Multi-Source-Driven Asynchronous Diffusion Model for Video-Sharing in Online Social Networks

Guolin Niu, *Student Member, IEEE*, Xiaoguang Fan, Victor O.K. Li, *Fellow, IEEE*, Yi Long, and Kuang Xu

***Abstract*—Characterizing the video diffusion in online social networks (OSNs) is not only instructive for network traffic engineering, but also provides insights into the information diffusion process. A number of continuous-time diffusion models have been proposed to describe video diffusion under the assumption that the activation latency along social links follows a single parametric distribution. However, such assumption has not been empirically verified. Moreover, a user usually has multiple activated neighbors with different activation times, and it is hard to distinguish the different contributions of these multiple potential sources. To fill this gap, we study the multiple-source-driven asynchronous information diffusion problem based on substantial video diffusion traces. Specifically, we first investigate the latency of information propagation along social links and define the single-source (SS) activation latency for an OSN user. We find that the SS activation latency follows the exponential mixture model. Then we develop an analytical framework which incorporates the temporal factor and the influence of multiple sources to describe the influence propagation process. We show that one's activation probability decreases exponentially with time. We also show that the time shift of the exponential function is only determined by the most recent source (MRS) active user, but the total activation probability is the combination of influence exerted by all active neighbors. Based on these discoveries, we develop a multi-source-driven asynchronous diffusion model (MADM). Using maximum likelihood techniques, we develop an algorithm based on expectation maximization (EM) to learn model parameters, and validate our proposed model with real data. The experimental results show that the MADM obtains better prediction accuracy under various evaluation metrics.**

***Index Terms*—Asynchronous diffusion process, exponential mixture model, measurement, online social network.**

I. Introduction

THE rapid development of online social networks (OSNs), such as Facebook and Twitter, renders them a powerful tool for information propagation. In OSNs, information is propagated in the so called "word-of-mouth" format, which greatly reshaped the access patterns of multimedia contents, especially the video contents. These videos shared in OSNs are originally hosted by video sharing sites (VSSes), such as YouTube. As a result, while the video requests of VSSes, e.g., YouTube, have imposed great demand on the Internet [1], a large number of requests to these VSSes are actually generated by OSNs with distinct access patterns [2]. The distinct URLs of these videos shared in OSNs allow us to identify and track the diffusion path of each shared video easily and explicitly without conducting topic identification and tracking [3]. Therefore, characterizing the video diffusion in OSNs is not only instructive for network traffic engineering, but also provides insights into the information diffusion process.

Many information diffusion models have been proposed to describe the information diffusion process. Among these various models, two important models are the Threshold Model [4] and the Cascade Model [5], and their special cases, i.e., the Linear Threshold Model and the Independent Cascade Model, respectively, have been applied to solve the influence maximization problems [6]. The above models all assume that the diffusion proceeds synchronously in discrete unit time steps, whereas in reality information propagates continuously and the diffusion rate varies with time. To fill this gap, a number of continuous-time diffusion models have been proposed to describe the diffusion more accurately [7], [8], or to solve specific problems, such as the network inference problem [9], [10]. However, the basic question of these studies, namely, "How does the inter-personal influence vary with time?" has not been empirically and extensively studied, and all these studies try to answer this question with a single parametric model assumption, including the exponential [7], power-law, or Rayleigh distributions [10], [9]. Such an assumption has been suggested to be impractical recently [11]. Moreover, a user usually has multiple activated neighbors with different activation times, and it is hard to distinguish the different contributions of these multiple potential sources [12], [13].

Motivated by the above issues, we select Renren,[1] the most popular Facebook-like OSN in China, to be our research platform for empirical studies on the video diffusion process. This is because compared with the strict privacy policies of other OSNs such as Facebook, Renren not only maintains a complete list of individual user's video sharing actions and makes it publicly accessible by default, but also provides the friendship network of its users until April 2011. We have collected substantial video diffusion traces as well as the underlying social network topology from Renren, covering the sharing actions of around 2.8 million Renren users for more than 3 years. In this paper, we study the video-sharing actions in Renren under multiple active sources. According to both theoretical [4], [5] and empirical [14] studies, for User u_i, other

Manuscript received August 21, 2013; revised March 23, 2014 and June 13, 2014; accepted July 08, 2014. Date of publication July 17, 2014; date of current version October 13, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan.

The authors are with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong (e-mail: glniu@eee.hku.hk; xgfan@eee.hku.hk; vli@eee.hku.hk; yilong@eee.hku.hk; xukuang@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2340133

[1]http://www.renren.com

than the Most Recent Source (MRS) neighbor, any other friend who shared the same video before u_i may exert cumulative influence on u_i. Hence, to analyze the influence of multiple sources on the current user, in this work we started by studying a single active source problem, and then further present an analytical framework to study the effect of multiple influential sources. Particularly, we consider the inter-personal influence as a time-varying stochastic process and deduce the information diffusion process under multiple influential sources. Moreover, we develop a novel asynchronous diffusion model based on the deduced stochastic process.

Our main contributions are summarized as follows:

- We collect a substantial dataset from Renren and analyze several basic properties to support our subsequent studies on information diffusion.
- We conduct a detailed empirical study on the collected information diffusion traces. We define the single-source (SS) activation latency as the time interval between the activation time of the target user and that of his/her single activated neighbor. We find that SS activation latency is characterized by the exponential mixture model.
- In addition to the SS active user, other existing active sources could also exert influence on the target user. We develop an analytical framework to study how multiple influential sources affect the target inactive user. We find that, even under multiple influential sources, one's activation latency can still be characterized by the exponential mixture model.
- Based on the analytical results given above, we develop a general Multi-source-driven Asynchronous Diffusion Model (MADM) to describe the information propagation behavior in social networks and predict the individual activation time. It incorporates both the aforementioned temporal features and the heterogeneous influential source features.
- Last but not least, using maximum likelihood techniques, we develop algorithms based on Expectation Maximization (EM) to learn the model parameters and conduct a comprehensive set of experiments on large scale real world datasets to evaluate the performance of MADM compared with four other models using two evaluation metrics, and demonstrate that MADM has better prediction accuracy on user's activation time.

The rest of the paper is organized as follows. Section II introduces related works. Section III describes the statistics of our crawled datasets and also presents measurement results. In Section IV, we study the characteristics of the diffusion traces and the distribution of SS activation latency. In Section V, we develop an analytical framework of the asynchronous diffusion model. Section VI presents the experimental results. Finally Section VII is the conclusion and future work.

II. Related Work

In this section, we survey literature on VSSes studies and information diffusion in online social network.

A. Video Sharing Sites Studies

With the advent of Web 2.0 technology, there are increasing interests on the VSSes. As the most successful VSS, Youtube has become a popular research topic since its establishment in early 2005. Gill *et al.* [15] presents a traffic characterization study of YouTube. Cheng *et al.* [16] not only presents a systematic and in-depth measurement study on the statistics of YouTube videos, but also investigate the social networking in YouTube videos. Xie *et al.* [17] uses a social graph to model YouTube, with people and content as nodes and meme postings as links, to track large-scale video remix in Youtube.

However, the video access patterns are different between VSSes and OSNs. In particular, videos in VSSes are mainly viewed via related videos, search engines and front page [18], whereas videos in OSNs are viewed via friends sharings. Such differences in access pattern have profound impacts on the workload of VSSes, e.g., the word-of-mouth diffusion in OSNs amplifies the skewness of video popularity compared with VSSes [19]. Hence understanding the video propagation behaviors in OSNs is indispensable.

B. Information Diffusion Studies in OSNs

The existing information diffusion models can be classified into two categories [20]: the aggregate-level model and the agent-based model. Among the aggregate-level diffusion models, the Bass diffusion model [21] is very popular and quantitatively describes how new products get adopted as an interaction between existing and potential users. However, such aggregate-level diffusion models suffer from considerable inherent limitations, e.g., the inability to reflect population heterogeneity and poor explanatory power [20]. Agent-based models are more commonly used today [22], and two representative models are the Threshold Model [4] and Cascade Model [5]. However, existing models "are based on assumed rather than measured influence effects" [23]. Motivated by this gap, many empirical studies have been conducted on the information diffusion process by focusing on various media contents. Cha *et al.* [24] measured the photo propagation in Flickr, one of the most popular photo sharing social network, and found that the spreading was slow. Sun *et al.* [25] conducted an empirical investigation of influence diffusion in Facebook by defining the influence behavior accurately using explicit page fanning diffusion data. However, compared with other types of media, the diffusion dynamics of videos have been demonstrated to hold some unique features [19], e.g., the popularity of videos in OSNs decays faster than photos [24]. It is, therefore, worthwhile to study video diffusion in OSNs.

Multiple-source influence in information diffusion is a critical issue, and a series of data-driven methods have been proposed. Goyal *et al.* [13] introduced a "credit distribution" model to learn activation probabilities of different sources by directly leveraging propagation traces. [12] also investigated the influence credit assignment issue among multiple sources to compute influence among Twitter. [14] discussed the activation probability of information diffusion across different topics under one or more exposures (multiple sources). In addition to lacking empirical verification on their approaches to assign the influence among multiple sources, such studies also do not

consider how the multiple influence probabilities vary with time. Moreover, the asynchronous activation modes of users' multiple neighbours are not fully addressed. In this paper, we propose an asynchronous temporal diffusion model to account for these issues.

We are not the first to investigate the interplay of OSNs and VSSes by studying how videos originally hosted by VSSes are diffused in OSNs. Most recently, Li *et al.* [19] compared studies on the characteristics of video requests from OSNs and from VSSes, and proposed a model to describe the growth process of videos overall popularity in OSNs. Then Wang *et al.* [26] discussed the video recommendation issue in OSNs by designing a joint social-content recommendation framework for users to import or re-share videos. Distinct from these studies, the aim of our study is to empirically investigate the temporal characteristics of influence with multiple sources, rather than to model videos' overall popularity [19] or to enhance the video recommendations [26].

III. Preliminaries

A. Dataset

The dataset in this study is collected from Renren, one of the most popular online social websites in China with more than 160 million registered users. It shares almost the same features, structure and layout as Facebook. Users could maintain their own profiles, photo galleries and blogs, and establish bidirectional friendship links with other users. However, two unique characteristics of Renren make it extremely attractive for studying information diffusion over OSNs. Firstly, unlike Facebook, the friendship relationship in Renren is accessible before April 2011, which enables us to acquire the topological data to create a real world social network graph. Secondly, and perhaps more importantly, well-organized information diffusion data in Renren is public to any registered users. This allows us to crawl substantial diffusion traces to study how information spreads temporally and spatially.

B. Mechanics of Renren Information Diffusion

Information diffusion in Renren is realized by "sharing" events, which includes sharing friends' notes, photos and video links. Diffusion of shared information occurs as follows. Firstly, user A shares a piece of information, which is broadcast to his entire friend list. Upon receiving this information, one or more of his friends may decide to share the information as well. Thus, we say that information is propagated from user A to some of his friends.

To study how information diffuses in Renren, we concentrate on the propagation of one particular event, namely, "sharing video." Since the video shared in Renren mainly come from several external video websites (see Table I, which shows, using statistics collected from our crawled dataset, that over 92% of all URLs accessed are from the top five websites), and the external video URLs can be used to identify different topics (videos) directly. There is no need to conduct text mining and topic identification, which is time consuming and probably inaccurate. Meanwhile, Renren provides the timestamp of when a user shares a video. This allows us to study how the propagation of videos evolves over time.

TABLE I
Statistics of External Websites in Renren

Rank	Video website	Fraction of videos accessed from web site
1	www.youku.com	50.07%
2	www.tudou.com	25.13%
3	www.ku6.com	6.44%
4	www.56.com	5.06%
5	www.youtube.com	3.76%
5	video.sina.com.cn	2.98%
6	Others (>25 websites)	6.56%

C. Crawling Methods

We design the following two crawlers to collect both user information data and the diffusion traces.

- The "user information crawler" analyzes the user profile webpage in order to acquire this user's unique user identities, affiliation information and friend list.
- The "video information crawler" obtains the given user's shared video information by parsing the user's "video-sharing" page.

Although it is computationally expensive and nearly impossible for us to acquire the entire Renren network topology, fortunately, just like Facebook, Renren evolves from a university-based social network and hence is divided into regional networks with affiliation information that represents universities or institutions. Therefore, following previous studies [27][28], we plan to focus our research on a university-based regional network in Renren as well. Moreover, as has been demonstrated by Choudhury *et al.* [29], our method, which incorporates both network topology and user-context, like the affiliation information, is better able to capture the information diffusion characteristics compared with a pure topology-based sampling method. Specifically, we perform affiliation-oriented crawls of the Xian Jiaotong University (XJTU) network, a famous college online social network community in Renren. Firstly, we use the "browse user" function to obtain the 50 most popular users (in terms of number of friends) in the XJTU community. Then, we seed the "user information crawler" with the 50 most popular users to preform an exhaustive crawling of the XJTU users as well as their friendship links. Finally, we obtain a connected social graph of XJTU users as well as their 1-hop friends, and it is not surprising that the social graph which has been collected from the 50 most popular seed users is a connected graph.

For users in the social graph, we further employ a "video information crawler" to collect the information of their shared videos, including the sharing time (namely, the activation time), video URL, video title, etc. To avoid the limitation of request frequency from a unique IP on the Renren server, the crawlers work in a parallel mode on more than 10 computers. For simplicity, in the rest of the paper we use the "XJ dataset" to represent data collected from the XJTU community.

D. Data Description

For the XJ dataset, the topology information is just a snapshot of a particular timestamp, and the diffusion traces do cover a long time period. The topology snapshot in the XJ dataset

TABLE II
PROPERTIES OF THE VIDEO DATASET

	XJ Dataset
Time period	March 2008 - July 2011
Nodes	2,808,681
Sharing events	209,409,778
Distinct URLs	6,416,745

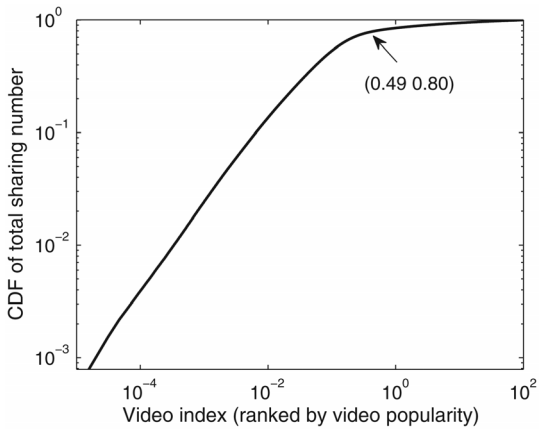


Fig. 1. Video popularity distribution.

is collected in February 2011. It includes 2,808,681 nodes and 52,685 of them are affiliated with XJTU, and the rest are friends of these XJTU affiliates. Then, for all nodes, we further collect their shared video information (see Table II). Although the friendship data is a snapshot at a particular time, the video information covers 1224 days in total, from March 2008 to July 2011. The reason is that each user's "video-sharing" page keeps a visible record of all his/her sharing history since this user account is created in Renren.

IV. EMPIRICAL STUDIES IN INFORMATION DIFFUSION

In this section, to capture the macro-level properties of the sharing events in Renren, we firstly analyze the video popularity. Then, we study the micro-level properties of the sharing events by investigating the inter-personal "activation latency" characteristics. We find that, in the macro-level, the diffusion behavior displays some common features already discovered in traditional social network studies; in the micro-level, the temporal properties could instruct us in building a more comprehensive and realistic diffusion model.

A. Video Popularity

The video popularity for a certain Video v, denoted $k(v)$, is defined as the number of users who have shared Video v. We study the popularity distributions for the XJ dataset. The results are compiled from over 6.4 million shared videos. The Cumulative Distribution Function (CDF) of the video popularity is shown in Fig. 1. Specifically, the horizontal axis of Fig. 1 ranks the videos from the most active ones to the least active ones, with the rank normalized between 0 and 100. Due to the high skewness of the data, we use the log-log scale to better display the figure. As the coordinate (0.49, 0.8) shows, the top 0.49% of the popular videos account for nearly 80% of video sharings.

TABLE III
THE POPULARITY OF THE TOP 5 VIDEOS

Video Rank	1	2	3	4	5
Popularity	158053	148340	131094	107897	100832

Such high skewness (in contrast with the top 8% of the popular videos accounting for 80% of video views in YouTube [30]) is mainly because the word-of-mouth diffusion in OSNs amplifies the skewness of video popularity, and hence demonstrates the uniqueness of video access pattern in OSNs. In particular, from the statistics in Table III, we observe that each top video enjoys a high popularity of over 100,000. Motivated by the high skewness of video popularity distribution, in addition to modeling the diffusion process for all videos, we will also specifically focus on the most popular videos to explore the influence of video popularity on the resultant model parameters in Section VI-A.

B. Influence Path

We define the influence path for a social network as follows. The Renren social graph can be represented in the form of an undirected graph $G = (V, E)$, where $V = \{u_1, u_2, \ldots, u_N\}$ is a set of nodes (users), and u_i is the unique ID of User u_i, and $E = \{\{u, v\} | u, v \in V\}$ is a set of edges. Note that friendship is mutual in Renren, so the relationship between User u and User v is bidirectional, which means $\{u, v\}$ is equivalent to $\{v, u\}$. Besides, we use T_C to denote the timestamp when we acquired the network topology, i.e, the timestamp of the topology snapshot.

After describing the social graph, we now focus on the video-sharing events for each user. Besides gathering a set of video objects $O = \{o_1, o_2, \ldots, o_m\}$ which have been shared among users in V, for any video object $o_m \in O$, we also collect the users who have shared Video o_m in the following set $P_m = \{(u_1, t_1), (u_2, t_2), \ldots, (u_k, t_k)\}$, where t_i is the time that user u_i shared the video object o_m, and is called the sharing time or activation time interchangeably.

For any two different elements (u_i, t_i) and (u_j, t_j), we assume that User u_j has influenced User u_i for sharing video Object o_m if and only if:

- User u_i and User u_j are friends, namely, $\{u_i, u_j\} \in E$;
- User u_j shared o_m earlier than u_i, namely, $t_j \leq t_i$;
- Note that video-sharing activity can only propagate along existing friendship links, so we extract sharing records which happen after the link is created, namely, $t_j \geq T_C$;
- The influence can only take effect within certain time interval θ, namely, $t_i - t_j \leq \theta$. θ can also be referred to as the Influence Effectiveness Time Window (IETW).

For any $(u_i, t_i) \in P_m$, we further collect the sharing times of users who are neighbors of User u_i for video object o_m to define the influence path:

$$TS_i^m = \{t_j | (u_i, t_i), (u_j, t_j) \in P_m; \{u_i, u_j\} \in E; t_j \geq T_C; 0 < t_i - t_j \leq \theta\} \quad (1)$$

We define the MRS activation latency of User u_i for video object o_m as: $t_i - \max(TS_i^m)$. For $|TS_i^m| = 1$, the MRS activation latency is reduced to a special case, which is defined as

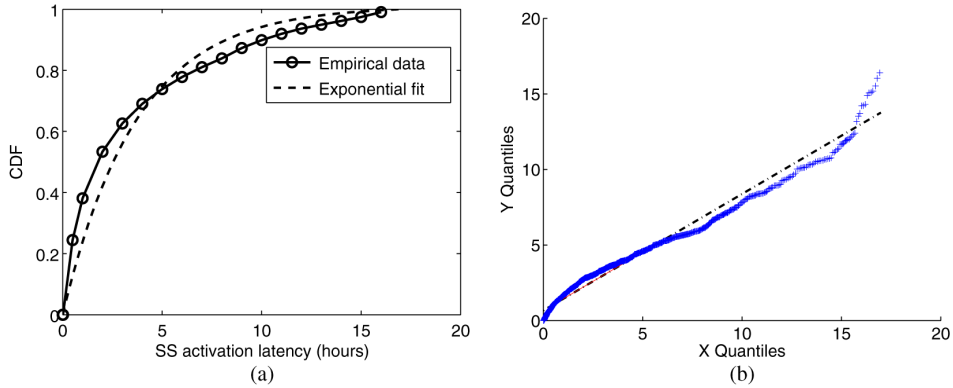


Fig. 2. Empirical data and exponential fitting.

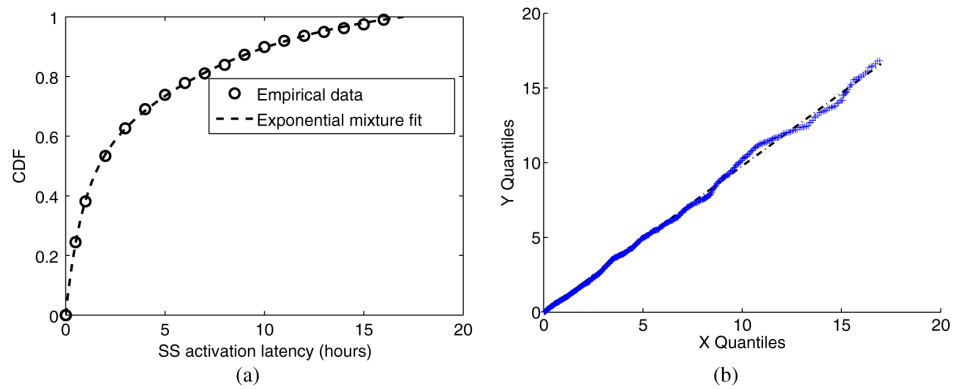


Fig. 3. Empirical data and exponential mixture distribution fitting.

the Single Source (SS) activation latency. In SS sharing, only one neighbour has shared the video.

In our data, we also find some zero-source sharing actions[2], which means that the corresponding user shares a video independently, i.e., without being influenced by his/her neighbours. As revealed by [31], such zero-source sharing actions are mainly driven by some external events, such as the mass media. In our studies, zero-source cases are excluded since we focus on modeling information diffusion within the OSN.

C. SS Activation Latency

What is the distribution of the SS activation latency? To answer this question, we present a large scale study for the aggregate propagation of all the crawled 6,416,745 distinct videos. In our previous work [8], we just analyze the MRS activation latency by assuming that sharing activities may happen within any time interval, so we did not set any time limit. In this work, we conduct a more detailed research on the SS activation latency using the parameters discussed in Section IV-B, T_C and θ, which limits the activation latency within the IETW θ and omits the sharing actions before T_C. Then we further study the case for multiple active sources.

1) Exponential Distribution Fitting: Fig. 2(a) shows the Cumulative Distribution Function (CDF) of the SS activation latency with the exponential distribution fitting. To provide goodness-of-fit, we use the Q-Q plot (Fig. 2(b)) to compare the empirical data and the theoretical distribution. Although the exponential model could reflect the distribution trend to some extent, the fitting result is not perfect.

2) Exponential Mixture Distribution Fitting: Since the SS activation latency cannot be fully described using pure exponential distribution, we adopt the exponential mixture model. Fig. 3(a) shows the CDF of the SS activation latency and the exponential mixture model fitting. Also, we use the Q-Q plot (Fig. 3(b)) to show the goodness-of-fit result. In probability and statistics, a mixture distribution [32] is a probability density function of the form: $f(x) = \sum_{k=1}^{K} \alpha_k f(x;\theta_k)$. Here, K is the number of components in the mixture model. For each k, $f(x;\theta_k)$ is the Probability Density Function (PDF) of the component number k. The scalar α_k is the proportion of component number k. In order for the mixture model to be a proper PDF, it must be the case that $\sum_k \alpha_k = 1$, where $\alpha_k \geq 0$ for all k. In this paper, we adopt the exponential mixture model, where $f(x;\theta_k)$ is an exponential distribution, and $K = 2$. The detailed formulation is

$$f(x) = \alpha_1 f(x;\theta_1) + \alpha_2 f(x;\theta_2) = \alpha_1\theta_1 e^{-\theta_1 x} + \alpha_2\theta_2 e^{-\theta_2 x}, \quad \alpha_1 + \alpha_2 = 1; \alpha_1 \geq 0, \alpha_2 \geq 0 \tag{2}$$

To provide rigorous goodness-of-fit result for the given distributions, we use the Kolmogorov-Smirnov test (K-S test) (see Table IV) to compare the empirical data and the theoretical distribution for the pure exponential model and the exponential mixture model. The results show that the exponential mixture model could explain the SS activation latency better.

Many different models have been adopted in modeling diffusion networks and social networks in the literature, including ex-

[2]In our dataset, the percentage of zero-source sharings, single-source sharings, and multiple-source sharings are 20%, 37%, and 43%, respectively.

TABLE IV
K-S TEST RESULT

Fitting type	Significance level α	p value	Result h
Exponential Mixture Model	0.05	0.4475	0 (accept)
Exponential Model	0.05	0	1 (reject)

ponential distribution, power law distribution and Weibull distribution [33][34]. We are the first to utilize the exponential mixture model to describe the interpersonal diffusion latency along social links. Why does the mixture model perform well? One intuitive explanation is as follows. Users are composed of two categories, i.e., active users and inactive users [35]. Active users are those who share videos more frequently, and inactive users are those who share videos less frequently. In these two groups, the average activation latencies will also be different, corresponding to the different parameters of the two components in our proposed exponential mixture model.

Meanwhile, due to the simplicity of the pure exponential fitting, we will explore its performances in the following sections. Under certain circumstances, it may provide a good approximation, and reduce the computational cost.

V. MULTI-SOURCE-DRIVEN ASYNCHRONOUS DIFFUSION MODEL

When studying the behaviour of a target user in information diffusion, it is common that the target user has multiple neighbours who all try to exert influence on his/her decision. Hence to model the diffusion process properly, we will derive a Multi-source-driven Asynchronous Diffusion Model (MADM). However, this is challenging since it is hard to distinguish the influences exerted by these multiple sources [14] [13] and to determine the activation latency accordingly. To overcome this challenge, rather than trying to measure the activation latency with multiple sources directly, we propose to deduce the activation latency with multiple sources from its special case, namely, the SS activation latency, which can be measured easily and explicitly as shown in Section IV-C. In this section, we not only illustrate the derivation, but also validate the derived conclusion.

A. Multi-Source Influence

According to the empirical results shown in Section IV, the activation latency under the single source influence scheme follows the exponential mixture distribution. In other words, the probability that a node is activated by its single active source neighbor decreases exponentially with time. Based on this significant observation, we try to further solve the multi-source influence problem by utilizing the time-shift linear superposition of asynchronous influences. Consider an inactive Node D, in a single source influence problem, Node D only has one active neighbor, say Node A. Suppose the total probability that Node D is activated by Node A is denoted as P_{AD}. This is the influential power of Node A on Node D, and obviously $P_{AD} \leq 1$. Next we try to integrate the continuous time factor to see how this probability is distributed across time. We use a probability density function $f_{AD}(t)$ to show the activation probability from Node A to Node D across time. Based on our empirical findings in Section IV, $f_{AD}(t)$ is defined as follows:

$$f_{AD}(t) = P_{AD}(\alpha_1\lambda_1 e^{-\lambda_1(t-t_A)} + \alpha_2\lambda_2 e^{-\lambda_2(t-t_A)}), \quad t \geq t_A; \alpha_1+\alpha_2=1; \alpha_1; \alpha_2 \geq 0 \tag{3}$$

where t_A is activation time of Node A, α_1 and α_2 are the parameters to control weights of the two exponential components, and the corresponding decreasing rates of the two components are represented by λ_1 and λ_2, respectively. Note that $\int_{t_A}^{\infty} f_{AD}(t)dt = P_{AD}$, which satisfies that the total probability equals P_{AD}. Next we discuss the two-source scenario in Fig. 4.

Suppose Node A is activated at time t_A, and starts to activate its neighbor Node D with the probability expressed in Equation (3). At time $t_B (t_B > t_A)$, another neighbor of Node D, namely Node B, is activated, but Node D has not been activated yet. Thus it has two influence sources A and B. The influences on Node D by A and B decrease exponentially with time shifts t_A and t_B, respectively

$$f_{AD}(t) = P_{AD}(\alpha_1\lambda_1 e^{-\lambda_1(t-t_A)} + \alpha_2\lambda_2 e^{-\lambda_2(t-t_A)}), \quad t \geq t_A; \alpha_1+\alpha_2=1; \alpha_1; \alpha_2 \geq 0 \tag{4}$$

$$f_{BD}(t) = P_{BD}(\alpha_1\lambda_1 e^{-\lambda_1(t-t_B)} + \alpha_2\lambda_2 e^{-\lambda_2(t-t_B)}), \quad t \geq t_B; \alpha_1+\alpha_2=1; \alpha_1; \alpha_2 \geq 0 \tag{5}$$

If the activation probability density function on Node D is noted by $f_D(t)$, we formulate it by the linear superposition of $f_{AD}(t)$ and $f_{BD}(t)$. It is shown in Equation (6) at the bottom of the page.

Here, P_D is the total probability that D is activated by the combined influential power of Node A and B. To simplify Equation (6), we introduce two symbols $\alpha_1{}'$ and $\alpha_2{}'$ as shown in Equations (7) and (8).

$$\alpha_1{}' = \frac{\alpha_1[P_{AD}e^{-\lambda_1(t_B-t_A)} + P_{BD}]}{P_{AD}[\alpha_1 e^{-\lambda_1(t_B-t_A)} + \alpha_2 e^{-\lambda_2(t_B-t_A)}] + P_{BD}} \tag{7}$$

$$\begin{aligned} f_D(t) &= P_D \cdot \frac{f_{AD}(t) + f_{BD}(t)}{\int_{t_B}^{\infty}(f_{AD}(t') + f_{BD}(t'))dt'} \\ &= P_D \cdot \frac{P_{AD}[\alpha_1\lambda_1 e^{-\lambda_1(t-t_A)} + \alpha_2\lambda_2 e^{-\lambda_2(t-t_A)}] + P_{BD}[\alpha_1\lambda_1 e^{-\lambda_1(t-t_B)} + \alpha_2\lambda_2 e^{-\lambda_2(t-t_B)}]}{\int_{t_B}^{\infty}[P_{AD}(\alpha_1\lambda_1 e^{-\lambda_1(t'-t_A)} + \alpha_2\lambda_2 e^{-\lambda_2(t'-t_A)}) + P_{BD}(\alpha_1\lambda_1 e^{-\lambda_1(t'-t_B)} + \alpha_2\lambda_2 e^{-\lambda_2(t'-t_B)})]dt'} \\ &= P_D \cdot \frac{\alpha_1\lambda_1 e^{-\lambda_1(t-t_B)}[P_{AD}e^{-\lambda_1(t_B-t_A)} + P_{BD}] + \alpha_2\lambda_2 e^{-\lambda_2(t-t_B)}[P_{AD}e^{-\lambda_2(t_B-t_A)} + P_{BD}]}{P_{AD}[\alpha_1 e^{-\lambda_1(t_B-t_A)} + \alpha_2 e^{-\lambda_2(t_B-t_A)}] + P_{BD}}, \qquad t > t_B \end{aligned} \tag{6}$$

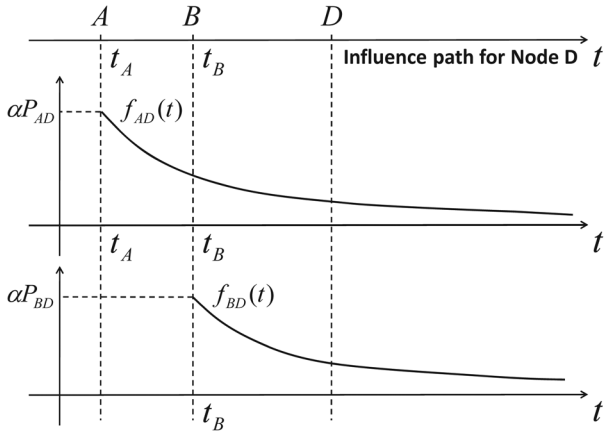


Fig. 4. Two-source influence problem.

and

$$\alpha_2{}' = \frac{\alpha_2[P_{AD}e^{-\lambda_2(t_B-t_A)} + P_{BD}]}{P_{AD}[\alpha_1 e^{-\lambda_1(t_B-t_A)} + \alpha_2 e^{-\lambda_2(t_B-t_A)}] + P_{BD}} \quad (8)$$

Equation (6) can then be written as:

$$f_D(t) = P_D[\alpha_1{}'\lambda_1 e^{-\lambda_1(t-t_B)} + \alpha_2{}'\lambda_2 e^{-\lambda_2(t-t_B)}], \qquad t \geq t_B \quad (9)$$

Note that $\alpha_1{}' + \alpha_2{}' = 1$. So we draw the conclusion that for the two-source case, the activation latency still follows the mixture model. And P_{AD}, P_{BD}, P_D can be easily derived from the parameters in the traditional LTM or ICM. In LTM, the influential power between a node pair is described as the edge weight. Thus P_{AD} and P_{BD} are the influence weights from Node A and Node B, respectively, to Node D. P_D should be the weight summation from all active neighbors of Node D, and thus is calculated as $P_{AD} + P_{BD}$. In ICM, the influential power is described as probability. Then P_{AD} and P_{BD} are the activation probabilities from Node A and Node B, respectively, to Node D and the total probability can be represented as: $P_D = 1 - (1 - P_{AD})(1 - P_{BD})$.

For the three-source case, with Node A, Node B and Node C activated at time t_A, t_B and t_C respectively ($t_A < t_B < t_C$), what is the activation probability density function for Node D? Following the same procedure as in the two-source case, the result under the three-source scenario is shown in Equation (10) at the bottom of the page. And the corresponding parameters are shown in Equation (11) and Equation (12) at the bottom of the page.

Following the same idea, we can derive the activation probability function for the multi-source influence problem. Suppose we wish to calculate the activation probability of Node u_i at time t, the first step is to find the neighbors of Node u_i activated before t. It includes all the active neighbors of Node u_i denoted $\{u_1, u_2, \ldots, u_j\}$, and their corresponding activation times, denoted $\{t_{u_1}, t_{u_2}, \ldots, t_{u_j}\}, t_{u_1} < t_{u_2} < \ldots < t_{u_j}$. Thus the activation probability function of Node u_i at time t could be expressed as follows:

$$f_{u_i}(t) = P_{u_i}[\Psi_1'\lambda_1 e^{-\lambda_1(t-\tau)} + \Psi_2{}'\lambda_2 e^{-\lambda_2(t-\tau)}], \qquad t \geq \tau \quad (13)$$

where

$$\tau = t_{u_j} \quad (14)$$

$$\Psi_1{}' = \frac{\alpha_1 \sum_{r=1}^{j} P_{u_r u_i} e^{-\lambda_1(t_{u_j} - t_{u_r})}}{\sum_{r=1}^{j} P_{u_r u_i}\left[\alpha_1 e^{-\lambda_1(t_{u_j} - t_{u_r})} + \alpha_2 e^{-\lambda_2(t_{u_j} - t_{u_r})}\right]} \quad (15)$$

$$\Psi_2{}' = \frac{\alpha_2 \sum_{r=1}^{j} P_{u_r u_i} e^{-\lambda_2(t_{u_j} - t_{u_r})}}{\sum_{r=1}^{j} P_{u_r u_i}\left[\alpha_1 e^{-\lambda_1(t_{u_j} - t_{u_r})} + \alpha_2 e^{-\lambda_2(t_{u_j} - t_{u_r})}\right]} \quad (16)$$

$$\Psi_1' + \Psi_2' = 1 \quad (17)$$

P_{u_i} has the following expressions:

$$P_{u_i} = \sum\nolimits_{r=1}^{j} P_{u_r u_i} \quad \text{(for LTM)} \quad (18)$$

$$P_{u_i} = 1 - \prod\nolimits_{r=1}^{j}(1 - P_{u_r u_i}) \quad \text{(for ICM)} \quad (19)$$

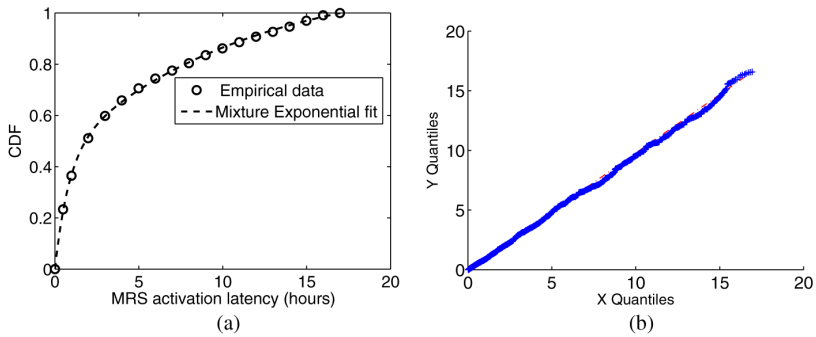
From Equation (13), we find several important results:

- The multi-source influence on Node u_i, represented by activation probability density function, decreases exponentially with time.
- The time shift on the exponential mixture model is τ. It corresponds to the MRS neighbor of Node u_i. Thus the shift is only determined by the MRS neighbor, but not other neighbors that are activated earlier.
- The weights for the two different components are controlled by all existing active sources, i.e., their activation times may influence the proportions of the two components, Ψ_1' and Ψ_2'.
- Note that $\int_{\tau}^{\infty} f_{u_i}(t)dt = P_{u_i}$, and the total probability P_{u_i} is not only determined by the MRS neighbor, but the combination of probabilities from all the neighbors activated before User u_i.

$$f_D(t) = P_D[\beta_1{}'\lambda_1 e^{-\lambda_1(t-t_C)} + \beta_2{}'\lambda_2 e^{-\lambda_2(t-t_C)}], \qquad t \geq t_C, \beta_1{}' + \beta_2{}' = 1 \quad (10)$$

$$\beta_1{}' = \frac{\alpha_1[P_{AD}e^{-\lambda_1(t_C-t_A)} + P_{BD}e^{-\lambda_1(t_C-t_B)} + P_{CD}]}{P_{AD}[\alpha_1 e^{-\lambda_1(t_C-t_A)} + \alpha_2 e^{-\lambda_2(t_C-t_A)}] + P_{BD}[\alpha_1 e^{-\lambda_1(t_C-t_B)} + \alpha_2 e^{-\lambda_2(t_C-t_B)}] + P_{CD}} \quad (11)$$

$$\beta_2{}' = \frac{\alpha_2[P_{AD}e^{-\lambda_2(t_C-t_A)} + P_{BD}e^{-\lambda_2(t_C-t_B)} + P_{CD}]}{P_{AD}[\alpha_1 e^{-\lambda_1(t_C-t_A)} + \alpha_2 e^{-\lambda_2(t_C-t_A)}] + P_{BD}[\alpha_1 e^{-\lambda_1(t_C-t_B)} + \alpha_2 e^{-\lambda_2(t_C-t_B)}] + P_{CD}} \quad (12)$$



(a) (b)

Fig. 5. Empirical data and exponential mixture distribution fitting for MRS activation latency.

TABLE V
K-S TEST RESULT FOR MRS ACTIVATION LATENCY FITTING

Fitting type	Significance level α	p value	Result h
Exponential Mixture Model	0.05	0.1177	0 (accept)

B. MRS Activation Latency

However, the derivation in Section V-A is based on the "independent user activation" assumption, and hence needs to be further verified in case this assumption may not hold. According to our derived results, the activation probability of target users is only determined by the MRS activation latency, which can be better characterized by the exponential mixture distribution. Therefore, we will empirically verify the distribution of MRS activation latency.

Fig. 5(a) shows the Cumulative Distribution Function (CDF) of the MRS activation latency and the exponential mixture model fitting. To provide goodness-of-fit for the given distribution, we use the Q-Q plot (Fig. 5(b)) to compare the empirical data and the theoretical distribution. To provide rigorous goodness-of-fit result for the given distributions, we also show the Kolmogorov-Smirnov test (K-S test) (see Table V) result. It shows that the MRS activation latency can be fitted well by the exponential mixture model, which accords with our derived results.

C. Multi-Source Influence Under the Exponential Model Approximation

In the SS activation latency measurement studies (Section IV-C), we found that although the exponential model cannot fully describe the SS activation latency, it could still provide a good approximation. Hence, in this section, we consider the exponential model, in which the SS activation latency follows the single exponential distribution. We will firstly focus on the two-source scenario as shown in Fig. 4, and then target the general multi-source influence problem.

We follow the same setting in Section V-A in which Node A and Node B are activated at time t_A and t_B respectively, with $(t_B > t_A)$. Their neighbor Node D has not been activated yet. Thus it has two influence sources A and B. The influences on Node D from A and B separately decrease exponentially with time shifts t_A and t_B, respectively. If the total activation probability density function on Node D is denoted by $g_D(t)$, we formulate it by the linear superposition of $g_{AD}(t)$ and $g_{BD}(t)$:

$$\begin{aligned} g_D(t) &= P_D \cdot \frac{g_{AD}(t) + g_{BD}(t)}{\int_{t_B}^{\infty}(g_{AD}(t') + g_{BD}(t'))dt'} \\ &= P_D \cdot \frac{\alpha P_{AD} e^{-\alpha(t-t_A)} + \alpha P_{BD} e^{-\alpha(t-t_B)}}{\int_{t_B}^{\infty}(\alpha P_{AD} e^{-\alpha(t'-t_A)} + \alpha P_{BD} e^{-\alpha(t'-t_B)})dt'} \\ &= P_D \cdot \frac{\alpha e^{-\alpha(t-t_B)}(P_{AD} e^{-\alpha(t_B-t_A)} + P_{BD})}{P_{AD} e^{-\alpha(t_B-t_A)} + P_{BD}} \\ &= \alpha P_D e^{-\alpha(t-t_B)}, \qquad t > t_B \end{aligned} \tag{20}$$

$$g_{AD}(t) = \alpha P_{AD} e^{-\alpha(t-t_A)}, \qquad t \geq t_A \tag{21}$$

$$g_{BD}(t) = \alpha P_{BD} e^{-\alpha(t-t_B)}, \qquad t \geq t_B \tag{22}$$

Here P_{AD}, P_{BD}, P_D follows the same definition as shown in Section V-A, which can be easily derived from the parameters in the traditional LTM and ICM model.

Following the same idea, we can derive the activation probability function for the general multi-source influence problem. Suppose we wish to calculate the activation probability of Node u_i at time t, we first find all the neighbors of Node u_i activated before t. It includes all the active neighbors of Node u_i denoted $\{u_1, u_2, \ldots, u_j\}$, and their corresponding activation times, denote $\{t_{u_1}, t_{u_2}, \ldots, t_{u_j}\}, t_{u_1} < t_{u_2} < \ldots < t_{u_j}$. Thus the activation probability density of Node u_i at time t could be expressed as follows:

$$g_{u_i}(t) = \alpha P_{u_i} e^{-\alpha(t-\tau)}, \qquad t \geq \tau \tag{23}$$

where $\tau = t_{u_j}$, and P_{u_i} has the following expressions:

$$P_{u_i} = \sum_{r=1}^{j} P_{u_r u_i} \quad \text{(for LTM)} \tag{24}$$

$$P_{u_i} = 1 - \prod_{r=1}^{j}(1 - P_{u_r u_i}) \quad \text{(for ICM)} \tag{25}$$

One important finding is that in the general formulation Equation (23), only the MRS activation latency affects the activation probability of the target user. Compared with the exponential mixture model, this largely reduces the computational cost. This is because in the exponential mixture model, we have to utilize all existing activated users' activation time records. As to the quantitative performance of the single exponential assumption, we will explore this in the model validation in Section VI-B.

D. Diffusion Model

To account for the aforementioned cumulative influence from the multiple active sources and the temporal information propagation feature, we propose a new diffusion model, namely, the Multi-source driven Asynchronous Diffusion Model (MADM). The specifics of MADM are as follows.

- Initially, a social graph $G = (V, E)$ is given, where V represents the set of nodes, and E is the set of edges. A group of seed nodes is initially activated.
- Time proceeds continuously. Any inactive node will be activated with the following probability function which varies with time:

$$f_{u_i}(t) = P_{u_i}[\Psi_1{}'\lambda_1 e^{-\lambda_1(t-\tau)} + \Psi_2{}'\lambda_2 e^{-\lambda_2(t-\tau)}], \quad t \in [\tau, \tau+\theta] \tag{26}$$

  where P_{u_i} is the total probability that node u_i is activated by the combined influential power of all of $u_i{}'s$ active sources, τ is the activation time of $u_i{}'s$ MRS neighbor, θ is the IETW, $\Psi_1{}'$ and $\Psi_2{}'$ can be calculated using existing activation time records as shown in Equation (15) and Equation (16) and λ_1 and λ_2 are the exponential component parameters.
- If any inactive node is activated at time t, update $f_{u_i}(t)$ for all the inactive nodes.
- The process ends if no more activation is possible.

The computational complexity of MADM includes two parts, i.e., model training and model validation. The first part is due to parameter learning, i.e., the computational cost of the EM algorithm. A rigorous proof of the finite convergence of the EM algorithm is given in [36]. According to the derived results, once we learned the required parameters under the explicitly measurable SS cases, we can further derive the parameters for multiple cases. Then the complexity of the EM algorithm performed on a training set is $O(T \times N)$, where N is the number of SS sharing actions and T is the maximum number of iterations. For the model validation part, we need to calculate the activation probability for all the sharing actions in the testing set following Equation (13). For each sharing action, the computation of Equation (13) will further depend on Equation (15) and Equation (16), and both of them need to traverse the active sources of this sharing action once. Accordingly, when applying the trained model on a testing set of M sharing actions, where each sharing action is, on average, driven by $\bar{d}$ active sources, the complexity is $O(M \times \bar{d})$. However, the average number of active sources $\bar{d}$ can be regarded as a fixed constant for a given dataset, and is usually much smaller than the average number of users' friends. In particular, $\bar{d}$ equals to 2.71 in our dataset, much smaller than the average number of friends which is 89. Therefore, the model validation complexity for MADM is actually $O(M)$. Thus the total complexity of MADM is the combination of these two parts, i.e., $O(M+T\times N)$. Moreover, in our following experiments, we set $T = 500$ according to previous studies [37], which is negligible compared with the values of M and N in the millions. Thus the computational complexity of training and testing MADM based on our experimental settings is $O(M + N)$, which grows linearly with the number of sharing actions investigated.

VI. Experiments

A. Model Training

Using our proposed MADM, given any User u_i, its activation probability can be calculated using a closed-form equation $f_{u_i}(t)$. Since User $u_i{}'s$ existing neighbor's activation time records are known parameters, we have to find a way to train the four unknown parameters in the exponential mixture model: α_1, α_2, λ_1 and λ_2. α_1 and α_2 are the mixing parameters, determining the proportion of the two exponential components. λ_1 and λ_2 are the inversion of means of the two exponential distributions, i.e., $\lambda_1 = \frac{1}{\mu_1}$ and $\lambda_2 = \frac{1}{\mu_2}$.

The expectation-maximization (EM) algorithm is an iterative procedure for model parameter estimation. It starts with any initial guesses for the values of the parameters, and then proceeds to repeat two phases that are called the expectation step (E-step), and the maximization step (M-step). The EM algorithm is the most widely adopted method for maximum-likelihood estimation of the parameters of a mixture distribution. In his work on distributions from exponential families, Hasselbald [38] derived the E-step and M-step for finite mixtures. Also, he compared the performance of the EM algorithm with another estimation method, the method of moments. Although it is difficult to draw definitive conclusions from his small study, his examples routinely showed that the EM algorithm produce estimates with smaller variances than the method of moments estimates.

Algorithm 1 EM Learning Algorithm

Require:
  Training set: $TS_1, TS_2, \ldots, TS_i, \ldots, TS_N$
Ensure:
  The final output parameters $\alpha_1, \alpha_2, \lambda_1, \lambda_2$
  Define SS activation latency set S
  for $TS_i = TS_1$ to TS_N **do**
    if $|TS_i| == 1$**then**
      Add TS_i to set S
    end if
  end for
  Initialize parameters:
  $\alpha_1 = 0.5, \alpha_2 = 0.5, \lambda_1 = 1, \lambda_2 = 1$
  $maxIter = 500, curIter = 0, threshold = 0.01,$
  $newL = -Inf, oldL = 2 * newL$
  while $curIter \leq maxIter$ **do**
    $[newL, W_1, W_2]$ = **Expectation**$(S, \alpha_1, \alpha_2, \lambda_1, \lambda_2)$
    $[\alpha_1, \alpha_2, \lambda_1, \lambda_2]$ = **Maximization**(S, W_1, W_2)
    if $|newL - oldL| < threshold$ **then**
      break
    end if
    $oldL = newL$
    $curIter++$
  end while
  return $\alpha_1, \alpha_2, \lambda_1, \lambda_2$

Hence, we also adopt the EM algorithm for estimation of our exponential two-mixture distribution. In each expectation step,

the probabilities that any given observation belongs to a particular exponential component are estimated using the currently fitted distribution parameters. In the maximization step, all the data incorporating the distribution parameters are re-maximized using the new estimates of the probabilities from the previous E-step. The details of the EM algorithm are described in Algorithm 1, incorporating the E-step (Algorithm 2) and the M-step (Algorithm 3).

Algorithm 2 Expectation Step

Require:
$S, \alpha_1, \alpha_2, \lambda_1, \lambda_2$
Ensure:
The final output parameters $newL, W_1, W_2$
$n = size(S)$
for $i = 1$ to n **do**

$$w_{i1} = \frac{\alpha_1\lambda_1 e^{-\lambda_1 x_i}}{\alpha_1\lambda_1 e^{-\lambda_1 x_i} + \alpha_2\lambda_2 e^{-\lambda_2 x_i}}$$
$$w_{i2} = \frac{\alpha_2\lambda_2 e^{-\lambda_2 x_i}}{\alpha_1\lambda_1 e^{-\lambda_1 x_i} + \alpha_2\lambda_2 e^{-\lambda_2 x_i}}$$

end for

$$newL = \sum_{i=1}^{n} \log(\alpha_1\lambda_1 e^{-\lambda_1 x_i} + \alpha_2\lambda_2 e^{-\lambda_2 x_i})$$

return $newL, W_1, W_2$

Algorithm 3 Maximization Step

Require:
S, W_1, W_2
Ensure:
The final output parameters $\alpha_1, \alpha_2, \lambda_1, \lambda_2$
$n = size(S)$

$$\alpha_1 = \frac{\sum_{i=1}^{n} w_{i1}}{n}, \alpha_2 = \frac{\sum_{i=1}^{n} w_{i2}}{n}$$
$$\lambda_1 = \frac{\sum_{i=1}^{n} w_{i1}}{\sum_{i=1}^{n} w_{i1} S_i}, \lambda_2 = \frac{\sum_{i=1}^{n} w_{i2}}{\sum_{i=1}^{n} w_{i2} S_i}$$

return $\alpha_1, \alpha_2, \lambda_1, \lambda_2$

According to the observations in Fig. 1, Section IV-A, we choose the top 0.49% videos as popular videos and the rest as non-popular videos, and then train our model for these two groups, respectively, to examine whether the resultant parameters are sensitive to video popularities. The results are shown in Table VI. There are two interesting observations. The first one is that compared with non-popular videos, the average activation latency of popular videos is larger, which means the diffusion rates along social links of popular video is, in general, slower. Although it seems counter-intuitive, it can be explained by the fact that popular videos have a much larger window of interest than non-popular ones, i.e., popular videos typically have a more constant interest rather than a brief surge of interest which is similar to the observations in [39]. Also, note that the video popularity in OSNs decays very fast [19], which means that the majority of the sharings is gathered soon after the video is first diffused. Then for the popular videos, the remaining small portion of video sharings will be distributed along a larger time window, making the activation latency of sharings which occur in the tail longer. The second observation is that the learned average activation latency for popular videos is similar to the value for all videos, i.e, the performance of popular videos is representative. This is because popular videos contribute most of the sharing actions.

TABLE VI
PARAMETER LEARNING FOR DIFFERENT VIDEOS

Video Types	α_1	α_2	$1/\lambda_1$	$1/\lambda_2$	mean
Popular Videos	0.3063	0.6937	0.6339	4.9093	3.5997
Non-Popular Videos	0.3507	0.6493	0.4851	4.8170	3.2978
All Videos	0.3150	0.6850	0.5979	4.8921	3.5394

B. Model Validation

In this section, we discuss the validation of the proposed MADM model on real datasets (i.e. the collected XJ dataset). In particular, we are interested in the predictive performance of the models, i.e., given the activation time of one's existing active sources, can we predict his activation time accurately? We consider two evaluation metrics and compare our results with three other prediction models.

1) Evaluation Metrics: To compare different diffusion models, we use two evaluation metrics, namely, prediction accuracy, and relative expected error. Considering that our model is continuous in time, it is impractical to predict a precise time point. Hence we introduce the tolerance level denoted by Δt, i.e. the model can predict an activation time point t_0 correctly within the given time interval $[t_0 - \Delta t, t_0 + \Delta t]$.

- The first metric is prediction accuracy, i.e. given the real activation time t_{u_i} for User u_i, the prediction accuracy metric is calculated as follows:

$$E(f_{u_i}(t)) = \frac{\sum_{TS_i^m \in S_{te}} (\int_{t_{u_i}-\Delta t}^{t_{u_i}+\Delta t} f_{u_i}(t)dt)}{k} \quad (27)$$

  where S_{te} is the set of influence paths in the test dataset, and $k = |S_{te}|$ is the number of influence paths in S_{te}. For the influence path TS_i^m of User u_i and video object o_m in S_{te}, t_{u_i} is User $u_i's$ real activation time. This metric measures how well the test set fits the given model.
- The second metric we consider is the relative expected error [40], which can be regarded as a soft version of accuracy. It does not insist on predicting the exact activation time, but takes into account the prediction error. For a real activation time in the test dataset, we measure the expected time error between the real activation time and the time predicted by the model.

$$\frac{\sum_{TS_i^m \in S_{te}} \left(\int_{\tau}^{\infty} |t - t_{u_i}| f_{u_i}(t)dt\right)}{k} \quad (28)$$

Compared with previous studies [41][7], which try to improve the accuracy of predicting whether an activation will

happen by learning the heterogeneous diffusion probability of each individual social link, the main concern of the performance evaluation here is the temporal aspect, i.e., the accuracy of predicting the occurrence time with a given activation probability. In our evaluation, we only consider those users who finally got activated, by setting $P_{u_i} = 1$[12]. Under this setting, the evaluation result will be the same for both LTM and ICM. Therefore, we do not differentiate LTM and ICM when presenting the evaluation result.

2) Comparison Models: Here, we use four non-trivial models for comparison.

- Random Model: Random Model (RM) is a baseline model, which assumes that the activation probability follows a uniform distribution as follows:

$$f(t) = \begin{cases} \frac{1}{b-a}, & a \le t \le b \\ 0, & t < a \| t > b \end{cases} \tag{29}$$

- Most Recent Model: Most Recent Model (MRM) is a way to describe how a human's historical behavior may influence his forthcoming actions. It considers that human behaviors are not purely random, but may follow certain patterns. Previous human dynamics studies in various areas, e.g. emails [42] and short messages [43], have demonstrated that human activities have the characteristics of following certain patterns or obey certain distributions. The specifications of MRM are described as follows. To predict User $u_i's$ activation time, we firstly collect all of $u_i's$ historical influence path records and order them by $u_i's$ activation times as the following sequence $TS_i^{m_1}, TS_i^{m_2}, \ldots, TS_i^{m_N}$. Here, N is the total number of historical records, and $u_i's$ activation times satisfy $t_{u_i}^{m_1} > t_{u_i}^{m_2} > \ldots > t_{u_i}^{m_N}$. Then MRM will estimate $u_i's$ future MRS activation latency as that in his last known influence path record $TS_i^{m_1}$.
- Exponential Model: Exponential Model (EXPM) is a variant of our proposed MADM, which assumes that the activation probability follows an exponential distribution as shown in Equation (23). The only unknown parameter α in EXPM can be easily calculated using the maximum likelihood technique.
- Rayleigh Model: the Rayleigh model (RAYM) is a well-known parametric model previously used in describing the temporal dynamics of diffusion in OSNs [9] and also in epidemiology [44], where the likelihood of influence or activation is modeled via the Rayleigh distribution. Specifically, $f(t_i|t_j,\omega_{j,i})$ is defined as the conditional likelihood of transmission between node u_j and node u_i. The transmission likelihood depends on the activation times (t_j, t_i) and transmission rate $\omega_{j,i}$. The detailed formulation is $f(t_i|t_j,\omega_{j,i}) = \omega_{j,i}(t_i - t_j)e^{-\frac{1}{2}\omega_{j,i}(t_i-t_j)^2}$. The only unknown parameter ω in RAYM can be easily calculated using the maximum likelihood technique.

3) Experimental Settings: In the experiments we set θ to be 24 hours, and select the influence path records within 120 days starting from January 1 2011. Additionally, we divide all data into a training dataset and a testing dataset. We use the influence path records within the time period [January 1 2011, March 31 2011] as the training set to learn model parameters, and the influence path records within the time period [April 1 2011, May 1 2011] as the testing set to validate the prediction power of MADM. We further consider the tolerance level Δt as

$$\Delta t = \{0.2, 0.4, 0.6, 0.8, 1.0, \ldots, 3.8, 4.0\} \tag{30}$$

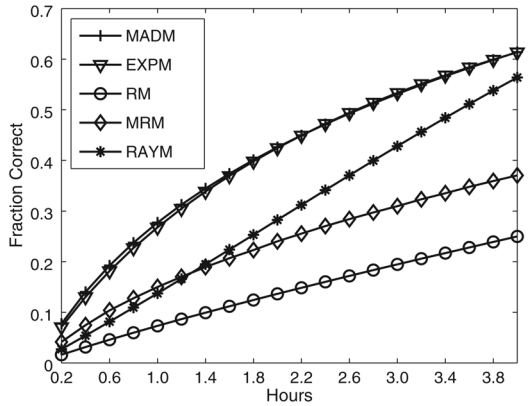


Fig. 6. Prediction accuracy.

TABLE VII
RELATIVE EXPECTED ERROR (HOURS)

Model	MADM	EXPM	RAYM	MRM	RM
Value	4.24	5.39	5.61	8.07	9.77

4) Performance: Fig. 6 compares the prediction accuracy of our model with other models. The X-axis corresponds to the tolerance level from 0.2 to 4.0. The Y-axis is the corresponding prediction accuracy value for the given tolerance level. Specifically, they show that the prediction accuracies of MADM, EXPM, RM, RAYM and MRM increase with the increase of tolerance levels. In general, our proposed MADM performs best with an accuracy around 63% when the tolerance level equals to 4 hours. Since exact prediction accuracy is hard to achieve, and there is much noise (like spam account and spam records) in the data, we think this is a remarkable result. Also note that here we only employed the general parameters learned from the whole population due to the computational simplicity as well as the lack of adequate training data for each individual user. If we take user heterogeneity into consideration, performance may be further improved by learning the parameters for each individual user.

Moreover, if we can accept a tolerance level larger than 1.8 hours, EXPM can perform as well as MADM. In this sense, EXPM could be treated as a good substitution for MADM due to its model simplicity. In addition, we can see that MADM, EXPM, RAYM and MRM outperform the baseline model RM under all tolerance levels. This is because RM does not use any historical influence diffusion records and just predicts the activation times randomly.

According to another performance metric, i.e., the Relative Expected Error performance (Equation (28)), as shown in Table VII, MADM still performs the best with 4.2 hours relative expected error. EXPM performs the second best, followed by RAYM, MRM and RM.

C. Discussion

Our work is not without limitations. We focus on the diffusion processes that are fully driven by inter-personal social links. In reality, diffusion can also benefit from external influential sources (e.g., mass media), and this has been previously observed in Twitter by Myers *et al.* [31]. They show that the information diffusion process in Twitter is determined not only by online activities but also by external activities which are not recorded online. This raises two interesting issues. How can we detect the external influence in the OSNs of our interest? How can our proposed MADM be combined with the external influence factors to generate a hybrid model? It would be interesting and meaningful to study these issues in our further work.

VII. Conclusion and Future Work

This paper introduces a Multi-source-driven Asynchronous Diffusion Model (MADM) to describe the information propagation process in OSNs. Through an influence measurement study of video-sharing propagation on a dominant Chinese online social network, we firstly characterize the temporal patterns of the information propagation process. Based on the empirical findings, we propose an analytical framework to study how multiple active sources affect the diffusion process. Next, we develop MADM to describe the information propagation behavior in social networks and predict the individual activation time. Using the datasets collected from the Renren social network, we conducted model training and model validation. The experimental results show that MADM performs better than other existing models.

One future research is to utilize the proposed asynchronous diffusion model to investigate the information diffusion related applications. For example, we may reconsider the influence maximization problems [45], [46] in OSNs by incorporating the temporal factors.

References

[1] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *Proc. ACM SIGCOMM 2010 Conf.*, 2010, pp. 75–86.

[2] H. Li, H. Wang, J. Liu, and K. Xu, "Video sharing in online social networks: Measurement and analysis," in *Proc. 22nd Int. Workshop Netw. Operat. Syst. Support Digit. Audio Video*, 2012, pp. 83–88.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 491–501.

[4] M. Granovetter, "Threshold models of collective behavior," *Amer. J. Sociol.*, vol. 83, no. 6, pp. 1420–1443, 1978.

[5] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Market. Lett.*, vol. 12, no. 3, pp. 211–223, 2001.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2003, pp. 137–146.

[7] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," in *Adv. Mach. Learn.*, 2009, pp. 322–337.

[8] G. Niu, V. O. K. Li, Y. Long, and K. Xu, "A measurement-driven temporal analysis of information diffusion in online social network," in *Proc. 2012 IEEE GLOBECOM*, Dec. 2012, pp. 2060–2065.

[9] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 561–568.

[10] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *Proc. Sixth ACM Int. Conf. Web Search Data Min.*, 2013, pp. 23–32.

[11] N. Du, L. Song, A. J. Smola, and M. Yuan, "Learning networks of heterogeneous influence," in *Proc. NIPS*, 2012, pp. 2789–2797.

[12] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *Proc. 4th ACM Int. Conf. Web Search Data Min.*, 2011, pp. 65–74.

[13] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *Proc. VLDB Endow*, vol. 5, no. 1, pp. 73–84, Sep. 2011.

[14] 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 *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 695–704.

[15] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 15–28.

[16] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Proc. 2008 IEEE 20th Int. Workshop Quality Service*, 2008, pp. 229–238.

[17] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith, "Visual memes in social media: Tracking real-world news in youtube videos," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 53–62.

[18] R. Zhou, S. Khemmarat, and L. Gao, "The impact of YouTube recommendation system on video views," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 404–410.

[19] H. Li, J. Liu, K. Xu, and S. Wen, "Understanding video propagation in online social networks," in *Proc. 2012 IEEE 20th Int. Workshop Quality Service*, 2012.

[20] E. Kiesling, M. Gunther, C. Stummer, and L. Wakolbinger, "Agent-based simulation of innovation diffusion: A review," *Central Eur. J. Oper. Res.*, vol. 20, no. 2, pp. 183–230, 2012.

[21] F. M. Bass, "A new product growth for model consumer durables," *Manage. Sci.*, vol. 15, no. 5, pp. 215–227, Jan. 1969.

[22] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, Jul. 2013.

[23] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, May 2007.

[24] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 721–730.

[25] E. Sun, I. Rosenn, C. A. Marlow, and T. M. Lento, "Gesundheit! Modeling contagion through Facebook news feed," in *Proc. 3rd Int. ICWSM Conf.*, 2009, pp. 146–153.

[26] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu, "Joint social and content recommendation for user-generated videos in online social network," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 698–709, Apr. 2013.

[27] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Y. Zhao, "Understanding latent interactions in online social networks," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 369–382.

[28] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. 4th ACM Eur. Conf. Comput. Syst.*, 2009, pp. 205–218.

[29] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media," in *Proc. ICWSM*, 2010, pp. 34–41.

[30] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.

[31] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2012, pp. 33–41.

[32] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2004.

[33] S. A. Myers and J. Leskovec, "On the convexity of latent social network inference," in *Proc. NIPS '10*, 2010, pp. 1741–1749.

[34] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2010, pp. 1019–1028.

[35] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf.*, 2009, pp. 49–62.

[36] C. J. Wu, "On the convergence properties of the em algorithm," in *Proc. Ann. Statist.*, 1983, pp. 95–103.

[37] C. Biernacki, G. Celeux, and G. Govaert, "Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models," *Comput. Statist. Data Anal.*, vol. 41, no. 3–4, pp. 561–575, 2003.

[38] V. Hasselblad, "Estimation of finite mixtures of distributions from the exponential family," *J. Amer. Statist. Assoc.*, vol. 64, no. 328, pp. 1459–1471, 1969.
[39] N. Sastry, A. Hylick, and J. Crowcroft, "Spinthrift: Saving energy in viral workloads," in *Proc. 2010 2nd Int. Conf. Commun. Syst. Netw.*, Jan. 2010, pp. 1–6.
[40] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, 2011, pp. 1082–1090.
[41] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Min.*, 2010, pp. 241–250.
[42] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, May 2005.
[43] Z.-D. Zhao, H. Xia, M.-S. Shang, and T. Zhou, "Empirical analysis on the human dynamics of a large-scale short message communication system," *Chin. Phys. Lett.*, vol. 28, no. 6, 2011.
[44] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *Amer. J. Epidemiol.*, vol. 160, no. 6, pp. 509–516, 2004.
[45] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Extracting influential nodes on a social network for information diffusion," *Data Min. Knowl. Discovery*, vol. 20, no. 1, pp. 70–97, 2010.
[46] D. Kempe, J. Kleinberg, and E. Tardos, "Influential nodes in a diffusion model for social networks," in *Proc. 32nd Int. Conf. Automata, Lang. Program.*, 2005, pp. 1127–1138.

Guolin Niu (S'12) received the B.S. degree in software engineering and the M.S. degree in system engineering from Xian Jiaotong University, Xian, China, in 2007 and 2010, respectively, and is currently working towards the Ph.D. degree with the Department Electrical and Electronic Engineering, University of Hong Kong (HKU), Hong Kong.

Her research interests include measurement studies of online social networks, modeling of information diffusions, and marketing strategies for social advertising.

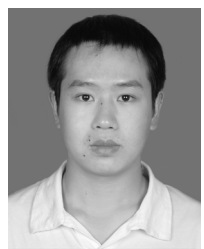
Xiaoguang Fan received the bachelor's degree in electronic engineering from Tsinghua University, Beijing, China, in 2009, and the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2013.

His research interests are opportunistic mobile network, social network analysis, and social data mining.

Victor O.K. Li (S'80–M'81–SM'86–F'92) received the B.S., M.S., E.E., and D.Sc. degrees in electrical engineering and computer science from MIT in 1977, 1979, 1980, and 1981, respectively.

Previously, he was a Professor of Electrical Engineering at the University of Southern California (USC), Los Angeles, CA, USA, and Director of the USC Communication Sciences Institute, Los Angeles, CA, USA. He has also served as Associate Dean of Engineering at the University of Hong Kong, Hong Kong, and Managing Director of Versitech Ltd., the technology transfer and commercial arm of the University of Hong Kong (HKU), Hong Kong. He also served on the board of China.com Ltd., Hong Kong. He is currently Chair Professor of Information Engineering and Head of the Department of Electrical and Electronic Engineering at HKU, Hong Kong, and serves on the board of Sunevision Holdings Ltd., Hong Kong, and Anxin-China Holdings Ltd., Hong Kong, listed on the Hong Kong Stock Exchange. Sought by government, industry, and academic organizations, he has lectured and consulted extensively around the world.

Dr. Li is a Registered Professional Engineer and a Fellow of the Hong Kong Academy of Engineering Sciences, the IAE, and the HKIE. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the U.K. Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the Hong Kong Special Administrative Region, China.



Yi Long received the B.S. and M.S. degrees from Xian Jiaotong University, Xian, China, in 2008 and 2011, respectively, and is currently pursuing the Ph.D. degree with the Department Electrical and Electronic Engineering, University of Hong Kong, Hong Kong.

He is currently working on measurement and modeling of user behaviors in online social networks.

Kuang Xu received the Ph.D. degree from the University of Hong Kong (HKU), Hong Kong, in 2010.

He is currently a research fellow with HKU.