m(t) = \int_0^t \beta_s ds$ is monotonically non-decreasing in t), the integrand in the lower bound in (13) is bounded away from zero, thus $\mathbb{E}\{\tau_j\} = \infty$. When $m(\infty) = \infty$, we can rewrite the term in the upper bound in (13) as

$$\int_0^\infty e^{-m(t)}(m(t))^i dt = \int_0^\infty \frac{e^{-x}x^i}{\beta_{m^{-1}(x)}} dx, \qquad (14)$$

by change of variable with $x = m(t)$.

When $\beta_t \sim \alpha t^{-\gamma}$ with $0 < \gamma < 1, \alpha > 0$ for large t, we have $m(t) \sim \frac{\alpha}{1-\gamma} t^{1-\gamma}$ and $\beta_{m^{-1}(x)} \sim \alpha^{\frac{1}{1-\gamma}}[(1-\gamma)x]^{-\frac{\gamma}{1-\gamma}}$. Thus, the integrand in (14) for large x is always in the form of $e^{-x}x^{i+\gamma/(1-\gamma)}$ and the integral is finite. Similarly, if $\beta_t \sim \alpha/t$ for large t, then $m(t) \sim \alpha \log t$ and $\beta_{m^{-1}(x)} \sim \alpha e^{-x/\alpha}$. Thus, the integrand in (14) for large x now is in the form of $x^i \exp(x(\frac{1}{\alpha} - 1))$. Clearly, if $\alpha > 1$, this integral converges, so we have $\mathbb{E}\{\tau_j\} < \infty$. The integrand in the lower bound in (13) is in the form of $t^{-|\partial(\mathbf{S}(0))|\alpha}$, and the integral diverges when $|\partial(\mathbf{S}(0))|\alpha \leq 1$, yielding $\mathbb{E}\{\tau_j\} = \infty$. This completes the proof. ■

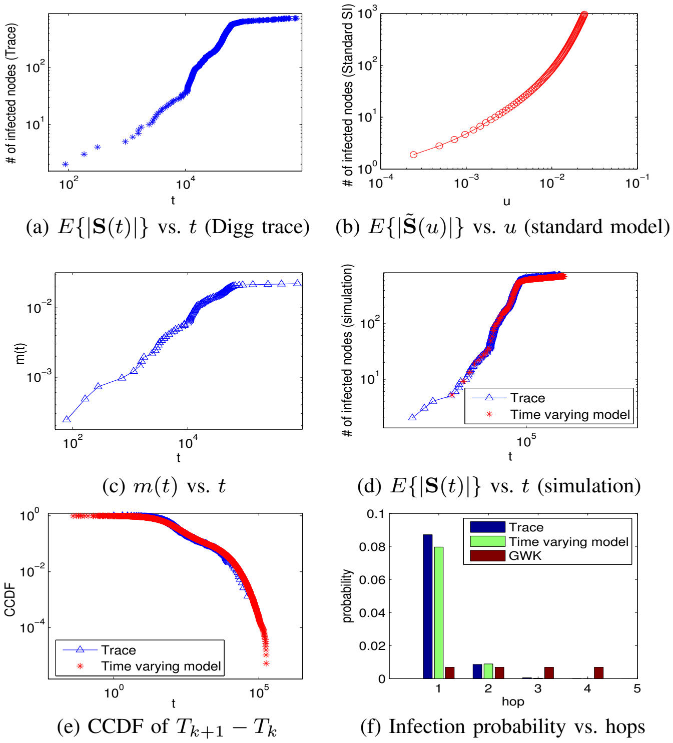
Proposition 2 asserts that when β_t is integrable, the diffusion stops as it takes infinite amount of time to reach any node j on average, and it goes pandemic when β_t decays slowly as $\alpha t^{-\gamma}$ with $0 < \gamma < 1$ and $\alpha > 0$ (thus $m(\infty) = \infty$), as every node will get infected within a finite amount of time on average. The situation is more subtle, however, when β_t decays as α/t for $\alpha \leq 1/|\partial(\mathbf{S}(0)|$. We still have $m(\infty) = \infty$, thus every node will get infected eventually, i.e., $|\mathbf{S}(\infty)| = n$ as expected from Theorem 1. But, this happens with $\mathbb{E}\{\tau_j\} = \infty$, or more precisely, $\mathbb{E}\{T_1\} = \infty$ as T_1 is heavy-tailed with $\mathbb{P}\{T_1 > t\} \sim t^{-|\partial(\mathbf{S}(0))|\alpha}$. A closer look into the proof of Proposition 2 reveals that the same argument goes through for any 'inter-infection' time, i.e., $T_{i+1} - T_i$ is heavy-tailed with infinite mean, as long as the total number of edges out of the currently infected set of nodes ($|\partial(\mathbf{S}(t))|$) remains bounded. This suggests that the shape of β_t largely governs whether or not the infection goes pandemic and the time till infection being heavy-tailed, whereas the exact distribution of τ_j and T_k will clearly depend on the graph structure and the resulting evolution of $\mathbf{S}(t)$ and $N(\mathbf{S}(t))$ over time.

The paper [16] also focused on the behavior that the infection usually reaches a limited number of nodes without going pandemic. They assumed β_t is constant for all generations on a tree graph, and then applied Galton-Watson branching process with modification that the infection might be "killed" at some random hop. We maintain that our framework is far more general and versatile, as our model not only works on any general graph, but also captures any possible time-dependent (thus hop-dependent) infection rate, not necessarily being constant and then turned off to zero.

IV. Simulation Results

In this section, we present numerical results to demonstrate that our continuous-time model with time-varying infection rate can well capture the dynamics of the information diffusion process in reality. Our experiments are performed over the Digg dataset [11]. Digg is a news aggregator website with an editorially driven front page. There are two mechanisms that the information spreads over the underlying Digg graph – through the connected friendship links, or through unconnected users' access to an outer source (like front page or other websites).

The dataset collected by K. Lerman [11] contains trace about 3553 popular stories promoted to Digg's front page over one month in 2009. For each story, they collected its related information including user IDs who voted for the story and the time stamp of each vote. In addition, they also collected the corresponding friendship graph composed of 270,535 distinct users. The first user who initiates a post is considered to be the source of this story. Since we are mainly concerned about the influence of friendship on information diffusion process, in our pre-processing procedure, we extract a "connected subgraph" originating from the source, by tracking only the votes from those who follow the ones that have voted earlier for this story. This way, we can eliminate other factors such as outer source problem, etc.



(a) $E\{|\mathbf{S}(t)|\}$ vs. t (Digg trace)

(b) $E\{|\tilde{\mathbf{S}}(u)|\}$ vs. u (standard model)

(c) $m(t)$ vs. t

(d) $E\{|\mathbf{S}(t)|\}$ vs. t (simulation)

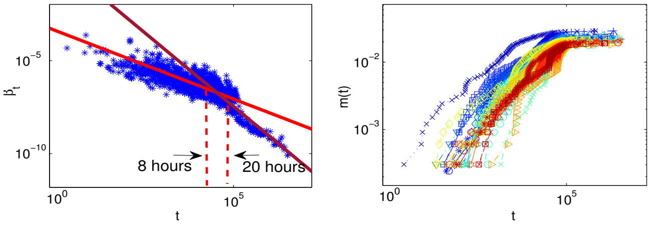
(e) CCDF of $T_{k+1} - T_k$

(f) Infection probability vs. hops

Fig. 2. Measurement and simulation results of story 696.

Due to page limit, we here include only the results of story 696 initiated by source node 129511 as an example. We have observed similar trends for all other stories. Figure 2(a) records the evolution of the number of infected nodes from the trace, where the time axis is in the unit of second. Since our goal is to construct a tractable model that can faithfully reproduce this process, here we consider the plot in Figure 2(a) as $\mathbb{E}\{|\mathbf{S}(t)|\}$. Figure 2(b) shows $\mathbb{E}\{|\tilde{\mathbf{S}}(u)|\}$ of the standard SI process with unit infection rate from the same source node, where each data point in this plot is the average over 1000 simulations. As explained in Section III-B and from (8), we can then derive $m(t)$ by simply letting $\mathbb{E}\{|\mathbf{S}(t)|\} = \mathbb{E}\{|\tilde{\mathbf{S}}(u)|\}$ and finding the corresponding time index pairs $u = m(t)$, which is shown in Figure 2(c). To be more specific, in Figure 2(a), we record the time sequence $\{t_k\}$ when k nodes get infected for all k, while in Figure 2(b), we record the corresponding times $\{u_k\}$. Then plotting (t_k, u_k) in Figure 2(c) will serve the purpose. Once we have extracted $m(t)$ (or equivalently β_t) for a given story, we then simulate our time-dependent information diffusion process starting from the same source on the same graph. Figures 2(d) and (e) show the resulting $\mathbb{E}\{|\mathbf{S}(t)|\}$ and $\mathbb{P}\{T_{k+1} - T_k > t\}$ (CCDF of the sojourn time between two consecutive retweets) over t, respectively, on a log-log scale where each data point is average over 1000 independent simulation runs and also over all k (for plot (d)).

In Figure 2(f), we compare the probability that users get infected at each hop away from the source. Here we also compare our algorithm with the Galton-Watson branching model with random "killing" in [16]. Note that the model in [16] doesn't serve the purpose to depict how the infection set evolves with time, but instead, tries to capture how infection spreads along each hop, and at which hop the information stops spreading. We can see from Figure 2(f) that our model captures the behavior that the nodes closer to the source have much higher probability to be infected than those far away, while the [16] assumes the same infection probabilities for all generations/hops until randomly killed (turned off to zero).



(a) β_t vs t

(b) $m(t)$ vs. t

Fig. 3. Aggregation of β_t and $m(t)$ over 30 most popular stories.

We next repeat our simulations for 30 most popular stories (Each story has more than 400 voters who get to know it through friendship links.), estimate the time-varying infection rate β_t and $m(t)$ as above, and plot their aggregate data on a log-log scale in Figure 3. We observe that β_t decays as approximately a piecewise power-law function of time t. The turning point appears in the time interval $[8, 20]$ hours after the information is first released. This observation precludes the feasibility of describing information diffusion process over OSNs with constant diffusion (infection) rate. Specifically, $\beta_t \sim \alpha t^{-\gamma}$ where the slope is measured to be $\gamma \approx 0.82$ with $\alpha \approx 0.0011$ for the first 10 hours or so, while $\gamma \approx 1.77$ and $\alpha \approx 25.1288$ is a good fit after 10 hours. Since $\gamma > 1$ for large t, it is clearly integrable and $m(\infty)$ is finite, as also shown in Figure 3(b). This clearly explains why the information reaches only a limited number of users in view of Proposition 2. Piecewise power-law shape of the infection rate β_t also suggests that most information tend to spread relatively fast within the first day (or daytime) after it is released, but people will soon lose their interests in following the information afterward, leading to faster decaying infection rate and ultimately stopping the information diffusion.

V. Conclusion

In this paper, we proposed a simple yet versatile diffusion model on a general OSN graph with time-dependent infection rate β_t. We have shown in theory and practice that the shape of β_t is the key factor in determining when and whether a message will reach a set of prescribed nodes. Our general framework and versatile time-rescaling relationship will allow us to reuse any result from well-studied SI model on a graph with constant infection rate, onto more realistic OSN setting in which users quickly lose their interests in spreading the information over time. One possible future work is to further investigate other factors in information diffusion, including spatially heterogeneous infection rate over different users (in addition to temporally decaying infection rate) and quantifying/predicting the size of infection and time till infection in terms of local network structure around the sources.

References

[1] S. Banerjee, A. Gopalan, A. Das, and S. Shakkottai. Epidemic spreading with external agents. *IEEE Trans. Information Theory*, 2014.

[2] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[3] M. Draief and L. Massouli. *Epidemics and rumours in complex networks*. Cambridge University Press, New York, 2010.

[4] N. Du, L. Song, M. G. Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *NIPS*, 2013.

[5] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *WOSN*, 2010.

[6] A. J. Ganesh, L. Massoulie, and D. F. Towsley. The effect of network topology on the spread of epidemics. In *Infocom*, 2005.

[7] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: a complex systems look at the underlying process of word-of mouth. In *Marketing letters*, 2001.

[8] M. Granovetter. Threshold models of collective behavior. In *American journal of sociology*, 1978.

[9] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[10] L. Kleinrock. *Queueing systems: Volume II Computer Applications*. Wiley Interscience, New York, 1976.

[11] K. Lerman. Digg 2009 data set. http://www.isi.edu/~lerman/downloads/digg2009.html.

[12] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[13] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*, 2011.

[14] A. V. Romualdo Pastor-Satorras. Epidemic spreading in scale-free networks. In *Phys. Rev. Lett*, April 2001.

[15] G. V. Steeg, R. Ghosh, and K. Lerman. What stops social epidemics? In *ICWSM*, 2011.

[16] D. Wang, H. Park, G. Xie, S. Moon, M.-A. Kaafar, and K. Salamation. A genealogy of information spreading on microblogs: a galton-watson-based explicative model. In *Infocom*, 2013.

[17] F. Wang, H. Wang, and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In *ICDCSW*, 2012.

[18] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: an eigenvalue viewpoint. In *Proc. IEEE SRDS*, 2003.

[19] V. A. M. F. Weng L, Flammini A. Competition among memes in a world with limited attention. In *Scientific Reports 2*, 2012.