Predicting Information Diffusion Initiated from Multiple Sources in Online Social Networks

Chuan Peng
School of Computer science, Wuhan University
Email: chuan.peng@asu.edu

Kuai Xu, Feng Wang, Haiyan Wang
Arizona State University
Email: {kuai.xu, fwang25, haiyan.wang}@asu.edu

***Abstract*—In recent years, online social networks such as Twitter and Facebook have become a major channel for information dissemination and communication. A number of prior studies apply mathematical approaches to characterize and model the complex dynamics of information dissemination, also called information diffusion over online social networks. Most of these work focus on the diffusion process of information that are posted by a single source, however few studies consider the diffusion patterns of information that come from multiple sources. As breaking news stories, emergency events, and controversial topics are often initiated by a number of sources, it is very important to understand the diffusion patterns of these multi-source information as well. In this paper we first study the basic characteristics of the diffusion process of multi-source informations via real data-sets collected from Digg, a social news aggregation site. Subsequently, we use a mathematical model to predict the information diffusion process of such multi-source news. Finally we use news in the same data-set to validate the accuracy of the proposed mathematical model. Our experiment results show that the model can describe the most representative news stories initiated from multiple sources with an accuracy higher than 90%, and can achieve an average accuracy around 75% across all multi-source news stories in the data-set. These results suggest that our approach is able to characterize and predict the spreading patterns of multi-source informations with high accuracy.**

***Keywords*-online social network; information influence; multiple sources; prediction;**

I. Introduction

Online social networks (OSNs) have recently played an increasingly significant role in information propagation for today's society. OSNs connect users via friendship or following/follwer relationships, and allow them to interact with each other including spreading information widely in social networks. A fundamental problem in understanding user interactions over OSNs is how and why influence spreads across such networks. A deep understanding of the process of information or influence diffusion can help us develop effective approaches for describing and predicting the spreading process of information, and more importantly could allow us to leverage OSNs as an appropriate way for propagating positive and emergency information as well as effectively containing the spreading of negative influence, e.g., rumors.

A rich body of research have recently studied information diffusion over OSNs via empirical analysis, and characterized its dynamics and complex natures [1], [2], [3]. On the other hand, a number of mathematical models have been proposed to quantitatively and accurately describe the process of information diffusion in OSNs [4], [5], [6], [7]. In general these models could be classified into three groups, i.e., local models, global models and hybrid models, based their focuses on user interactions in local communities or global networks. Two most representative local models are linear threshold model and independent cascade model [7], which analyze the behavior of a given users according to the behavior of immediate friends or followers. Some other models prefer to consider the behavior of users from a global perspective, e.g., [8] developing a model to capture users' behavior based on the behaviors of the rest of user population. In addition, [9], [10], [11], [12], [13] also studied several mathematical models over the temporal aspect from a global point of view. A few recent studies [8] explore hybrid models to investigate information diffusion from both local and global perspectives, in particularly, focusing on the interactions between local communities and the entire social network graphs to characterize the behavior of a given user.

Although many research have been done to model the process of information diffusion over OSNs, these models only focus on describing and predicting information diffusion along the temporal dimension [9], [10], [11], [14]. Our prior studies recently introduced PDE-based diffusion models, namely *diffusive logistic model* [15] and *linear diffusive model* [16], to precisely characterize diffusion dynamics from not only the temporal dimension, but also the spatial dimension from a global point of view, thus gaining a deep understanding of information diffusion over OSNs.

The PDE-based diffusion models address the spatial-temporal problem of information diffusion. Specifically, the models are able to describe and predict the density of influenced users, $d(x,t)$, at a distance of x from the information source after a certain period of time t. The experimental results based on Digg news stories demonstrate the capability and high accuracy of the models in describing the process of information diffusion. However, these prior studies only consider the news stories that are initiated from

a *single* source.

As online social networks have played an increasingly important role in disseminating news stories, promoting news products and political campaigns thanks to their growing popularity, it becomes very common to observe a popular news story, e.g, the final result of a popular sport game, to be posted by multiple fans at the same time, and then forwarded or retweeted by other users. Unlike single-source information, multiple-source information are originated from a set of initiators in OSNs. For example, two Digg users firstly submit a certain news story at the same time before the rest of other users in OSNs, and then the same information cascades in parallel along both temporal and spatial dimensions. Thus this paper extends prior effort to investigate the diffusion process of information that are initiated simultaneously by two or more sources, referred to as *information diffusion from multiple-sources problem*: given an information initiated from a set of multiple sources $S = s_1, s_2, ..., s_m$, what is the density of influenced user, $d(x, t)$, at the distance of x from the multi-sources after a period of time t.

As OSNs users have a variety of distances to multiple-sources who initiate the same information, this paper first designs a simple approach to quantify the distance values between every OSNs user and multiple-sources. Subsequently, we propose an algorithm to effectively choose 1433 multiple-source news stories from Digg data-set which are approximately initiated from two or more sources, and divide these stories into different groups based on the number of multiple-sources.

With thousands of multiple-source news stories, we characterize the temporal patterns of information diffusion for these news stories, and analyze spatial distribution of influenced users from these multiple-sources. To understand whether the proposed models in our prior studies are able to describe the process of information diffusion for multiple-source news stories, we apply linear diffusion model to characterize and predict the information diffusion process of these 1433 news stories. Our experiment results show that linear diffusion model is able to achieve an average of 75% prediction accuracy on the density of influenced users across all groups of news stories. In particularly, this model achieves over 90% accuracy for the most popular news stores across all groups. Thus these results verify that our PDE-based diffusive models can accurately describe the influence spreading over both spatial and temporal dimensions not only for single-source news stories but also for multiple-source news stories.

The remainder of the paper is organized as follows. Section II describes our method of calculating distances from influenced users from multiple users, while Section III presents a simple algorithm for finding multiple-sources news stories. Section IV presents our experimental results, while Section V concludes this paper and outlines our future work.

II. Measuring Distance Between An Influenced User and Multiple Sources

Previous studies [15] have shown that distance between an initiator of a news story and influenced users plays an essential role in the process of information diffusion over online social networks. Thus a key question in this study is to measure the distance between a user and multiple sources that initiate a particular news at approximately the same time. The distance between a single pair of users in online social networks is often measured by the number of hops in the shorted path between them, which is also refereed to as friendship hops in [15]. Based on this definition, one could divide all users of an online social network into a number of disjoint groups according to their distances from a news initiator. For example, the immediate followers of the initiator have a distance of 1, and the followers of the initiator's immediate followers have a distance of 2. Continuing this process will cluster all users into distinctive groups.

However in this study we focus on the process of information diffusion initiated simultaneously from multiple users, therefore the first step of our approach is to measure the distance from a users to a set of initiators rather than a single one. Specifically, we consider the distance between a given users and a set of multiple sources as the minimum value among all the friendship hops calculated between the user and each of the sources. Let U represent the entire user population in the online social network, and $S = \{s_i | i = 1, 2, ..., n\}$ denote the set of n initiators of a given news story. Given a user u, let $d(s_i, u)$ denote the distance from s_i to the user u. Then, $d_{min} = min\{d(s_i, u) | i = 1, 2, ..., n\}$ is defined as the distance between u and these multiple sources.

Our intuition of choosing the minimum shortest path as the distance between a user and multiple sources is based on a simple yet common observation. When a user of online social networks could potentially be influenced by a set of sources that initiate a news story via different multiple paths, the nearest source to which the user has the minimum friendship hop has the highest probability of influencing the behavior of the given user due to the smallest number of friendship hops.

Given this distance definition, we divide all online social network users, U, into a set of groups, $U = \{U_i | i = 1, 2, ..., m\}$, based on their distances to the multiple sources of a news story, where m is the maximum distance among all users to this set of initiators, and the group U_i consists of users that have a distance of i to the multiple sources.

III. Identifying News Stories Initiated by Multiple Sources from Digg Data-Set

In order to characterize the diffusion process of information initiated from multiple sources, our next step is identify

such information for in-depth analysis. In this study, we use the data-set collected from Digg, a major news aggregation site. Digg users submit Web links of news stories which they read in news sites or blogs to www.digg.com, such that the other Digg users could read, vote (also called digg) or comment on these news stories. In this paper, we refer to the first Digg users who bring the news stories to the Digg site as the news initiators or sources. Besides sharing news stories, the Digg users also form friendship relationships via following each other.

The Digg data-set consists of 3553 most popular new stories on Digg site during June 2009. These news stories have received a total of over 3 millions votes from 139,409 Digg users. For each news story, the data-set contains the ID of all the users who have voted on the news, and the timestamps when each of the vote was cast. The time granularity is measured in sections. Due to the fine time granularity, it is very difficult to find news stories that are simultaneously initiated by multiple users at the exact same time. Therefore, we develop a simple yet effective approximate algorithm, as illustrated in Algorithm 1, to identify news stories initiated with multiple sources.

Algorithm 1 An approximate algorithm for identifying news stories with multiple sources from Digg Data-Set

1: Parameters: a news stories s, the very first user who digged the news story, u, and time threshold T;
2: search the set of direct followers of u in the friendship graph, represented as F_u;
3: identify Digg users in F_u who also digged the news story s, denoted as V_s;
4: sort users in V_s according to the voting timesstamp in a non-decreasing order, represented as V_s';
5: locate a set of Digg users in V_s' who do not have any following relationships, denoted as M;
6: select the first n users voters, M_n, who vote no late than T away from start timestamp from M as the initial set of multiple initiators or sources for the given news story s.

After we identify the news story s and the set of multiple sources, M_n using the above algorithm, we assume that this news story is initiated by multiple users, and only study the votes that are received after the votes by these multiple sources M_n. In other words, we consider these sources as the initial submitters, and study the process of information diffusion starting from these multiple sources. In this study, we consider that non-neighbor users who vote for the given information within the first early 5 minutes can be approximately seen as simultaneous submitters, i.e. the multiple sources. The approximation algorithm provides us 1433 news stories that are initiated by multiple sources in our Digg data-set.

IV. Experiment Results

The aforementioned algorithm leads to a collection of 1433 news stories that are simultaneously initiated from two more more sources. Based on the number of the sources each story has, we classify these 1433 news stories into different groups. Specifically, we classify these news into six groups: 2-source, 3-source, 4-source, 5-source, 6-source and 8-source, which include 1045, 304, 64, 16, 3 and 1 news stories, respectively.

In this section, we perform an empirical analysis on the spreading patterns of these news stories along temporal and spatial dimensions. Then, we apply *linear diffusive model* [16] to characterize and predict the spreading patterns using these 1433 multiple-source news stories.

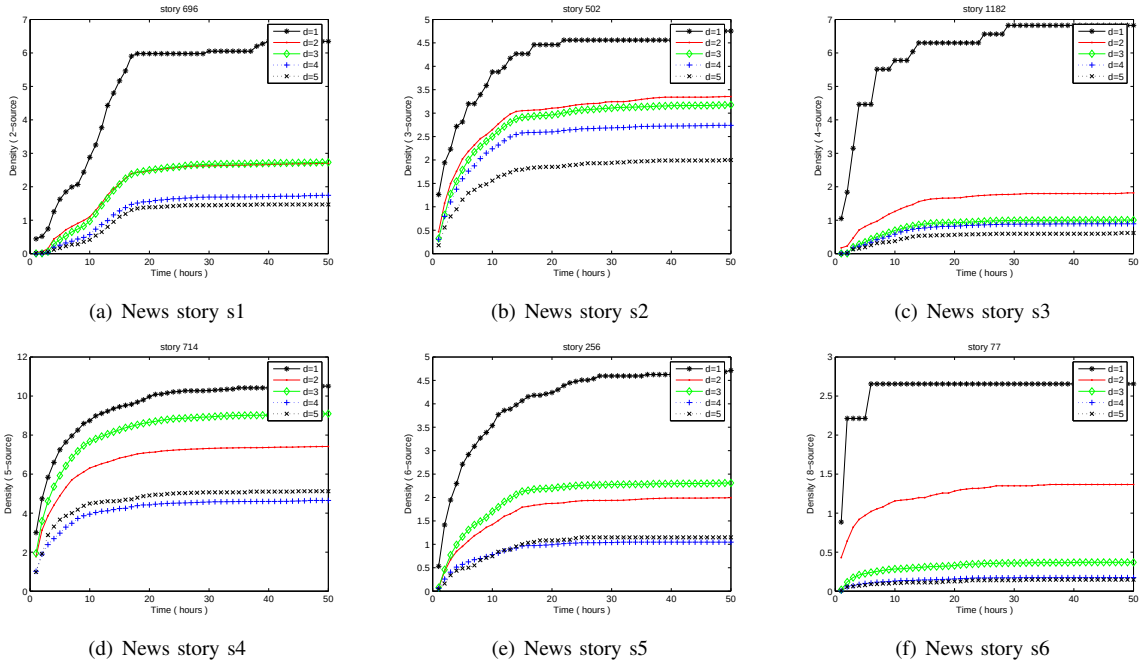
A. Temporal Patterns of Information Diffusion for News Stories Initiated from Multiple-Sources

When an information starts to circulate or spread over online social networks, some users expresses their interests and opinions through certain actions, e.g., digging, forwarding, retweeting, while some users may choose to simply ignore the information. If a user takes certain actions on the information, we refer to this user as an *influenced user* of this information. Note that we use U_i to denote the group of users that have a distance of i to the multiple-sources of the information. Similarly we could use $U_i'(t)$ to denote the users in U_i, who have been influenced by the information at a given time t. Thus, we could calculate $d(x,t)$, the density of the influenced users at distance x at the time t, as $\frac{U_i'(t)}{U_i}$. In other words, $d(x,t)$ represents the ratio of the total influenced users in U_i over the total users of U_i at a given time t.

To understand the patterns of information diffusion for news stories initiated from multiple-sources, we first select the top news story from each group of multiple-source news stories based on the number of diggs as case studies. Let s1, s2, s3, s4, s5, s6 denote the most popular new stories from each group. These news stories have received 7549, 8492, 2753, 24099, 5251, 780 diggs, respectively. Figures 1[a-f] demonstrate the densities of influenced users with distance 1 to 5 from multiple-sources for these six news stories over 50-hour time-span. Each line in the figures represents the density of influenced users at a given distance from multiple-sources that initiate the news stories simultaneously.

As shown in these figures, the densities of influence users across six news stories initiated from multiple-sources exhibit several interesting patterns along the temporal and spatial dimensions. From the temporal perspective, we can see that the densities of influenced users increase fairly fast during the initial few hours and then slow down gradually. Along the spatial perspective, we find that the densities of influenced users with smaller distance are higher than those with larger distances. This observation is not surprising since the initiators have stronger influences on users that are closer to them in the network topology of friendship graphs.

More importantly, these patterns are consistent with our prior studies [15], [16] that analyze the densities on influenced users for news stories initiated from a single source. The similar patterns of information diffusion for news stories initiated from multiple-sources and single-sources lead us



(a) News story s1

(b) News story s2

(c) News story s3

(d) News story s4

(e) News story s5

(f) News story s6

Figure 1. Density of influenced users over 50 hours for the top news story from six groups based on the number of diggs

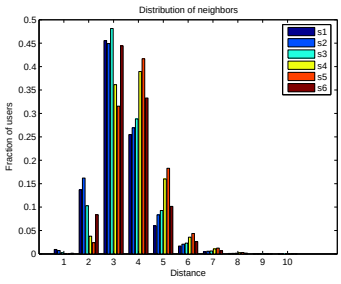


Figure 2. Distance distribution of influenced users from multiple-source initiators for six news stories

to validate the performance of our proposed linear diffusive model in [16] in characterizing and predicting information diffusion for news stories initiated from multiple-sources.

B. Spatial Distributions of Influenced Users From Multiple-Sources

To further examine the impact the distance, i.e., friendship hops in the social network graph, on the density of influenced users, we next study the spatial distributions of influenced users from multiple-sources using the same set of news stories. Note that we use the shortest distance to denote the distance between any given user in the online social networks and multiple-sources that have initiated the same news story.

Figure 2 illustrates the distance distribution of influenced users from multiple-source initiators for the news stories s1, s2, s3, s4, s5 and s6. In general, most influenced users have a distance of 2 to 5 to the group of multiple-sources, and the distances of 3 and 4 have the highest percentages of influenced users. As the distance increases to 6 or 7, there is only a few influenced users due to a very small user population at such distances. Thus we only consider the users with a distance of 1 to 5 when validating linear diffusion model against news stories with multiple-source initiators. The heterogeneity in distance distribution of influenced users further indicates that density growth of influenced user depends on distance to a great extent.

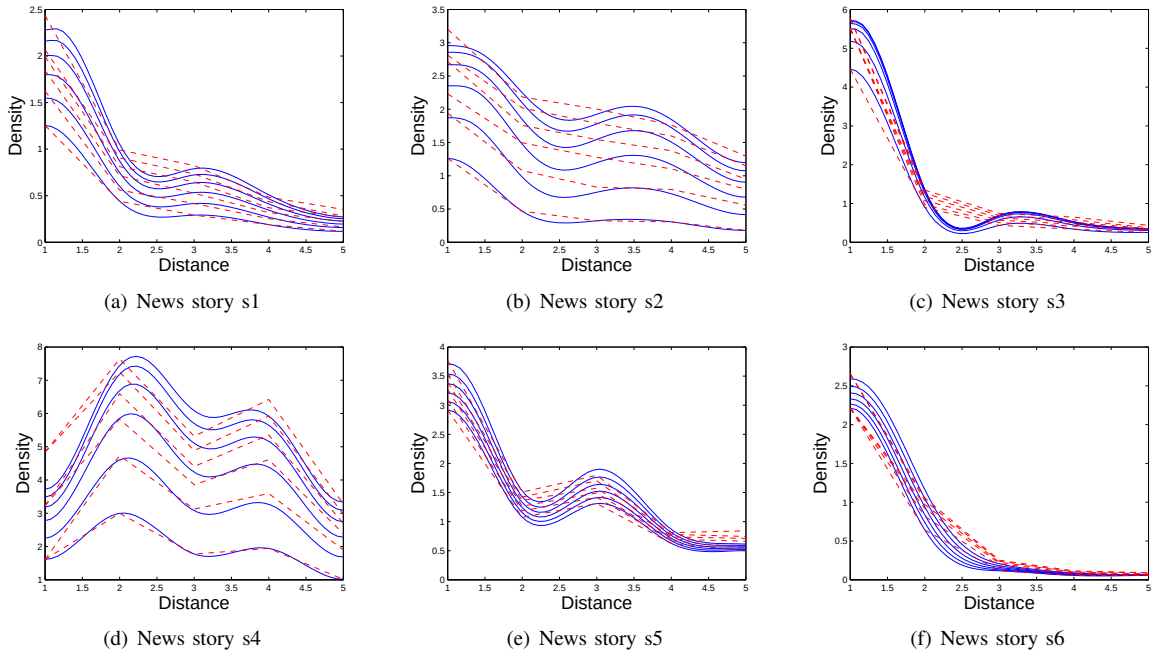
C. Predicting Information Diffusion Initiated From Multiple-Sources

In this subsection we validate the accuracy of *Linear Diffusive Model* by comparing the densities calculated by the model with the actual values observed in the Digg data-set. We quantitatively measure the predicting accuracy of the model, $f_{accuracy}$ as follows:

$$f_{accuracy} = 1 - \frac{|v_p - v_a|}{v_a}, \tag{1}$$

where v_p denotes the density predicted by the model while v_a denotes the actual value from the real Digg data-set. Clearly, $0 \leq f_{accuracy} \leq 1$.

Figures 3[a-f] illustrate the accuracy of the model for the most popular news stories s1 to s6 in each group of multiple-source news stories. As shown in these figures, the model achieves a high accuracy of predicting the density of influenced users over time across six different news stories that are initiated from different numbers of sources.



(a) News story s1 (b) News story s2 (c) News story s3

(d) News story s4 (e) News story s5 (f) News story s6

Figure 3. The prediction accuracy of linear diffusion model on six popular news initiated from multiple-sources. The dashed line denotes the actual observation for the density of influenced users over time, while the solid one denotes the density calculated by the model.

To evaluate the accuracy of the model on all other news stories, we run the model on all 1433 news stories in our data-set. Table I illustrates the results on the accuracy of the model on all these news stories as well as on the most popular news for each group. The first column denotes the group of multiple-source news stories, the second and third columns show the most popular news story for each group and the prediction accuracy. The last two columns summarize the total number of news stories in each group and the average prediction accuracy among all the news stories in the same group. Apparently, the model is able to achieve over 90% accuracy for the top news from all groups. More importantly, this model achieves very high prediction accuracy for other news stories as well. The average accuracy of all 1433 news stories is 76.25%, and the average accuracies for all groups are higher than 70%. These findings confirm the prediction capability of the linear diffusion model on news stores initiated from single sources as well as news stories initiated from multiple-sources.

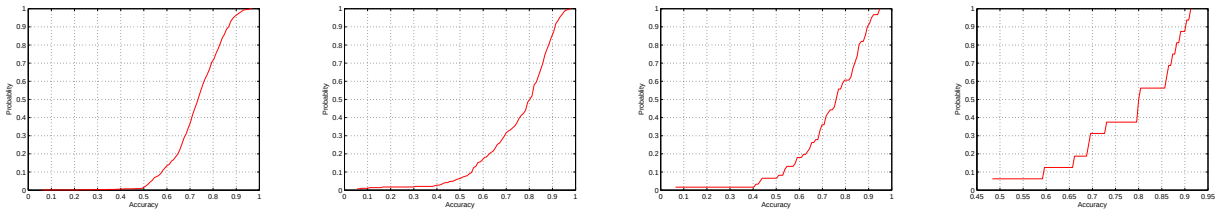
In addition, we also study the CDF of prediction accuracy for news stories in each group. Figures 4[a-d] show the CDF of prediction accuracy for 2-source, 3-source, 4-source and 5-source groups, respectively. Due to the small number of news stories, we do not include 6-source and 8-source groups here. As illustrated in Figure 4, the linear diffusion model achieves very high accuracy in predicting the density of influenced users. For example, for the news stories in 2-source group, the model exhibits 70% or higher prediction accuracy for nearly 65% news stories. Therefore, our experiment results indicate that the linear diffusion model can well predict the process of information diffusion for news stories initiated from multiple-sources over Digg network.

Table I
PREDICTION ACCURACY ON ALL NEWS STORIES AND THE MOST POPULAR NEWS STORIES FOR EACH GROUP

Group	Story	Accuracy	Total sews stories	Average accuracy
2-Source	s1	93.43%	1045	76.55%
3-Source	s2	93.69%	304	75.70%
4-Source	s3	92.61%	64	73.63%
5-Source	s4	90.97%	16	77.82%
6-Source	s5	94.49%	3	76.41%
8-Source	s6	95.55%	1	95.55%

V. CONCLUSION

As online social networks have played an increasingly important role in disseminating news stories, promoting news products and political campaigns thanks to their growing popularity, it becomes curtail to understand the patterns of information diffusion over these networks. Prior studies have extensively analyzed the process of infuriation diffusion for information initiated from a single source. This paper extends prior effort to characterize the diffusion process of information that are initiated simultaneously by two or more sources, since it is common to observe a popular news story, e.g, the final result of a popular sport game, to be posted by multiple fans at the same time, and then forwarded or retweeted by other users. We first introduce a



(a) 2-source (b) 3-source (c) 4-source (d) 5-source

Figure 4. CDF of prediction accuracy for 2-source, 3-source, 4-source and 5-source groups. The x axis represents the accuracy the model can achieve with new stories news, while the y axis represents the cumulative distribution of the accuracies for all news stories among the the group.

simple algorithm to extract 1433 news stories that are approximately initiated from multiple sources from Digg datasets. Subsequently, we analyze the diffusion patterns of these news stories from both temporal and spatial perspectives. In addition, we use the linear diffusion model proposed in prior studies to characterize and predict the information diffusion process of these news stories. The experiment results show that linear diffusion model is able to achieve an average of 75% prediction accuracy on the density of influenced users, and indicate that the model could effectively characterize and predict the process of information diffusion for news stories initiated from single sources as well as from multiple-sources. Our future work lies in understanding the diffusion patterns of controversial news stories, e.g., environmental issues or political debates, over online social networks.

REFERENCES

[1] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[2] G. Steeg, R. Ghosh, and K. Lerman, "What stops social epidemics?" *arXiv preprint arXiv:1102.1985*, 2011.

[3] M. Cha, A. Mislove, and K. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 721–730.

[4] S. Liu, L. Ying, and S. Shakkottai, "Influence maximization in social networks: An ising-model-based approach," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 570–576.

[5] R. Kumar, M. Mahdian, and M. McGlohon, "Dynamics of conversations," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 553–562.

[6] A. Goyal, F. Bonchi, and L. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 241–250.

[7] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[8] C. Budak, D. Agrawal, and A. El Abbadi, "Diffusion of information in social networks: Is it all local?" in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 121–130.

[9] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 599–608.

[10] R. Ghosh and K. Lerman, "A framework for quantitative analysis of cascades on networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 665–674.

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

[12] W. An, "Models and methods to identify peer effects," *The Sage Handbook of Social Network Analysis. London: Sage*, pp. 515–532, 2011.

[13] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Efficient discovery of influential nodes for sis models u in social networks," *Knowledge and information systems*, vol. 30, no. 3, pp. 613–635, 2012.

[14] T. Hogg and K. Lerman, "Social dynamics of digg," in *Proc. Int. Conference on Weblogs and Social Media (ICWSM10)*. Springer, 2010.

[15] F. Wang, H. Wang, and K. Xu, "Diffusive logistic model towards predicting information diffusion in online social networks," in *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on*. IEEE, 2012, pp. 133–139.

[16] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia, "Characterizing information diffusion in online social networks with linear diffusive model," in *Proceedings of International Conference on Distributed Computing Systems (ICDCS)*, 2013.