Characterizing Information Diffusion in Online Social Networks with Linear Diffusive Model

Feng Wang*, Haiyan Wang*, Kuai Xu*, Jianhong Wu†, Xiaohua Jia‡
* *School of Mathematical and Natural Sciences, Arizona State University*
Email: {fwang25, haiyan.wang, kuai.xu}@asu.edu
†*Department of Mathematics and Statistics, York University*
Email: wujh@mathstat.yorku.ca
‡*Department of Computer Science, City University of Hong Kong*
Email: csjia@cityu.edu.hk

***Abstract*—Mathematical modeling is an important approach to study information diffusion in online social networks. Prior studies have focused on the modeling of the temporal aspect of information diffusion. A recent effort introduced the spatio-temporal diffusion problem and addressed the problem with a theoretical framework built on the similarity between information propagation in online social networks and biological invasion in ecology [1]. This paper examines the spatio-temporal characteristics in further depth and reveals that there exist regularities in information diffusion in temporal and spatial dimensions. Furthermore, we propose a simpler linear partial differential equation that takes account of the influence of spatial population density and temporal decay of user interests in the information. We validate the proposed linear model with Digg news stories which received more than 3000 votes during June 2009, and show that the model can describe nearly 60% of the news stories with over 80% accuracy. We also use the most popular news story as a case study and find that the linear diffusive model can achieve an accuracy as high as 97.41% for this news story. Finally, we discuss the potential applications of this model towards finding super spreaders and classifying news story into groups.**

***Keywords*-information diffusion, mathematical modeling, PDE, spatio-temporal, online social network.**

I. Introduction

In light of the significant role online social networks have played in recent events and crisis, it is important to study how information travels over such networks. Several papers have studied the characteristics of information diffusion over various online social networks using empirical approaches [2], [3], [4], [5], [6]. These efforts revealed intrinsic patterns of information diffusion dynamics. However, it is indispensable to develop mathematical models in order to formally and quantitatively describe the essential features of the diffusion dynamics. A well-calibrated mathematical model of information diffusion can help predict the popularity of a positive news story before it is published on the web. In addition, we can use mathematical models to simulate the evolution of different systems controlled by their corresponding system parameters.

Modeling information diffusion dynamics in online social networks is challenging due to the intricacy of human dynamics and social interactions, the obscure underlying diffusion network structures, the vast scale of online social networks, and the heterogeneity and diversity of social media. A few recent research efforts use mathematical models, either deterministic or stochastic models, to describe and predict information diffusion in temporal dimension in online social networks [7], [8], [9], [10].

One recent paper proposed the first Partial Differential Equation (PDE) based diffusion model called diffusive logistic model to characterize both temporal and spatial dynamics of information diffusion in online social networks from a macroscopic perspective [1]. Studying the diffusion process from both spatial and temporal dimensions can provide more details in terms of speed and coverage of information diffusion which cannot be addressed by investigating from temporal dimension alone. It can also shed light on the underlying diffusion structures of online social networks. Specifically the diffusive logistic model addresses the *spatio-temporal diffusion problem*: for a given information m initiated from a particular user called *source* s, after a time period t, what is the density of influenced users at distance x from the source? An influenced user is a user who actively votes or likes the information. [1] demonstrates the feasibility of modeling information diffusion in both spatial and temporal dimensions and achieves high accuracy for describing the information diffusion process of the most popular news story in a Digg data set. Although the ability to precisely describe a news story with a large number of votes with a simple model and few parameters is noteworthy, whether there exists a model to precisely describe news stories with different voting scales is left unanswered.

In this paper, we study the same spatio-temporal diffusion problem. Our goal is to investigate three questions as follows: "Does there exist some regularities of information diffusion dynamics through spatial and temporal dimensions in online social networks?", "Can we quantitatively describe the process with a deterministic mathematical model?", and

"How precise is the model?". We first provide empirical results and insights into the spatio-temporal patterns in a Digg data set containing dissemination information of 3553 news stories collected during June 2009. More importantly, we propose a simple linear PDE-based model called Linear Diffusive Model which takes account of the influence of spatial spreading power and temporal news decay on the information diffusion process of a news story. By separating the influence of distance and time, the model can reveal the inherent news decay pattern and provide an approach to search for super spreaders in online social networks. Additionally, we use least square fitting technique to find the parameters of the model to best approximate the data and examine the accuracy of the linear diffusive model for all news stories with more than 3000 votes in the Digg data set. Following is the list of our major findings:

- The empirical study of the Digg data set exhibits strong evidence of the existence of regularities in information diffusion over time and distance. This supports the predictability of spatio-temporal information diffusion dynamics. Specifically, out of 3553 news stories, over 90% of them show consistent diffusion patterns where the density of influenced users at hop i from the submitter of a news story is consistently higher than the density of influenced users at hop $i+1$ over 50 hours (when the news influence saturates) for $i = 1$ and 2;
- The linear diffusive model can achieve better accuracy than the diffusive logistic model. For the most popular news story with 24,099 votes, linear diffusive model achieves an accuracy of 97.41% when predicting its diffusion process over 6 hours after its first submission for users at hops 1 to 5 from the submitter while the accuracy of logistic diffusive model for the same news story over the same period is 92.08%;
- The linear diffusive model can also describe news stories with smaller scales of votes with high precision. For all 133 news stories with more than 3000 votes in Digg, the linear diffusive model achieves higher than 80% accuracy for more than 60% of the news stories when predicting the densities of influenced users at hops 1 to 5 from the news submitters for 20 hours.

The remainder of this paper is organized as follows. Section II presents the empirical analysis of temporal and spatial information diffusion patterns in Digg. Section III introduces the linear diffusive model and describes the construction of the initial density function. The validation of the proposed linear diffusive model is presented in Section IV. Section V discusses the applications of the linear diffusive model. Section VI gives a brief literature review on related work, and Section VII concludes the paper and outlines our future work.

II. Spatio-Temporal characteristics in Digg

In this section, we first briefly introduce the functionality of Digg site, then present the spatial and temporal information diffusion patterns revealed by empirical analysis on the Digg data set.

A. Digg Social News Aggregation Site

Digg is one of the most popular news aggregation sites. Digg users can submit links of news stories that they find in professional news sites and blogs to Digg, and can vote (also called digg in the Digg context) and comment on the submitted news story. Digg users form friendship links through "following" each other. The first user who brings the news story to the Digg site is referred as the submitter or source. Information propagates dynamically in three major channels: 1) friendship relationship. A Digg user can view the news story submitted by the friends he follows, and then digg the news story. After a user diggs the news story, all his followers are able to see and digg the news story, and so on; 2) a news story could be promoted to the front page if it ranks as one of the top news stories based on the number of diggs it received. After a news story is promoted, all Digg users, including those who are not direct friends of the news submitter and voters, can view and vote for the news story; 3) a user can discover and digg news through search functions on Digg. The last two propagation channels introduce *randomness* in information diffusion over online social networks. Due to the complexity of user-to-user interactions, we study information diffusion with deterministic mathematical models from the macroscopic rather than the microscopic perspective.

The Digg data set investigated in this paper was collected in [11]. It consists of 3553 news stories that were digged and promoted to the front page of `www.digg.com` due to vote popularity during June 2009. In total, there are more than 3 millions votes cast on these news stories from $139,409$ Digg users. In addition, the data set also includes the directed *friendship* links among the Digg users who have voted these news stories. Based on these friendship links, we construct a directed social network graph among these Digg users. For each of the news stories, the data set includes the user id of all the voters during the collection period, and the timestamps when votes were cast. The time granularity is in seconds. The availability of the voting history, timestamp and the social network graph give us the opportunity to study the temporal and spatial patterns of information propagation in Digg.

B. Temporal and Spatial Patterns of Information Diffusion

In order to model the information diffusion process in temporal and spatial dimensions, we conduct a series of empirical analysis to discover the information spreading patterns in the Digg dataset. Studying information diffusion in the spatial dimension requires a definition of distance.

A natural distance metric between two users in Digg is the length of the shortest path, measured by the number of hops from one user to another in the social network graph. This distance metric is called *friendship hops* in [1]. With this definition of distance, all users can be divided into distance groups based on their distance from the news submitter. Clearly, the direct followers of the submitter have a distance of 1, while their own direct followers have a distance of 2 from the submitter, and so on. As a news story propagates through the Digg network, users express their interests in the news by voting for it. We call such users as *influenced users* of the information.

1) Information Diffusion Dynamics: In this subsection, we first explore the information diffusion dynamics in four individual news stories $s1$ to $s4$, then investigate whether the observations in the four stories are valid for other news stories in the Digg data set. Story $s1$, $s2$, and $s3$ are the three most popular news stories in the Digg data set with 24,099, 8521, and 8492 votes respectively. Story $s4$ is a news story with 1618 votes. Figure 1 illustrates the densities of influenced users at distances 1 to 5 from the corresponding submitter over 50 hours for these four news stories. Each line represents the density of influenced users at a certain distance. It is clear that the density evolution of these four news stories exhibits consistent patterns as follows: 1) The densities of influenced users at different distances show consistent evolving patterns rather than increase with random fluctuations. Densities of influenced users increase rapidly at the first few hours, then saturate after about 50 hours. This is consistent with the observation in [7] which studies the temporal evolution dynamics in Digg data set collected from July 1, 2007 to December 18, 2007; 2) The density of influenced users at distance 1 (represented by the top line) is significantly higher than that of users at distances greater than 1. It is consistent with the observation in [8] that the friends of the news submitter are more interested in the news than the friends of the voters of the news. This also indicates that friendship link is an important channel of information spreading; 3) For news stories with less votes, the diffusion process starts after some delays. In addition, the densities of influenced users at all distances increase consistently after these delays. For example, for $s4$, after the news story is posted for 14 hours, the densities of influenced users at distance 2 to 5 all start to increase at approximately the same time.

In order to investigate whether the observed evolution patterns hold for other news stories in the Digg data set, we calculate the densities of influenced users at distance 1 to 5 over 50 hours for all news stories. Let $\delta^t_{i,i+1}$ represent the difference of the densities of influenced users between distance i and distance $i+1$ at time t. For each $1 \leq i \leq 4$, we count the number of times that δ is non-negative during 50 hours and call it *consistency counts*, denoted as $\delta_{i,i+1}$. If $\delta_{i,i+1}$ of a news story is 50, it indicates the densities of influenced users at distance i are always greater or equal to the densities of influenced users at distance $i+1$. On the contrary, if $\delta_{i,i+1}$ is 0, it means the densities of influenced users at distance i are always smaller than those at distance $i+1$.

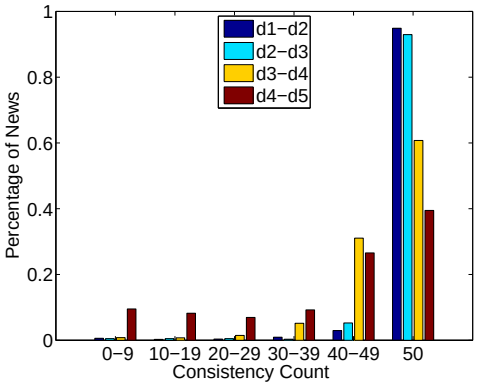
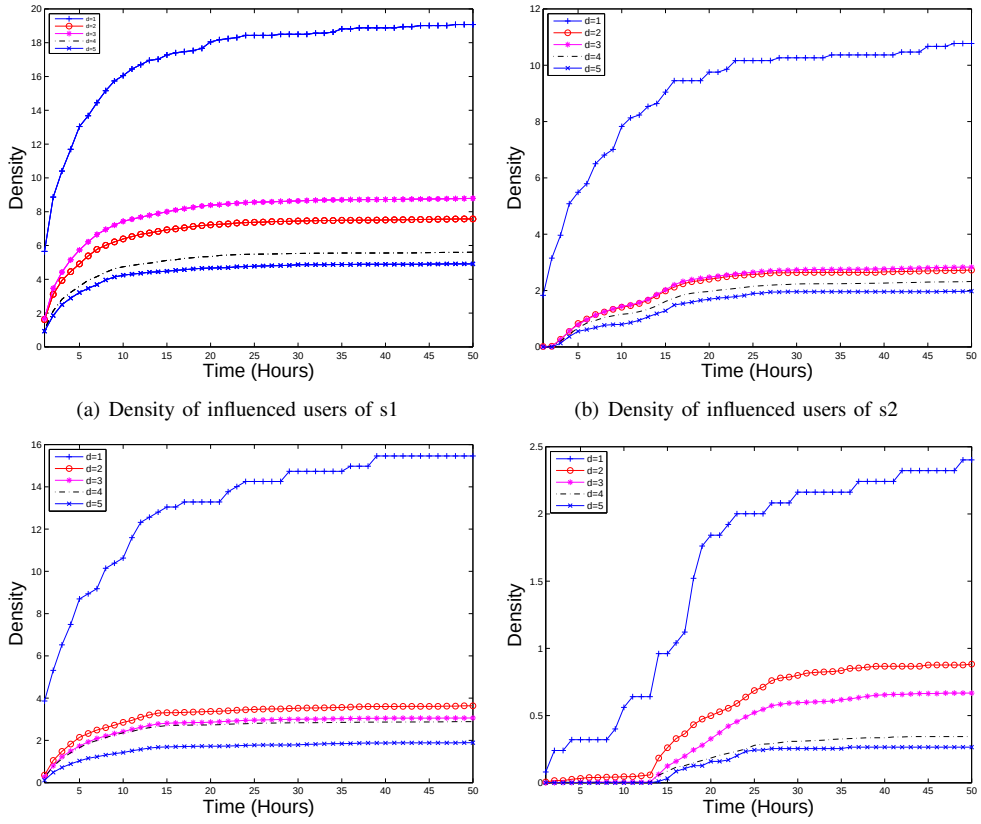


Figure 2. Percentage of consistency counts of 3553 news stories in the Digg data set

Figure 2 shows the percentage of all 3553 news stories in the Digg data set whose consistency counts fall in a certain range. As can be seen, 94.9% of all news stories have consistency counts $\delta_{1,2}$ of 50. This means that for 94.9% of news stories, the densities of their influenced users at distance 1 are always higher than or equal to the densities of their influence users at distance 2. In the figure, 92.9% of news stories have consistency counts $\delta_{2,3}$ of 50. This means 92.9% of news stories have densities of influenced users at distance 2 always higher than or equal to the densities of influence users at distance 3. Even the percentage of news stories with $\delta_{3,4}$ of 50 is relatively low as 60.8%, the combined percentage of news stories with $\delta_{3,4}$ in the range of 40 to 50 is as high as 90.8%, thus it still strongly indicates that for most news stories, their influenced user densities at distance 3 are higher than those at distance 4 for most of the time. The lower consistency counts $\delta_{4,5}$ is due to the small densities of influenced users at distance 4 and 5 for less popular news stories. This figure shows that the evolution of news stories over users at different distances are consistent for the majority of the news stories. Therefore, mathematical models can be used to describe the evolution dynamics. For most of the news stories, densities of influenced users decrease as the distances of the users increase, reconfirming that friendship is an important channel of information spreading. The phenomenon of the high ratio of news with consistency counts of 50 can be described by spatial-temporal PDE, which leads us to the linear diffusive model.

2) Spatial Heterogeneity of Users: To understand the impact of distance on the increase of the density of influenced users, we study the distance distributions of the direct and



(a) Density of influenced users of s1

(b) Density of influenced users of s2

(c) Density of influenced users of s3

(d) Density of influenced users of s4

Figure 1. Densities of influenced users over 50 hours for s1 to s4

indirect followers of all news submitters in the Digg data set. We define neighbors of a submitter as the set of all users that are reachable from the submitter in the social network graph. Figure 3 shows the neighbor distance distribution of the submitters of the four news stories and the average neighbor distance distribution for all submitters. From the figure, we can tell that users at distance 3 from a submitter consist of the largest number of its neighbors. In addition, the majority of a submitter's neighbors are at a distance of 2 to 5 from the submitter. For example, for the submitters of all four stories, their neighbors at distance 3 account for over 40% of all their neighbors. In average of all 3553 news stories, the users at distance 3 are 44.18% of all the users. As the distance increases from 6 to 10, the number of a submitter's neighbors drops sharply. In average, distance 1 to 5 users account for 96.67% of all users. Due to the small size of users at distance 6 to 10, we only present the results for users at distance 1 to 5 in the remaining of this paper.

This heterogeneity in neighbor distribution of submitters implies that the growth of influenced user densities should be dependent on distance x. The concavity of the shape of Figure 3 further suggests us to use a concave down quadratic function to naturally describe this heterogeneity in distance

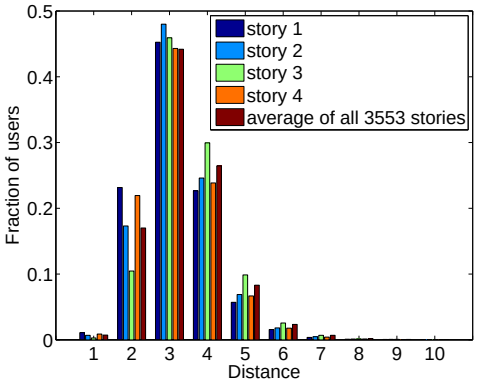


Figure 3. The average distribution of neighbors of the submitters for all news stories

in linear diffusive model proposed in section III.

3) Temporal News Decay: To understand how the user interests in news stories decay over time, we draw Figure 4 to illustrate the density of influenced users of the most popular story $s1$ from a different perspective than Figure 1[a]. Each of the 50 lines represents the density at time t where t varies from 1 to 50 hours. It shows that as

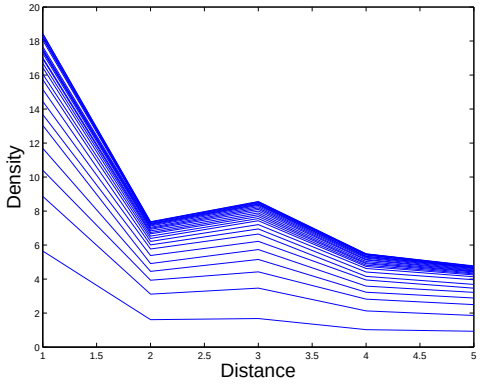


Figure 4. Density distribution of voted users over 50 hours

time progresses, the density of influenced users increases. However, the increasing rate of densities at t and $t+1$ drops exponentially as time elapses. This observation leads us to use an exponentially decreasing function in linear diffusive model proposed in section III to capture the decay rate of interests in news over time.

Our empirical results confirm the positive answer to the question "Does there exist some regularities of information diffusion dynamics through spatial and temporal dimensions in online social networks?" and motivate the design of the linear diffusive model.

III. Linear Diffusive Model

In this section, we propose a PDE based linear diffusive model to characterize the information diffusion process in both temporal and spatial dimensions. This model takes into account the heterogeneity of the spreading power of users at different distances from the submitter and the temporal decay of user interests in the news stories. This model is a simpler *linear* model than the *non-linear* diffusive logistic model in [1]. In addition, it achieves higher precision than the diffusive logistic model in [1].

A. Mathematical Formulation of Linear Diffusive Model

Let U denote the user population in an online social network, and s is the submitter of a news story. For each story, the user population U is broken down into groups based on a user's distance from the submitter s. The group U_x consists of all users at distance x from the submitter. As information propagates through online social networks, users express their interests in the information by voting for it.

Many PDE models come from a basic balance law, or conservation law. A conservation law is a mathematical formulation of the basic fact that the rate at which a quantity changes in a given domain must be equal to the rate at which the quantity flows across the boundary of the domain plus the rate at which the quantity is created, or destroyed, within the domain. In the context of information diffusion, the conservation law indicates that the rate of change of the influenced users with respect to time t comes from two processes: 1) social process, which corresponds to the quantities flowing across different groups, and 2) growth process, which corresponds to the quantities created within each group. Social process happens because the users in U_y can influence users in U_x where $x \neq y$ through direct or indirect friendship links. Growth process exists due to the fact that the users in U_x also influence each other. The growth process dominates the rate of change in the density of influenced users at a distance due to the following reasons: first, distance metric in online social networks has limited discrete values because of the small world scenario. For example, in Digg's context, users are clustered within a few hops from the submitter, and are divided into at most 12 distance groups. Therefore, the contribution of social process is limited; second, social triangles, also called triads formed by high clustering of users, are very common in online social networks [12]. Thus users in the same distance have high probability of being direct friends and having strong influence on each other. Compared with the context of infectious disease, the social process is similar to a disease spreading to a new community by infectious travelers, and the growth process is similar to the spreading of the disease within the new community.

To quantify the conservation law for information diffusion, let $I(x,t)$ denote the density of influenced users at distance x and time t. As widely used in spatial biology and epidemiology [13], the social process can be modeled by the following mathematical expression:

$$d\frac{\partial^2 I(x,t)}{\partial x^2} \tag{1}$$

where d describes the rate of information diffusion among users in different distance groups. In general, d may be dependent on I, x, t. In this paper, we choose d as a positive constant to represent the average of the diffusion rate. We call d social capability in the context of online social networks.

In order to precisely model the growth process, we take into accounts two characteristics observed in Digg data set. First, as shown in Figure 3, the distribution of the user populations at different distances from the submitter is not homogenous. The majority of users are in the groups of distance 2, 3 and 4. It is intuitive that the integrated spreading power of users (regardless of news content and time) at a distance depends on the number of users in a distance group. In other words, the more users in a distance group, the more integrated spreading power this distance group users can have on the growth process. Therefore, the growth function is dependent on distance x. The concavity of the shape of Figure 3 further suggests that we can use a concave down quadratic function $h(x)$ to describe this

heterogeneity in distance where

$$h(x) = -(x - \rho)(x - \sigma) \tag{2}$$

The coefficient of x^2 in $h(x)$ is scaled to -1. $h(x)$ reflects the rate of the change of influenced user density with respect to distance x.

Second, news stories are time-sensitive and the interests in news decay as time elapses. Figure 4 shows that the interests in news decay exponentially over time. This exponential news decay can be modeled by the following ordinary differential equation:

$$\begin{aligned} &\frac{dr(t)}{dt} = -\alpha r(t) + \beta \\ &r(1) = \gamma \end{aligned} \tag{3}$$

where $\frac{dr(t)}{dt}$ is the rate of change of r with respect to time t, α is the decay rate, γ is the initial rate of influence. β represents the residual rate after a news story becomes stable, which can be very small. Solving r in Equation (3), we obtain

$$r(t) = \frac{\beta}{\alpha} - e^{-\alpha(t-1)}(\frac{\beta}{\alpha} - \gamma) \tag{4}$$

In this setting, the simplest mathematical expression to model the growth process is

$$r(t)h(x)I(x,t) \tag{5}$$

which states that the rate of growth of the influenced users with respect to time within a distance group is proportional to the current density of influenced users I. This model is called linear model since the rate of change $I(x,t)$ with respect to time t in the growth process is proportional to $I(x,t)$. Thus, combining the social process (1) and the growth process (5) together, the conservation law for information diffusion can be formulated by the following linear diffusive equation with appropriate initial and boundary conditions:

$$\begin{aligned} &\frac{\partial I(x,t)}{\partial t} = d\frac{\partial^2 I(x,t)}{\partial x^2} + r(t)h(x)I(x,t) \\ &I(x,1) = \phi(x), \;\; l < x < L \\ &\frac{\partial I}{\partial x}(l,t) = \frac{\partial I}{\partial x}(L,t) = 0, \;\; t > 1 \end{aligned} \tag{6}$$

where

- $I(x,t)$ represents the density of influenced users with distance x at time t;
- d represents the social capability measuring how fast the information travels over friendship links in online social networks;
- $r(t)$ represents the average of news decay rate over all distances;
- $h(x)$ represents the heterogeneity of the spreading power of users in different distances;
- L and l represent the lower and upper bounds of the distances between the submitter and other social network users. In Digg network, $l = 1$ and $L = 12$;
- $\phi(x)$, which is greater or equal to zero but not identical to zero, is the initial density function, which can be constructed from historical data of information. Each information has its own unique initial function;
- $\frac{\partial I(x,t)}{\partial t}$ represents the first derivative of $I(x,t)$ with respect to time t;
- $\frac{\partial^2 I(x,t)}{\partial x^2}$ represents the second derivative of $I(x,t)$ with respect to distance x.

$\frac{\partial I}{\partial x}(l,t) = \frac{\partial I}{\partial x}(L,t) = 0$ is the Neumann boundary condition [13], which means no flux of information across the boundaries at $x = l, L$. Therefore, we assume that an online social network is a closed environment without external input.

B. Initial Density Function Construction

Now we present the method to construct the initial density function ϕ. In general, the initial function is constructed using the data collected from the initial stage of information diffusion. Specifically, ϕ is a function of distance x which captures the density of influenced user at distance x at the initial time when a news story is submitted.

In online social networks, it is only possible to observe discrete values for the initial density function because the distance x is discrete. As in [1], we apply an effective mechanism available in Matlab cubic spline package, called *cubic splines interpolation* [14], to interpolate the initial discrete data in constructing $\phi(x)$. Using this process, a series of unique cubic polynomials are fitted between each of the data points, with the stipulation that the obtained curve is continuous and smooth. Hence $\phi(x)$ constructed by the cubic splines interpolation is a piecewise-defined function and twice continuous differentiable. After cubic splines interpolation, we simply set the two ends to be flat to satisfy the second requirement since in this way the slopes of the density function $\phi(x)$ at the left and right ends are zero.

IV. Model Validation

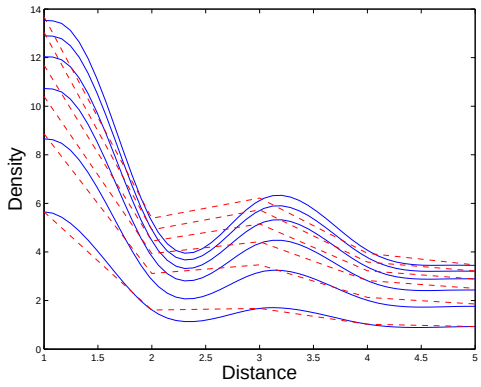
In this section, we evaluate the performance of the proposed linear diffusive model by comparing the density calculated by the model with the actual observations in the Digg data set. We first present the model accuracy for the most popular news story, then examine the overall accuracy of the proposed linear diffusive model for all news stories with more than 3000 votes in Digg. Lastly, we study the correlation between the number of the submitter's followers and the accuracy of the model.

The construction of initial density function follows the method outlined in Section III.B. Specifically, we create the initial density function for each news story using the density of influenced users captured at the first hour when the news

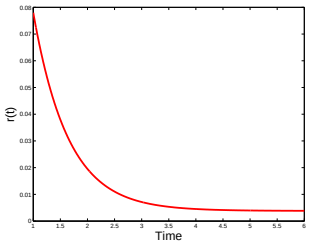
density at distance 2 starts to grow. This is also the time when news stories start to spread to users at distances beyond 2. This time varies for each news story. For example, for story $s1$, the initial condition is built with the density at first hour after the news story is submitted while the first hour for story $s4$ is the fourteenth hour after the news submission. This is similar to the concept of Digg hour defined in [7]. The model accuracy is defined as follows:

$$\text{model\_accuracy} = 1 - \frac{|\text{predicted\_value} - \text{actual\_value}|}{\text{actual\_value}} \tag{7}$$

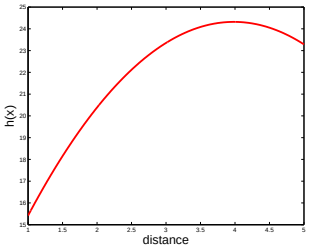
A. Model Accuracy: A Case Study



(a) Predicted (blue, solid) vs. Actual data (red, dotted)



(b) $r(t)$



(c) $h(x)$

Parameter	value
d	0.0020
α	1.5526
β	0.0059
γ	0.0780
ρ	-0.9478
σ	8.9149

(d) Parameter values

Distance	Average
1	97.88%
2	97.27%
3	97.44%
4	96.20%
5	98.25%
Overall	97.41%

(e) Model Accuracy

Figure 5. Model accuracy for the most popular news story in the Digg data set

In this subsection, we show the accuracy of the linear diffusive model for the most popular news story in Digg as a case study. Examining this specific news story is necessary since this news story has statistically a large number of votes thus the abnormal behavior of individual users can be diluted. Compared with the spreading of infectious disease, an occurrence of such a popular news story resembles a massive outbreak of an epidemic of which the accurate description has been heavily investigated in epidemiology.

Figure 5[a] illustrates the performance of the linear diffusive model for the most popular story $s1$. The dashed lines denote the *actual* observations for the density at different times, while the solid lines illustrate the density *calculated* by the linear diffusive model. Note that the lowest line representing $t = 1$ is the initial density function. In online social networks, the density is only meaningful when distance is integer. It is clear that the calculated values closely follow the actual values over time and distance. Figure 5[b] gives the shape of $r(t)$ and Figure 5[c] gives the shape of $h(x)$. As can be seen, $r(t)$ is an exponentially decreasing function as expected and $h(x)$ is a concave down function with peak between 3 and 4, which deviates from the peak of the neighbor distribution illustrated in Figure 3. This is an indicator that there exist highly influential users at distance 4 from the submitter of news $s1$. The corresponding parameters are listed in Figure 5[d]. The parameter are adjusted manually to best fit the actual data. Social capability d is relatively small, which is consistent with our discussion in the model that growth process has dominating impact on the information diffusion process. α, β, γ decide the shape of $r(t)$, and ρ and σ decide the peak of $h(x)$. The average accuracy at different distances are calculated for time $t = 2, ...6$, and are provided in Figure 5[e]. The model can achieve high accuracy across all distances. The average accuracy for $s1$ is 97.41%, much higher than that in [1], where the average accuracy for $s1$ is 92.08%.

B. Overall Accuracy of Linear Diffusive Model

In order to study the capability of the model for describing all news stories in the Digg data set and examine whether the model can capture the intrinsic features in information diffusion over Digg network, we explore the overall accuracy of the linear diffusive model for all 133 news stories with over 3000 votes in the Digg data set. We choose to study news stories with over 3000 votes to avoid statistically skewed results: first, news stories with a smaller number of votes are statistically less meaningful; second, news stories with a small number of votes have very low densities of influenced users across all distances. For example, in Figure 1[d], for the news story with 1618 votes, the density of users at distance 1 is less than 2.5%, and the density of users at distance 2 to 5 are all less than 0.8%. This low density can cause high error rate since even a small deviation between the calculated value and the observed value accounts for a large percentage of the observed value.

To automate the parameter selection of a batch of news stories, we use the least square fitting technique provided in Matlab to search for parameters to best fit the actual data, where the accuracy of the model depends on the number

of searches in the parameter space. The more iteration of search, the higher the accuracy. The results presented here use an iteration of 100000 and fit the observed data for 20 hours.

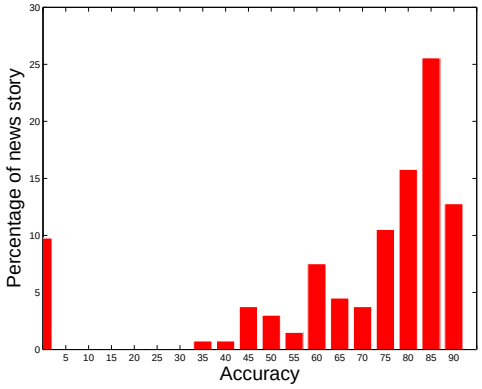


Figure 6. Accuracy histogram of news stories with more than 3000 votes. x-axis is the accuracy, y-axis is the percentage of the news stories that can be predicted with the accuracy.

Figure 6 illustrates that about 13% of news stories can be described with accuracy higher than 90%. In total, about 60% of news stories can be described with accuracy higher than 80%. Note that if manually adjust parameters for each individual news story, higher accuracy can be achieved. For example, for the most popular news story, with manually adjusted parameters, the average accuracy can reach to 97.41%, while with the automated parameter selection, the average accuracy is only about 91%. The high accuracy across all news stories with over 3000 votes shows a strong evidence that the linear diffusive model captures the intrinsic diffusion patterns of news and can be used as an effective approach to describe the news in Digg.

C. Influence of the Number of Submitter's Followers

When studying the diffusion pattern of news in the Digg data set, we observe the spreading dynamics of a news story with 8507 votes is much different from other news stories of the same scale. Correspondingly, linear diffusive model achieves extremely low accuracy for this news story. This special case raises an interesting question, "What caused the abnormal behavior of the news stories which do not show consistent evolution patterns?" We investigate the relationship between the model accuracy and the number of votes a news story receives, and find no obvious correlation between these two metrics. With further investigations on this news story, we discover that the submitter of this news story has only 10 followers. This motivates us to investigate the relationship between the number of direct followers of the submitter and the model accuracy.

Figure 7 shows that there exists correlation between accuracy and the number of the direct followers of the news submitter for the news stories with more than 3000 votes. In

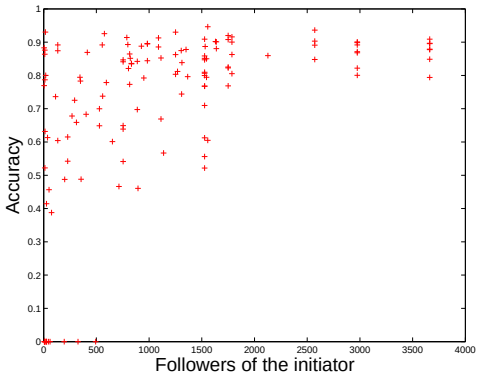


Figure 7. Correlation between accuracy and the number of hop one neighbors of the news submitter

this figure, we can see that the number of the direct followers of the submitter has an impact on the accuracy (which indicates the evolution dynamics of a news story) when the submitter is an influential user with a large number of followers. For all the news stories initiated from submitters with more than 2000 followers, the linear diffusive model can achieve over 80% accuracy. For all the news stories with submitters that have more than 500 direct followers, only two news stories have accuracy lower than 50%. However, there is little correlation between the accuracy and the number of the direct followers of the submitter when the submitter has less direct followers. For example, the accuracies of news stories initiated from a submitter with 500 direct followers vary from 0 to 92%. This indicates that news stories initiated from influential submitters follow consistency evolving patterns and are more predictable.

In summary, the experiment results show that the PDE-based linear diffusive model can achieve high accuracy of 97.41% for the most popular news story and show consistent high accuracy for news stories with more than 3000 votes. We also reveal the correlation between the number of submitter's direct followers and the spreading dynamics of the news initiated from the submitter. Thus, we can confidently conclude that the linear diffusive model can characterize the process of information dissemination and can be used as an effective method to describe the process.

V. Applications and Discussions

In this section, we sketch some applications of the proposed linear diffusive model in online social networks. Firstly, through introducing friendship hops as a distance metric into the model, we build the connection between the linear diffusive model and the topology of the social network graph. Therefore, the experimental results of the linear diffusive model can help discover new patterns in the underlying topology. For example, in section IV, the low model accuracy of a news story with large number of votes motivates us to investigate the relationship between the

number of first hop neighbors of the submitter of a news story and the accuracy, and leads us to the discovery that information spreading from a powerful submitter usually follows a consistent evolving pattern. More importantly, the linear diffusive model provides a viable approach to search for super spreaders, which has significant importance in online social networks. In linear diffusive model, the impact of time and distance on growth function is modeled separately as $r(t)$ and $h(x)$. This dictates that the peak of $h(x)$ is related to the largest growth rate per capita and therefore the corresponding location x is the group where the super spreaders locate. An interesting observation from $h(x)$ of the most popular news story in Digg is that the peak is at around distance 4 instead of expected distance 3 where the user population is the highest. This indicates that there might exist super-spreaders in distance 4 users.

Secondly, the linear diffusive model also provides an approach to classify news stories in online social networks based on their parameters. The parameters that best fit the observed data reveal the spreading patterns of the news story. For example, we can group news stories with similar decay patterns $r(t)$ which are decided by α, β, and γ into one clusters. On the other hand, the model can be used as a verification of the clusters identified by other approaches, e.g., a data mining approach to discover the clusters of news stories and users in Digg network in [15]. The quality of these clusters can be verified by calculating whether the news stories in the same clusters have similar parameters in the model.

Thirdly, since the linear diffusive model studies information diffusion from both temporal and spatial dimensions, we can build a visualization tool to demonstrate the speed and coverage of real or simulated information. This cannot be achieved by only studying the diffusion in temporal dimension.

Finally, an important application of a mathematical model is its ability to predict the future events. This paper proposes a simple model with only six parameters to precisely describe information spreading over time and distance in Digg. This capability of describing the past events is the first and important step towards the prediction. With more detailed study on parameter estimations, we can provide guidelines on parameter selections in order to predict the evolution of future news stories.

VI. Related Work

Information diffusion over online social networks has drawn much attention from the networking and data mining research communities [16][17]. The existing work can be classified into three categories: 1) empirical study, 2) macroscopic modeling, and 3) microscopic modeling.

Empirical study: There exist empirical studies on various online social networks which reveal intrinsic patterns of information diffusion dynamics in the corresponding networks. [18] studied the dynamics of information propagation in weblogs. [5], [6], [19] studied cascade patterns of disseminating popular photos over Flickr social networks, while [2] examined how the interest in news stories spreads among the users in Digg and Twitter social networks based on empirical data extracted from these networks. [3] studied the factors that prohibit the epidemic transmission of popular news posted on Digg. In addition, Tang *et al.* presented a large-scale empirical study on network structure, user characteristics, and content dissemination process of Digg social network [20].

Macroscopic modeling: A few research efforts have focused on modeling information diffusion in social networks from a global point of view. [21] provided a survey on mathematical models of information diffusion. [22] proposed an information diffusion model to leverage interpersonal diffusion rate based on continuous time Markov chain, while [9] introduced a *Linear Influence Model* to predict the number of newly influenced users based on the time in which the previous set of users are influenced. The SIS (susceptible, infected, and susceptible) epidemic model was used in [23] to characterize the process of information diffusion over social networks during a given time period. Similarly, [7], [9], [10] studied different mathematical models to understand information diffusion in social networks over a time period without considering the underlying network structure. A recent paper [1] proposed the first model to characterize information diffusion in both spatial and temporal dimensions and verified the model with the most popular news in a Digg data set. [24] studied a free boundary problem for the logistic model in [1] where one side of the boundaries can change with respect to time and obtained the moving speed of the boundary.

Microscopic modeling: Understanding microscopic user interaction is important since both predicting of information diffusion and choosing influential individuals rely on the precision of the interpersonal interaction model. [25] proposed various models to predict whether an individual can have positive or negative impact on a neighbor. [26] introduced user-to-user interaction principles which can produce macroscopic user behavior matching the observed Weblog dynamics. An earlier work [27] used two basic diffusion models, namely *Linear Threshold* and *Independent Cascade Models*, to search the most influential users in online social networks. A recent work [28] developed a link-based latent variable model to describe a weighted friendship between two individuals rather than traditional binary friendships.

This paper belongs to the *macroscopic modeling* category and studies the spatio-temporal diffusion problem. Different from [1], this paper investigates the diffusion property in Digg network in further depth and proposes a simpler mathematical model considering both the decay of interests in news over time and the heterogeneity of spreading power

of users at different distances. We also provide statistical results of the accuracy of the linear diffusive model over all news with more than 3000 votes in the Digg data set.

In parallel with research in computer science, there are also research in biology, sociology, economics, and physics to model time evolution systems [29], [30], [31], [32], [33] and use dynamic mathematical models including ordinary differential equation and partial differential equation models to translate local assumptions or data on the movement and reproduction of individuals into global conclusions on the population. This paper is different from dynamic mathematical modeling in other disciplines in that the concepts such as social distance, information diffusion, and influence growth are abstract and unique for online social networks.

VII. Conclusions and Future Work

This paper performs empirical studies on the characteristics of information spreading in temporal and spatial dimensions in Digg social network and introduces a linear diffusive model to describe the spatio-temporal diffusion characteristics. The linear diffusive model takes into account the heterogeneity of spreading powers of users at different distances from the source in online social networks and the decay of news over time. The proposed model is validated and evaluated with the real data collected from Digg social network. Our experiment results show that the model achieves high accuracy for the majority of news with more than 3000 votes in Digg and achieves higher precision than the diffusive logistic model proposed in [1]. Since the linear diffusive model captures essential factors shaping information diffusion in online social networks, it can be applied to characterize other social media, such as Twitter, with adjusted parameters. For future work, we plan to: 1) study the parameter estimations of the model, and 2) study the spatio-temporal pattern in Twitter social network and adjust the linear diffusive model for Twitter.

Acknowledgments

Feng Wang, Haiyan Wang, and Kuai Xu are supported in part by the National Science Foundation under the grant CNS #1218212.

References

[1] F. Wang, H. Wang, and K. Xu, "Diffusive logistic model towards predicting information diffusion in online social networks," in *Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking (HOTPOST)*, 2012.

[2] K. Lerman, and R. Ghosh, "Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks," in *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[3] G. V. Steeg, R. Ghosh, and K. Lerman, "What Stops Social Epidemics?" in *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2011.

[4] S. Ye and F. Wu, "Measuring Message Propagation and Social Influence on Twitter.com," in *Proceedings of the Second international conference on Social informatics (SocInfo)*, 2010.

[5] M. Cha, A. Mislove, B. Adams, K. Gummadi, "Characterizing Social Cascades in Flickr," in *Proceedings of ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2008.

[6] M. Cha, 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 (WWW)*, 2009.

[7] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of The ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[8] T. Hogg and K. Lerman, "Social dynamics of digg," *http://arxiv.org/pdf/1202.0031v1.pdf*, 2012.

[9] J. Yang and J. Leskovec, "Modeling Information Diffusion in Implicit Networks," in *Proceedings of IEEE International Conference on Data Mining*, 2010.

[10] R. Ghosh and K. Lerman, "A Framework for Quantitative Analysis of Cascades on Networks," in *Proceedings of Web Search and Data Mining Conference (WSDM)*, 2011.

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

[12] D. Easley and J. Kleinberg, "Networks, crowds, and markets: Reasoning about a highly connected world," *Cambridge University Press*, 2010.

[13] J. D. Murray, *Mathematical Biology I. An Introduction*. Springer-Verlag, 1989.

[14] C. Gerald, and P. Wheatley, *Applied Numerical Analysis*. Addison-Wesley, 1994.

[15] F. Wang, K. Xu, and H. Wang, "Discovering shared interests in online social networks," in *Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking (HOTPOST)*, 2012.

[16] D. Agrawal, C. Budak, and A. E. Abbadi, "Tutorial: Information diffusion in social networks: Observing and influencing societal interests," in *Proceedings of International Conference on Very Large Data Bases*, 2011.

[17] J. Leskovec, "Tutorial: Analytics & predictive models for social media," in *Proceedings of International Conference on World Wide Web (WWW)*, 2011.

[18] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of International Conference on World Wide Web (WWW)*, 2004.

[19] B. Yu, and H. Fei, "Modeling Social Cascade in the Flickr Social Network," in *Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2009.

[20] S. Tang, N. Blenn, C. Doerr, and P. V. Mieghem, "Digging in the Digg Social News Website," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1163–1175, 2011.

[21] W. An, "Models and methods to identify peer effects," *The SAGE Handbook of Social Network Analysis*, 2010.

[22] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in *Proceedings of International Conference on World Wide Web*, 2007.

[23] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Efficient Discovery of Influential Nodes for SIS Models in Social Networks," *Knowledge and Information Systems*, 2011.

[24] C. Lei, Z. Lin, and H. Wang, "The free boundary problem describing information diffusion in online social networks," *J. Differential Equations*, vol. 254, pp. 1326–1341, 2013.

[25] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of International Conference on World wide web (WWW)*, 2010.

[26] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos, "Modeling blog dynamics," in *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2009.

[27] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[28] R.J.Xiang, J. Neville, and M.Rogati, "Modeling relationship strength in online social networks," in *Proceedings of International Conference on World wide web (WWW)*, 2010.

[29] R. Cantrell and C. Cosner, *Spatial Ecology via Reaction-Diffusion Equation*. Wiley, 2003.

[30] P. Fife, "Spatial Ecology via Reaction-Diffusion Equation," *Mathematical Aspects of Reacting and Diffusing Systems*, 1979.

[31] M. J. Keeling and K. T.D Eames, "Networks and epidemic models," *Journal of The Royal Society Interface*, pp. 295–307, 2005.

[32] W. Goffman and V. A. Newwill, "Generalization of epidemic theory: An application to the transmission of ideas," *Nature*, pp. 225–228, 1964.

[33] Alain Barrat, Marc Bathelemy, and Alessandro Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.