IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 19, 2020, accepted February 17, 2020, date of publication February 24, 2020, date of current version March 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976126

Detection of Rapidly Spreading Hashtags via Social Networks

YOUNGHOON KIM[1] **AND JIWON SEO**[2]

[1]Devision of Computer Science, Hanyang University, Ansan 15588, South Korea
[2]Department of Computer Science, Hanyang University, Seoul 04763, South Korea

Corresponding authors: Younghoon Kim (nongaussian@gmail.com) and Jiwon Seo (seojiwon@hanyang.ac.kr)

This work was supported in part by the Next-Generation Information Computing Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) under Grant2017M3C4A7063570, in part by the Research Fund of Hanyang University under Grant HY-2014-N, in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korea Government (MSIT) (WiseKB: Big Data Based Self-Evolving Knowledge Base and Reasoning Platform) under Grant 2013-0-00109.

ABSTRACT Social network services (SNSs) such as Twitter and Facebook have emerged as a new medium for communication. They offer a unique mechanism of sharing information by allowing users to receive all messages posted by those whom they "follow". As information in today's SNSs often spreads in the form of hashtags, detecting rapidly spreading hashtags in SNSs has recently attracted much attention. In this paper, we propose realistic epidemic models to describe the probabilistic process of hashtag propagation. Our models take into account the way how users communicate in SNSs; moreover the models consider the influence of external media and separate it from internal diffusion within networks. Based on the proposed models, we develop efficient inference algorithms that measure the propagation rates of hashtags in social networks. With real-life social network data including hashtags and synthetic data obtained by simulating information diffusion, we show that the proposed algorithms find fast-spreading hashtags more accurately than existing algorithms. Moreover, our in-depth case study demonstrates that our algorithms correctly find internal diffusion rates of hashtags as well as external media influences.

INDEX TERMS Social network, information diffusion, hashtag, probabilistic modeling, EM algorithm.

I. INTRODUCTION

With the rise of social network services (SNSs) such as Twitter and Facebook, their role as media of information is becoming increasingly important. These SNSs commonly provide the concept of *following* other users as a way of sharing information; a user receives all the messages from those whom they follow. With the scale-free network structures of the SNSs, a user's postings can easily reach large numbers of people via his followers and followers of followers, etc., by content-sharing features, such as "retweet" in Twitter.

In this paper we are interested in detecting rapidly spreading hashtags in SNSs. A hashtag is a user-generated tag in text messages to make it easy for others to find their messages. Usually, a hashtag represents a specific topic or idea of a message. As more users find a hashtag, or its representing idea, compelling, the hashtag becomes viral and used by many people. For example in 2015, the hashtag `#ILookLikeAnEngineer`, initially used by a few female engineers, was quickly adopted by many other female engineers to break the gender stereotypes and highlight diversity in the engineering industry.

Hashtags in SNSs propagate mainly in two distinct mechanisms: 1) users learn a new topic from external media such as BBC and New York Times, and share them in SNSs with a hashtag for the topic, or 2) users read the postings of their followees (those whom they follow) then share the hashtag on their messages. Note that it is already hard to find rapidly spreading hashtags in SNSs, because there are too many hashtags; simply counting the hashtags to detect the ones with rapidly growing frequency does not scale. It is even harder to identify the diffusion mechanisms (from external media or from followee to follower) for rapidly spreading hashtags.

This paper presents a novel model-centric approach for detecting rapidly spreading hashtags in social networks. We design realistic models to describe the probabilistic process of hashtag propagation. Based on our diffusion models, inference algorithms are proposed to estimate diffusion

The associate editor coordinating the review of this manuscript and approving it for publication was Xiping Hu.

rates of hashtags. Our proposed models separate diffusion by external influence from diffusion over the connections of networks; hence the inference algorithms accurately estimate the diffusion rates over networks and identify the *hashtags that primarily spread via networks*.

To the best of our knowledge, this paper presents *the first practical technique to find spreading hashtags in SNSs scalable to large graphs*. The contributions of this paper are as follows:

- We propose realistic diffusion models considering the user's pattern of posting messages with hashtags in SNSs, such as the recency of posted messages and the influence of external media.
- Based on our probabilistic diffusion models, we design efficient inference algorithms. Using the Expectation Maximization (EM) technique, our inference algorithms estimate the diffusion rates as well as external influences, which maximize the likelihoods in the models.
- To evaluate how accurately our inference algorithms identify rapidly spreading hashtags, we use a real-world Twitter dataset, collected in a large metropolitan area in Korea for three months. The tweet data is examined by active Twitter users, and viral hashtags are manually annotated. With the annotated data, we perform an in-depth analysis of the viral hashtags reported by our algorithms.
- We develop a parallel/distributed version of the inference algorithms running on Pregel, a well-known parallel graph processing model. Our evaluation demonstrates that the distributed inference algorithm scales to a large network with 80 million connections; for the network it takes less than six minutes for eight machines to compute the diffusion rates for a hundred hashtags.

Case Study (Viral Hashtags Discovered): In our in-depth analysis with real-world Twitter dataset, our algorithm identified interesting viral hashtags. Among the identified hashtags, a few notable topics for them are `#MeToo` movement and political scandal. In October 2016, a year before the global `#MeToo` movement begins, there was a similar movement in Korea in artistic communities; a few artists held a public interview to reveal the prevalent sexual harassment in the communities. Soon the related hashtags started circulating in social media as users share their own `#MeToo` stories. Those hashtags discovered by our algorithm spread fast over many local communities in SNSs. However, at that time, it did not become very well-known in Korea like the world-wide `#MeToo` movement.

Another group of viral hashtags we identified is about the political scandal in Korea, the one that resulted in the impeachment of the former president. Although we identified many widely used hashtags for this topic, most of them became viral due to the influence of external media. However, our algorithm did successfully discover some hashtags that became viral primarily because of the network effect – i.e., with little influence from the external media. For instance, the hashtag `#by_the_way_Choi` (Choi is the person at the center of the political scandal) is used jokingly in the messages of unrelated topics to draw people's attention to the political scandal. The SNS users found it fun to use the hashtag and other similar ones in unrelated messages, thus those hashtags became viral and widely used. These examples show that our algorithm effectively captures the user's behavior of using hashtags in SNSs and discovers viral hashtags that primarily spread via the networks.

II. PRELIMINARIES

In this section, we describe the notations that are used in our hashtag diffusion models in SNSs and in our inference algorithms to estimate the diffusion rates. We then formally define the problem of hashtag diffusion in SNSs.

Notations: Many popular SNSs such as Twitter, Facebook and Instagram offer a mechanism of sharing information by allowing each user to receive all messages from those who he *follows*. We will refer to those who follow a user as *followers* and those whom a user follows as *followees*. When users log into those services, they may see a list of all messages posted by their followees in reverse chronological order of the posting times in their homepages, which is called a *timeline*.

Let H be the set of hashtags appearing in the messages posted in SNSs. If a message posted by a user u with a hashtag $h \in H$ is followed by another message including h posted by a user v who follows u, with the influence of the preceding one by u, we say that u *infected* v *with* the hashtag h. Note that we will present formal definitions of infection later in Section IV. Furthermore, we call those who are either recovered or never infected, and thus can be infected with a hashtag, a *susceptible* user. Let $\mathbb{G}=(V, E)$ denote a directed graph that represents relationships between users in a SNS. A vertex $u \in V$ is an SNS user and a directed edge $(u, v) \in E$ exists if a user v follows another user u, which implies that all messages posted by u are published on the timeline of v. In other words, u can infect v with a hashtag.

We consider SNS message logs with hashtags collected in a short term of time such as a week or a month as input. For convenience in modeling, each infection time is recorded in uniform discrete time segments, each of which is about several hours to a day long. We call this time segment a *day* in this paper and T denotes the number of *days* in the observed data. Let $\tau_h(u)$ be the set of *day*s when a user u posted a message with a hashtag h. Especially we use $\tau_h^0(u)$ to refer to the first day when u mentioned h. If u is never infected with h in our observed period, we set $\tau_h^0(u) = \infty$. $\mathbb{IN}_h(u, t)$ denotes the number of u's followees who have already posted one or more messages with h no later than t. Similarly, we refer to the number of u's followees who have mentioned h between $t - \Delta$ and t as $\mathbb{IN}_h(u, t, \Delta)$. We summarize the notations used in our paper in Table 1.

Our problem is then defined as follows.

Problem Definition: Given a set H of hashtags, a social network graph $\mathbb{G} = (V, E)$ and the infection days $\tau_h(v)$ for every hashtag $h \in H$ and user $v \in V$, our goal is to find

TABLE 1. List of notations.

Notation	Description
H	The set of hashtags appearing in the collection of SNS messages
$\mathbb{G}=(V,E)$	The directed graph representing follower-and-followee relationships between SNS users
T	The number of dates for SNS message collection
$\tau_h(u)$	The set of dates a user u posted a message including hashtag h
$\tau_h^0(u)$	The earliest date in $\tau_h(u)$
$\mathbb{IN}_h(u,t)$	The number of u's followees who have posted at least a message with h no later than t
$\mathbb{IN}_h(u,t,\Delta)$	The number of u's followees who have mentioned h between $t-\Delta$ and t
$(\delta, 1-\delta)$	The probability that users access social media or external media during a day
ρ_h	The probability with which users are infected with the hashtag h during a day by one of their followees who already have infected
ϵ_h	The probability with which users mention h in their message after accessing external media

the top-k most rapidly spreading hashtags among H and to identify the influence from external media for those hashtags.

III. HASHTAG DIFFUSION MODELS AND INFECTION RATE INFERENCE

We propose three generative models that statistically capture the behavior of social network users who post with hashtags influenced by their followees in this section. The models are commonly based on the following assumptions. It is generally accepted in many information diffusion models [1]–[3] that infection occurs with a uniform probability each time people are exposed to information. Thus we assume the followings:

Assumption 1 (Social Infection Rate): For a hashtag h, each uninfected user becomes infected with probability ρ_h, which is called infection rate (or diffusion rate) *of h, every time he/she sees a message with h posted by his followees in the timeline.*

Based on the above assumption, each day a user stays susceptible for h with probability $(1-\rho_h)^n$ despite of having n already-infected followees, and should be infected with h with probability $1-(1-\rho_h)^n$.

Many classical diffusion models have assumed a discrete time period at which a single chance is allowed for an information to be transmitted [4]. Thus, the next assumption follows:

Assumption 2 (Periodic Media Access): We simply assume that people check any media such as new papers and SNSs at least once in a day, and post a message on SNSs about what they read from those media. Note that throughout this paper we use the term day *to denote the period people access media, rather than its literal meaning. Furthermore, even if a user posts a hashtag multiple times in a day, we regard it as a single posting to simplify the different access rates between SNS users.*

Algorithm 1 EM-IPSI: EM Algorithm for IPSI

```
Input  : $\mathbb{G} = (V, E)$ and for a $h \in H$, $\tau_h(v)\ \forall v \in V$
Output: $\rho$ and $\epsilon$
Randomly initialize $\rho$ and $\epsilon$;
while $logL$ does not converge do
    Set $\rho_1$, $\rho_2$, $\rho_3$, $\epsilon_1$ and $\epsilon_2$ to 0;
    for each vertex $v \in V$ do
        for each $t$ from 1 to $min(\tau^0(v) - 1, T)$ do
            Compute $x_{v,t,1}$ and $x_{v,t,2}$ in Eqn. (2) ;
            $\epsilon_2 \leftarrow \epsilon_2 + x_{v,t,1}$, $\rho_2 \leftarrow \rho_2 + x_{v,t,2} \cdot \mathbb{IN}(v,t)$;
        end
        if $\tau(v) \neq \emptyset$ then
            Compute $y_{v,1}$ and $y_{v,2}$ in Eqn. (3) ;
            $\epsilon_1 \leftarrow \epsilon_1 + y_{v,1}$, $\rho_1 \leftarrow \rho_1 + y_{v,2}$;
            Compute $z_{v,k}$ for $k$=0,...,$\mathbb{IN}(v, \tau^0(v))-1$ in
            Eqn. (4);
            $\rho_3 \leftarrow \rho_3 + y_{v,2} \cdot \sum_{k=0}^{\mathbb{IN}(v,\tau^0(v))-1} z_{v,k} \cdot k$;
        end
    end
    $\epsilon \leftarrow \epsilon_1/(\epsilon_1+\epsilon_2)$, $\rho \leftarrow \rho_1/(\rho_1+\rho_2+\rho_3)$;
end
```

We next consider the influence from external media as following:

Assumption 3 (External Infection Rate): Whenever users access media once a day, they check external media (other than the SNS) with probability $1-\delta$ (or read their timeline in the SNS with δ). They then may be infected with a hashtag h from those external media with probability ϵ_h.

Myers et al reported that information diffusion in a social media service may be influenced by external out-of-network sources such as New York Times and CNN [5], [6]. With the above assumption of *external infection rate*, we can estimate both the internal infection rate of a hashtag and its external infection rate separately.

Based on the above three assumptions, we propose hashtag diffusion models by adopting the traditional state transition models (e.g., SI, SIR and SIR) and generalizing independent cascade model [7]. Some studies claim that information diffusion follows complex contagion process rather than simple traditional models [8]; for instance, they argue that the trends of diffusion (e.g. infection rates) may change over time. However, the models that we are based on can still describe the diffusion process with reasonably high accuracy, especially when we focus on a small time span. Because we aim to assess the infection rates of hashtags in a relatively small time span, the claims in those work do not undermine our assumptions.

A. AN OVERVIEW OF THE PROPOSED MODELS

While a user's posting with a hashtag may infect other users and trigger them to use the hashtag, it is not generally feasible

TABLE 2. Summary of our proposed models.

	IPSI	IPSI+	IPSI+S
Chance of being infected	Never be infected again once after the user mentions a hashtag	May be infected after Δ days since a user is infected previously	May be infected after Δ days since a user is infected previously
Ability of infecting neighbors	∞ days after the user mentions a hashtag	∞ days after the user mentions a hashtag	Δ days after the user mentions a hashtag
Equations for EM	(2)~(6)	The same as IPSI except $\mathbb{IN}_h(v,t)$ is replaced with $\mathbb{IN}_h(v,t,\Delta)$ in Equations (2) and (6)	(2)~(4), (9) and (10)
Algorithm	EM-IPSI	EM-IPSI+	EM-IPSI+S
Pseudocode	Algorithm 1	The same as Algorithm 1 except using $\mathbb{IN}_h(v,t,\Delta)$ instead of $\mathbb{IN}_h(v,t)$	Algorithm 2

in online social media to pinpoint the causal effect; i.e., it is hard to determine which a hashtag in a user's posting infected another user or how long the posting is able to influence neighbors since it is posted. For our analysis, we design three infection models that adopt different definitions of infection events. We first divided two cases about the chance of being infection; a user can be influenced by the hashtag only once when the user first mentioned the hashtag or user can be infected repeatedly every Δ days. For the ability of infecting neighbors, we assume that a message of a user including a hashtag can influence his/her neighbors forever once after it is posted, or a message can influence its owner's neighbors for only Δ days.

According to these assumptions we develop three models named IPSI, IPSI+ and IPSI+S models. Note that IPSI+S model is based on more complicated and realistic assumptions than the others, but it does not guarantee that it estimates the infection rate of a hashtag more accurately than simpler models such as IPSI. We provide a summary of differences between these models and their corresponding infection rate estimation algorithms in Table 2.

B. IPSI MODEL: INDEPENDENT PROPAGATION MODEL WITH SUSCEPTIBLE-INFECTED STATES

The first model is based on the assumption that *users can be influenced by one of their followees who previously have ever mentioned the hashtag at least once*. Furthermore, since the state of infection lasts indefinitely once a user is infected, the event of infection can happen only once when the user mentions the hashtag the first time.

Consider an infected vertex $v \in V$ (an SNS user) with $\tau_h(v) \neq \emptyset$ (or $\tau_h^0(v) \neq \infty$). For a hashtag h, the user v had survived the infection of h from the independent contagions of its infected followees for each time $t < \tau_h^0(v)$ (i.e., v is not infected with probability $(1-\rho_h)^{\mathbb{IN}(v,t)}$ for each time $t < \tau_h^0(v)$). If v is a susceptible user, we can regard that v survives for every t from 1 to T, which is the number of days in the data. Thus, for each day when user v remains uninfected, v survives the infection from the hashtag h by following the statistical process below:

- For each day $t = 1, \ldots, \min(\tau_h^0(v){-}1, T)$,
  - With probability $1-\delta$, v is exposed to an external source, and v survives the infection with probability $1-\epsilon_h$.
  - With probability δ, v is exposed to his timeline listing his followee's messages, and v survives the infection with probability $(1-\rho_h)^{\mathbb{IN}(v,t)}$.

After surviving the infection for $t < \tau_h^0(v)$, v is finally infected with h at time $\tau_h^0(v)$ following the next steps:

- At $\tau_h^0(v)$ (i.e., the day when first v mentions h),
  - With probability $1-\delta$, v is exposed to an external source, and v is infected with probability ϵ_h.
  - With probability δ, v is exposed to his timeline, and v is infected with $1-(1-\rho_h)^{\mathbb{IN}(v,\tau_h^0(v))}$.

The probability δ with which a user accesses the social media in a given time period, or a *day*, is set to 0.7 by default, which is empirically detemined. According to the above probabilistic process, the likelihood $\mathbb{L}$ given parameters ρ_h and ϵ_h can be formulated as

$$\mathbb{L} = \left[\prod_{v\in V}\prod_{t=1}^{min(\tau_h^0(v)-1,T)}\left\{(1-\delta)(1-\epsilon_h) + \delta(1-\rho_h)^{\mathbb{IN}(v,t)}\right\}\right] \cdot \prod_{v\in V:\,\tau_h(v)\neq\emptyset}\left\{(1-\delta)\epsilon_h + \delta\left(1-(1-\rho_h)^{\mathbb{IN}(v,\tau_h^0(v))}\right)\right\} \tag{1}$$

1) EM INFERENCE ALGORITHM (EM-IPSI)

We apply the Expectation Maximization technique to optimize the likelihood in equation (1). We can derive Eqns. (2), (3) and (4) that are necessary for *E-step*, and Eqns. (5) and (6) for *M-step* with Mean Field Approximation [9]. Note that we omitted h from the subscription of the symbols for concise presentations.

$$x_{v,t,1} = \frac{(1-\delta)(1-\epsilon)}{(1-\delta)(1-\epsilon)+\delta(1-\rho)^{\mathbb{IN}(v,t)}}, \tag{2}$$

$$y_{v,1} = \frac{(1-\delta)\epsilon}{(1-\delta)\epsilon+\delta\left(1-(1-\rho)^{\mathbb{IN}(v,\tau^0(v))}\right)}, \tag{3}$$

$$z_{v,k} = \frac{\bar{\rho}^k}{1+\bar{\rho}+\ldots+\bar{\rho}^{\mathbb{IN}(v,\tau^0(v))-1}} = \frac{\rho\cdot\bar{\rho}^k}{1-\bar{\rho}^{\mathbb{IN}(v,\tau^0(v))}}, \tag{4}$$

$$\epsilon = \frac{\sum_{v\in V:\ \tau(v)\neq\emptyset} y_{v,1}}{\sum_{\substack{v\in V:\\ \tau(v)\neq\emptyset}} y_{v,1} + \sum_{v\in V}\ \sum_{t=1}^{min(\tau^0(v)-1,T)} x_{v,t,1}}, \tag{5}$$

$$\rho = \frac{\sum_{v\in V:\ \tau(v)\neq\emptyset} y_{v,2}}{\left(\begin{array}{l}\sum_{\substack{v\in V:\\ \tau(v)\neq\emptyset}} y_{v,2}+\sum_{v\in V}\ \sum_{t=1}^{min(\tau^0(v)-1,T)} x_{v,t,2}\cdot\mathbb{IN}(v,t)\\ +\sum_{v\in V:\ \tau(v)\neq\emptyset}\sum_{k=0}^{\mathbb{IN}(v,\tau^0(v))-1} y_{v,2}\cdot z_{v,k}\cdot k\end{array}\right)} \tag{6}$$

where $\bar{\rho}$ represents $1-\rho$ and $k = 0, \ldots, \mathbb{IN}(v, \tau^0(v)) - 1$ in Eqn. (4). The terms $x_{v,t,2}$ and $y_{v,2}$ in Eqn. (4) are simply the complements of $x_{v,t,1}$ and $y_{v,1}$ respectively (i.e., $1-x_{v,t,1}$ and $1-y_{v,1}$). To yield a factorization of the term with $1-(1-\rho)^{\mathbb{IN}(v,\tau^0(v))}$, we need a non-trivial trick.

2) A TRICK FOR APPLYING JENSEN's INEQUALITY

To apply Jensen's inequality to the logarithm of the likelihood $\mathbb{L}$ in Eqn. (1), we introduce the arbitrary weights $x_{v,t,\cdot}$ and $y_{v,\cdot}$, which can be interpreted as the probability distributions of the hidden variables about whether the node v survived at t and is infected at $t(v)$ from external contagion, and obtain a lower bound of log-likelihood $\log\mathbb{L}$ as follows:

$$\begin{aligned}&\log\mathbb{L}\\ &\geq \sum_{v\in V}\ \sum_{t=1}^{min(\tau_h^0(v)-1,T)} \left\{\log x_{v,t,1}(1-\delta)(1-\epsilon_h)-x_{v,t,1}\log x_{v,t,1}\right\}\\ &\quad+\sum_{v\in V}\ \sum_{t=1}^{min(\tau_h^0(v)-1,T)} \left\{x_{v,t,2}\log\delta\bar{\rho}_h^{\mathbb{IN}(v,t)} - x_{v,t,2}\log x_{v,t,2}\right\}\\ &\quad+\sum_{v\in V:\ \tau_h(v)\neq\emptyset}\left\{y_{v,1}\log(1-\delta)\epsilon_h - y_{v,1}\log y_{v,1}\right\}\\ &\quad+\sum_{v\in V:\ \tau_h(v)\neq\emptyset}\left\{y_{v,2}\log\delta\left(1-\bar{\rho}_h^{\mathbb{IN}(v,\tau_h^0(v))}\right)-y_{v,2}\log y_{v,2}\right\}\end{aligned} \tag{7}$$

Since we cannot apply Jensen's inequality to the term $1-(\bar{\rho}_h)^{\mathbb{IN}(v,\tau_h^0(v))}$ directly, we substitute the term with $\rho_h(1+(\bar{\rho}_h)+(\bar{\rho}_h)^2+\ldots+(\bar{\rho}_h)^{\mathbb{IN}(v,\tau_h^0(v))-1})$. By introducing arbitrary weights $z_{v,k}$ with $k = 0, \ldots, \mathbb{IN}(v, \tau_h^0(v)) - 1$ in Eqn. (4), we finally obtain the lower bound $\mathbb{F}$ of log-likelihood $\mathbb{L}$ as

$$\begin{aligned}\mathbb{F} = &\sum_{v\in V}\ \sum_{t=1}^{min(\tau_h^0(v)-1,T)} \left\{\log x_{v,t,1}(1-\delta)(1-\epsilon_h)\right.\\ &\qquad\qquad\left.- x_{v,t,1}\log x_{v,t,1}\right\}\\ &+\sum_{v\in V}\ \sum_{t=1}^{min(\tau_h^0(v)-1,T)} \left\{x_{v,t,2}\log\delta\bar{\rho}_h^{\mathbb{IN}(v,t)}-x_{v,t,2}\log x_{v,t,2}\right\}\\ &+\sum_{v\in V:\ \tau_h(v)\neq\emptyset}\left\{y_{v,1}\log(1-\delta)\epsilon_h - y_{v,1}\log y_{v,1}\right\}\\ &+\sum_{v\in V:\ \tau_h(v)\neq\emptyset}\left\{y_{v,2}\log\delta\rho - y_{v,2}\log y_{v,2}\right\}\\ &+\sum_{v\in V:\ \tau_h(v)\neq\emptyset}\ \sum_{k=0}^{\mathbb{IN}(u,\tau_h^0(v))-1}\left\{y_{v,2}z_{v,k}k\log\bar{\rho}_h\right.\\ &\qquad\qquad\left.- y_{v,2}z_{v,k}\log z_{v,k}\right\}\end{aligned} \tag{8}$$

3) PSEUDOCODE OF EM-IPSI

The algorithm *EM-IPSI* in Algorithm 1 works as follows: For each vertex v in V, $x_{v,t,\cdot}$'s are computed for every t when v is susceptible (i.e., $1 \leq t \leq min(\tau^0(v)-1, T)$, and aggregated into ϵ_2 and ρ_2 to compute the second terms in the denominators of ϵ and ρ in Eqns. (5) and (6) respectively. Similarly, for $t = \tau^0(v)$ when v is infected, $y_{v,\cdot}$'s are computed and summed into ϵ_1 and ρ_1, which are the first terms in the denominators of ϵ and ρ respectively. Furthermore, $z_{v,k}$ for every k=0,...,$\mathbb{IN}(v, \tau^0(v))-1$ is used to compute ρ_3. Once they are calculated for every vertex, we update ρ and ϵ, and repeat the steps until the log-likelihood $logL$ converges.

4) TIME COMPLEXITY OF EM-IPSI

For every node v in V, we compute $x_{v,t,\bullet}$'s in Eqn. (2) at most T times and $y_{v,\bullet}$'s in Eqn. (3) at most once if v is an infected user. Furthermore, $z_{v,t,k}$'s in Eqn. (4) is calculated $\mathbb{IN}(v, \tau^0(v))$ times for every node. Thus, each iteration of the EM-steps in Algorithm 1 (in lines 4–15) has the time complexity of $O(|V|\cdot(T+\max_{v\in V}\ \mathbb{IN}(v, \tau^0(v)))$.

C. IPSI+ MODEL: INDEPENDENT PROPAGATION MODEL WITH TIME-LIMITED INFECTION BASED ON SI STATES

In IPSI model, we assumed that the users in SNSs read all messages in their timelines and have a uniform chance to be infected by any of the messages with a hashtag, no matter how long ago they are posted. This is unrealistic, even with the short period – a week or a month – of infection data observed for our considered problem; considering the posting rate in today's SNSs, a user is not possibly exposed to all the previous messages in his timeline. Thus, we improve the model by assuming that only the messages posted recently within a given time interval Δ can be found in users' timelines; this model is named *IPSI+ model*. In *IPSI+*, a message disappears from the timeline after Δ days since it is posted, and thus cannot infect the followers with its hashtags. Still, as in IPSI model, the event of infection may happen only once when a user first mentions the hashtag, because the infection for the hashtag lasts indefinitely.

Under the above assumption, the only change from IPSI model is to use $\mathbb{IN}_h(v, t, \Delta)$ instead of $\mathbb{IN}_h(v, t)$ in line 7 of Algorithm 1. As defined in Section II, $\mathbb{IN}_h(v, t, \Delta)$ can be computed as $|\{u \in V|(u, v) \in E \wedge \exists d \in \tau_h(u) \text{ s.t. } t-\Delta < d \leq t\}|$.

Time complexity of EM-IPSI+: As we discussed above, the time complexity of EM-IPSI+ is exactly the same as that of EM-IPSI since the algorithm replaces $\mathbb{IN}_h(v, t)$ with $\mathbb{IN}_h(v, t, \Delta)$ both of which are constants given to the models.

Algorithm 2 EM-IPSI+S: EM Algorithm for IPSI+S

```
   Input  : $\mathbb{G} = (V, E)$ and for a $h \in H$, $\tau_h(v)\ \forall v \in V$
   Output: $\rho$ and $\epsilon$
1  Randomly initialize $\rho$ and $\epsilon$;
2  while logL does not converge do
3      Set $\rho_1, \rho_2, \rho_3, \epsilon_1$ and $\epsilon_2$ to 0;
4      for each vertex $v \in V$ do
5          for each t from 1 to T do
6              if $t \notin \tau(v)$ then
7                  Compute $x_{v,t,1}$ and $x_{v,t,2}$ in Eqn. (2);
8                  $\epsilon_2 \leftarrow \epsilon_2 + x_{v,t,1}$,
                   $\rho_2 \leftarrow \rho_2 + x_{v,t,2} \cdot \mathbb{IN}(v, t, \Delta)$;
9              end
10             else
11                 Compute $y_{v,1}$ and $y_{v,2}$ in Eqn. (3);
12                 $\epsilon_1 \leftarrow \epsilon_1 + y_{v,1}$, $\rho_1 \leftarrow \rho_1 + y_{v,2}$;
13                 Compute $z_{v,t,k}$ for
                   $k$=0,...,$\mathbb{IN}(v, \tau^0(v), \Delta)-1$ in Eqn. (4);
14                 $\rho_3 \leftarrow \rho_3 + y_{v,2} \cdot \sum_{k=0}^{\mathbb{IN}(v,\tau^0(v),\Delta)-1} z_{v,t,k} \cdot k$;
15             end
16         end
17         $\epsilon \leftarrow \epsilon_1/(\epsilon_1 + \epsilon_2)$, $\rho \leftarrow \rho_1/(\rho_1 + \rho_2 + \rho_3)$;
18     end
19 end
```

D. IPSI+S MODEL: INDEPENDENT PROPAGATION MODEL WITH TIME-LIMITED INFECTION AND SUSCEPTIBLE-INFECTED-SUSCEPTIBLE STATES

Lastly in our *IPSI+S model*, we drop the assumption that a user is considered to be infected once and for all, and never be *uninfected* after he mentions the hashtag the first time. In *IPSI+S model*, we assume that users infected with a hashtag may forget it after some time (hence *uninfected* with the hashtag) and may be infected again to post a message with the hashtag, when he is reminded by their followees who have used the hashtag recently within Δ days.

By the new assumption, in our E-step, $x_{v,t,\cdot}$'s are computed for every t at which v does not mention a hashtag h (i.e., $t \notin \tau_h(v)$) even after he mentions the hashtag the first time, which was not calculated again once he mentions it the first time in *IPSI* and *IPSI+ models*. Furthermore, we compute $y_{v,t,\cdot}$ and $z_{v,t,\cdot}$ for every day when v mentions the hashtag h in messages (i.e., for all t's in $\tau_h(v)$) while we compute them once when v mentions h for the first time in the previous models. The equations of $x_{v,t,\cdot}$, $y_{v,t,\cdot}$ and $z_{v,t,\cdot}$ is similar to Eqns. (2)–(4) except we put $\mathbb{IN}_h(v, t, \Delta)$ instead of $\mathbb{IN}_h(v, t)$.

In M-step, ϵ and ρ is calculated as

$$\epsilon = \frac{\sum_{v\in V:\ \tau(v)\neq\emptyset} \sum_{t\in\tau_h(v)} y_{v,t,1}}{\sum_{\substack{v\in V:\\ \tau(v)\neq\emptyset}} \sum_{t\in\tau_h(v)} y_{v,1} + \sum_{v\in V}\ \sum_{t\notin\tau_h(v)} x_{v,t,1}}, \tag{9}$$

$$\rho = \frac{\sum_{v\in V:\ \tau(v)\neq\emptyset} \sum_{t\in\tau_h(v)} y_{v,t,2}}{\left(\begin{array}{c}\sum_{\substack{v\in V:\\ \tau(v)\neq\emptyset}} \sum_{t\in\tau_h(v)} y_{v,t,2} + \sum_{v\in V}\ \sum_{t\notin\tau_h(v)} x_{v,t,2}\cdot\mathbb{IN}(v,t,\Delta) \\ + \sum_{v\in V:\ \tau(v)\neq\emptyset} \sum_{t\in\tau_h(v)} \sum_{k=0}^{\mathbb{IN}(v,t,\Delta)-1} y_{v,t,2}\cdot z_{v,t,k}\cdot k\end{array}\right)} \tag{10}$$

Algorithm 3 Distributed EM-IPSI on Pregel

```
1  Function vertex-compute (u, messages)
2      if terminate then
3          voteToHalt();
4      end
5      for each neighbor v of u do
6          send(v, $\tau^0(u)$);
7      end
8      for each t from 1 to min($\tau^0(u) - 1, T$) do
9          $\mathbb{IN}(u, t) \leftarrow \mathbb{IN}(u, t-1)$ + count of $t$ in messages
10     end
11     $\rho_1, \rho_2, \rho_3, \epsilon_1$ and $\epsilon_2$ are defined as aggregate
       variables;
12     (the same as lines 4–15 in Algorithm 1)
13 Function global-aggregate ()
14     $\epsilon \leftarrow \epsilon_1/(\epsilon_1 + \epsilon_2)$, $\rho \leftarrow \rho_1/(\rho_1 + \rho_2 + \rho_3)$;
15     if logL converged then
16         terminate ← True
17     end
```

The changes in the pseudocode of *EM-IPSI+S* from that of *EM-IPSI* are presented in Algorithm 2. The difference to *EM-IPSI* is that for each user v, not only ϵ_2 and ρ_2 are updated with all day $t \notin \tau_h(v)$ in lines 3–6, but also ϵ_1 and ρ_1 are updated for every day $t \in \tau_h(v)$ in lines 7–12.

Time complexity of EM-IPSI+S: For every node v in V, $x_{v,t,\bullet}$'s are computed at most T times and at most once if v is an infected user. Furthermore, $y_{v,t,\bullet}$'s and $z_{v,t,k}$'s with every k=1, ..., $\mathbb{IN}(v, t, \Delta)$ may be calculated at most T times. Thus, each iteration of the EM-steps in Algorithm 2 (in lines 4–18) runs in $O(|V| \cdot T \cdot \max_{v\in V,\ t\in[1,T]} \mathbb{IN}(v, t))$ time.

IV. DISTRIBUTED EM ALGORITHM WITH PREGEL

The EM algorithms proposed in the previous section compute $x_{v,t,\cdot}$, $y_{v,\cdot}$ and $z_{v,\cdot}$ independently for each vertex v, and aggregate those values to update the model parameters ϵ and ρ. Thus it can be simply and efficiently parallelized using Pregel, a parallel framework for large-scale graph processing system supporting a vertex-centric computatio model [10].

In each iteration of Pregel programs, each vertex processes messages from other vertices with a user-defined *vertex function*, updates its status, and sends messages to other vertices. The iterations are separated by a global synchronization, when all messages in the previous iteration are communicated. Pregel provides *aggregators* that, with user-defined reduction operators, globally aggregate values from vertices after each iteration.

Parallel EM-IPSI Using Pregel: For brevity we only describe the inference algorithm for IPSI model, but it can be simply extended for IPSI+S and IPSI+ model. In actual systems monitoring hashtags, every message posted by a user u is input to the vertex function for u, which processes all hashtags in the message and timestamps in a certain time window (e.g., T days) – typically a week or a month for our considered problem. Pregel EM-IPSI then runs periodically and detects the top-k hashtags that have been propagated in the recent T days with the k largest infection rates.

When distributing the calculation of EM-IPSI algorithm to run vertex-by-vertex separately, the greatest obstacle in scaling for large graphs is to, for each vertex, compute and maintain $\mathbb{IN}_h(v, t)$'s with all candidate hashtags and days in the time window; we need to know all timestamps when h is mentioned by v' followees to compute $\mathbb{IN}_h(v, t)$. In our Pregel version of the EM algorithms, each vertex stores timestamped hashtags of its own only; the EM algorithms are implemented in the two user-defined functions as follows:

- **Vertex function**: Each vertex v receives hashtags with their timestamp from its followee vertices and computes $\mathbb{IN}_h(v, t)$ (or $\mathbb{IN}_h(v, t, \Delta)$ for EM-IPSI+ and EM-IPSI+S) for each hashtag h and day t with $1 \leq t \leq min(\tau_h^0(v)-1, T)$ (or every $1 \leq t \leq T$ for EM-IPSI+S). With the $\mathbb{IN}_h(v,t)$, we calculate $x_{v,t,\cdot,\cdot}$, $y_{v,\cdot}$ and $z_{v,\cdot}$ in Equations (2), (3) and (4) respectively, which are then sent to the aggregator assigned to h.
- **Aggregator**: The aggregator incrementally aggregates the values for each vertex and calculates the maximum likelihood estimations of ρ_h and ϵ_h for the given hidden parameters.

The pseudocode of distributed *EM-IPSI* is shown in Algorithm 3. The code is written to estimate ρ and ϵ for a single hashtag, but it can be simply extended to work for multiple hashtags.

Because the infection information of a vertex needs to be sent to all of its neighbor vertices, our distributed algorithm may have large volume of communication for extremely large-scale networks. For such cases, we make it possible to reduce the communication volume by dividing the overall days in the time window into N sub-ranges, so that N Pregel iterations correspond to a single iteration in our algorithm. In the ith Pregel iteration, infection information of a vertex is sent as a message only if the corresponding infection time is within $\frac{T}{N}(i \mod N)$ and $\frac{T}{N}(i \mod N + 1)$. This requires each vertex u to additionally store $\mathbb{IN}_h(v, t)$ – the number of infected followee for h in the previous time range.

V. EXPERIMENTS

We empirically demonstrated the *accuracy* of the proposed algorithms in detecting rapidly spreading hashtags. The experiments are done on a commodity machine with Intel(R) Core(TM)2 Duo CPU 2.66GHz and 8GB of main memory. All the tests are completed in a reasonable amount of time – less than 0.5 hour at maximum.

A. QUANTITATIVE QUALITY EVALUATION

To evaluate the accuracy of the estimated infection rate, we need to have ground truth infection rates of hashtags or the rankings of highly infectious hashtags in real social networks. However, such information is not available even in social network services because there have been no technique to calculate infection rates of hashtags yet. Thus, we performed two types of experiments for quantitative quality evaluation; 1) we run user studies to evaluate the agreement between the infection rates estimated by real social network users and those by our inference algorithms, and 2) we simulate the diffusion process to generate synthetic data with known infection rates; with the synthetic data we evaluate the accuracy of estimations by our inference algorithms. Furthermore, we utilize the nDCG measure [11] for the quality measure because it is impossible to predict the exact infection rates, and finding relatively rankings of highly infectious hash tags is our goal.

We describe the details of two strategies for quantitative evaluation in the following.

B. IMPLEMENTED ALGORITHMS & DATA SETS

We implemented **EM-IPSI**, **EM-IPSI+** and **EM-IPSI+S** proposed in the previous section. Furthermore, we implemented the following baseline algorithms for accuracy performance study:

- **CM**: This denotes the estimation method which computes the weighted average of infection rate β using the deterministic compartmental SI model [12]. It models the infected (I) and susceptible (S) populations using the differential equations $dS/dt = -\beta(t)SI/N$ and $dI/dt = \beta(t)SI/N$, where $\beta(t)$ is the infection rate at time t. For an overall infection rate, we averaged $\beta(t)$ over all days t from the first day of infection, weighted by 0.5^{t-1} since it has been shown that the infection rate generally decreases rapidly over time according to their models.
- **NV**: This is a naive heuristic inference algorithm. In NV, we compute the infection ratio ρ_v for each node v by solving the equation $1 - (1 - \rho_v)^{\#\text{ of infected neighbors}} = r$ where r is the observed ratio of the number of v's infected neighboring nodes to the number of nodes with the same number of infected neighbors. Then, the infection ratio of a hashtag is inferred to be the harmonic average of ρ_v for all v's. The external infectiousness ϵ is estimated as the ratio of infected nodes among the nodes without any previously infected neighbors.

Datasets used for experiments include both synthetically generated graphs as well as real social networks, and both simulated and real hashtag diffusion data.

1) SOCIAL GRAPHS

For evaluating the prediction accuracy of detected hashtags, we use the online social network of students at the University of California, Irvine, originally collected in [13] including 1, 899 nodes and 20, 296 directed edges. It is called *UC-NET*. For a larger graph, a social graph sampled from Epinions.com provided by [14], with 75, 879 vertices and 508, 837 directed edges, is utilized and we call it *EPINION*. Because *UC-NET* and *EPINION* are social graphs among users without messages, we synthetically simulate the hashtag diffusion based on them.

To assess the accuracy with real hashtag diffusion, we collected a social graph from Twitter using its developer API for 4 months from Oct. 2016 to Jan. 2017 posted in Seoul, Korea. We call the dataset *TWT*. It contains 45, 161 users and 317, 032 follow relationships as well as 6, 821, 168 mentions of 1, 000 hashtags. The hashtags are randomly selected among all hashtags that appeared in *TWT* at least 50 times.

2) HASHTAG DIFFUSION

We simulated the diffusion process in three different scenarios. For each virtual hashtag h, we first randomly select an infection rate ρ_h and a set of starting nodes of random size (at most 10) and the infection date of each starting node, which is also randomly chosen between 1 to 50. With growing t from 1 to 50, an infected vertex transmits h to each follower v with probability ρ_h, and if infection occurs then we append the pair (v, t) into the data set. The three scenarios differs with respect to each other concerning the behavior of users after the first mention of h as following.

- *SIMUL-SI:* A vertex may mention h at most once. A mention of h may infect the followers at any time after it is mentioned, i.e., infinitely.
- *SIMUL-SI+:* A vertex infected with h mentions it once a day with probability 0.5. A mention of h older than Δ is not infectious; only a recent mention of h within Δ time may infect the followers.
- *SIMUL-SIS:* A vertex infected with h becomes susceptible again after one day. The condition for the infection is the same as *SIMUL-SI+*.

We repeat the diffusion simulation for a given number of virtual hashtags. To make the problem challenging, we additionally generate 4 times more noise hashtags mentioned by each vertex with a randomly chosen probability. Note that such simulation has been widely used in many studies of information diffusion [15], [16].

3) QUALITY MEASURES

To measure the quality of estimated infection reates compared to pseudo-ground truth data, we use Precision and Normalized Discounted Gain (nDCG) [11], widely accepted metrics for measuring the quality of ranking. Precision@k is simply the ratio of infectious hashtags correctly found among the top-k hashtags retrieved by an algorithm. Using the ground-truth relevance of hashtags, nDCG measures the accuracy of estimated rank of hashtags sorted in the decreasing order of their ground-truth ρ_h. Because we are interested in highly infectious hashtags and their rankings, we use nDCG@k that measures the accuracy of top-k item's ranking. The value of nDCG@k is computed in the following way. Let $R(h)$ be the known relevance score of a hashtag h. It may denote either the score rated through user study or the infection rate used for the simulation in the synthetic data test. For top-k hashtags with the k largest ρ_hs, the hashtag at the ith position is denoted by h_i. DCG@k is calculated as $\sum_{i=1}^{k} R(h_i)/log(i+1)$ and then, normalized by dividing the value by the ranking of the known relevance scores, which is the the upper bound of DCG@k.

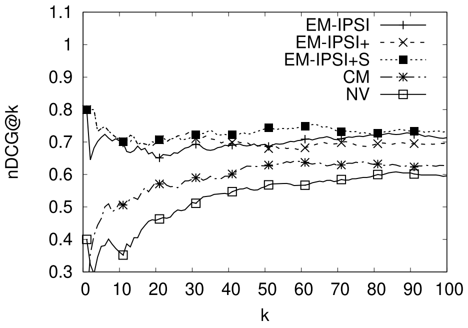


FIGURE 1. Agreement between user rates and estimated infectiousness.

C. PERFORMANCE TEST AND USER STUDY WITH TWT

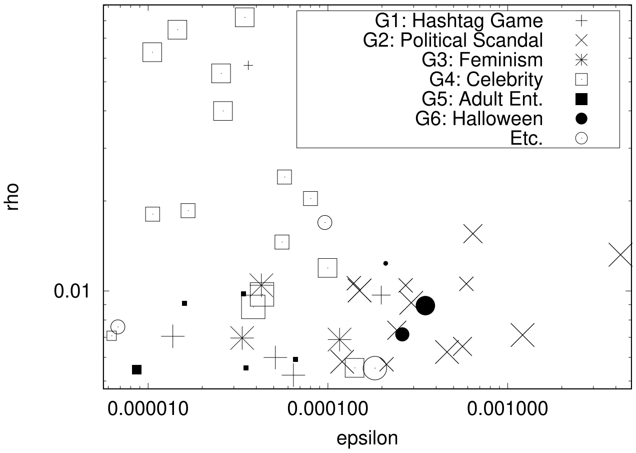
To obtain ground-truth relevance scores for hashtags' contagiousness, we employed 5 active Twitter users to grade hashtags for their contagiousness. We randomly selected 1, 000 hashtags in TWT and asked them to grade each hashtag in a 5-point scale; higher points are given to more infectious (rapidly spreading in Twitter) hashtags. We asked the human annotators to carefully grade the contagiousness by examining the messages of infected users for each hashtag, as well as those posted by his/her followers.

1) PERFORMANCE TEST

Using the contagiousness scores collected from the Twitter users, we plotted nDCG@k of the top-100 infectious hashtags detected by the algorithms with varying k in Figure 1. With growing k, the figure demonstrates that our proposed algorithms *E-IPSI*, *EM-IPSI+* and *EM-IPSI+S* steadily yield high accuracy with *EM-IPSI+S* being the best performer among them. For this test, the default value of probability δ, with which a user accesses the social media each *day*, was empirically set to 0.7. The time period for a *day* was set to 6 hours.

2) USER STUDY

To further verify the detected hashtags for their infectiousness, we investigated 47 common hashtags included in all the top-100 results by the three algorithms; we discovered that most of them can be grouped into 7 categories as shown in Figure 2. Group 1 has the hashtags used for games between Twitter users such as #followme and #1RT (Users share their wishlist tagged with #1RT, #2RT, ...) which is observed to be highly infectious. Group 2 is about the political scandal in South Korea that began in late Oct. 2016. Group 3 consists of hashtags for the feminist issue raised in the communities of Korean artists during the period as a Korean #MeToo movement. Group 4 and Group 5 are the hashtags used for the promotion of celebrities and adult entertainment respectively. Since Halloween was in the middle of our data collection,



(a) Top-47 hashtags plotted with ρ and ϵ

Group	P1	P2	P3	P4	P5	P6	P7
G1	5	4.2	395.6	753.6	2.194	150.2	0.001637
G2	13	4.154	1845	5338	1.897	1129	0.0004961
G3	3	5	98.67	223.7	1.995	36.67	0.004382
G4	15	3.667	83.93	317.1	4.813	36.53	0.02065
G5	5	1.2	96.2	734.4	11.71	242	0.06427
G6	3	2.667	1157	1926	1.757	304.3	0.0003017
Etc.	3	3.667	489.7	1039	4.546	232	0.007682

(b) Statistics of hashtags in 6 groups (P1: the num. of hashtags, P2: avg. rates, P3: avg. num. of infected vertices, P4: avg. num. of messages with each hashtag, P5: avg. num. of hashtag usages per vertex, P6: avg. num. of edges between the vertices in each group, P7: avg. clustering coeff. of the vertices in each group)

FIGURE 2. **Analysis on the top-47 hashtags.**

we observed a few spreading hashtags for Halloween (Group 6).

In Group 2, about the political scandal in Korea that resulted in the impeachment of the former president, our algorithm did successfully discover some hashtags that became viral primarily because of the network effect – i.e., with little influence from the external media. For instance, the hashtag `#by_the_way_Choi` (Choi is the person at the center of the political scandal) is used jokingly in the messages of unrelated topics to draw people's attention to the political scandal. The SNS users found it fun to use the hashtag and other similar ones in unrelated messages, thus those hashtags became viral and widely used. These examples show that our algorithm effectively captures the user's behavior of using hashtags in SNSs and discovers viral hashtags that primarily spread via the networks.

The hashtags in Groups 3 and 6 also support the virtue of our model since those hashtags went viral primarily in SNSs without being highlighted by mass media; Korean `#MeToo` movement was rarely handled in such media and Halloween was not popular event in Korea either. Moreover, the hashtags of Group 1 are definitely ones became viral only in SNSs.

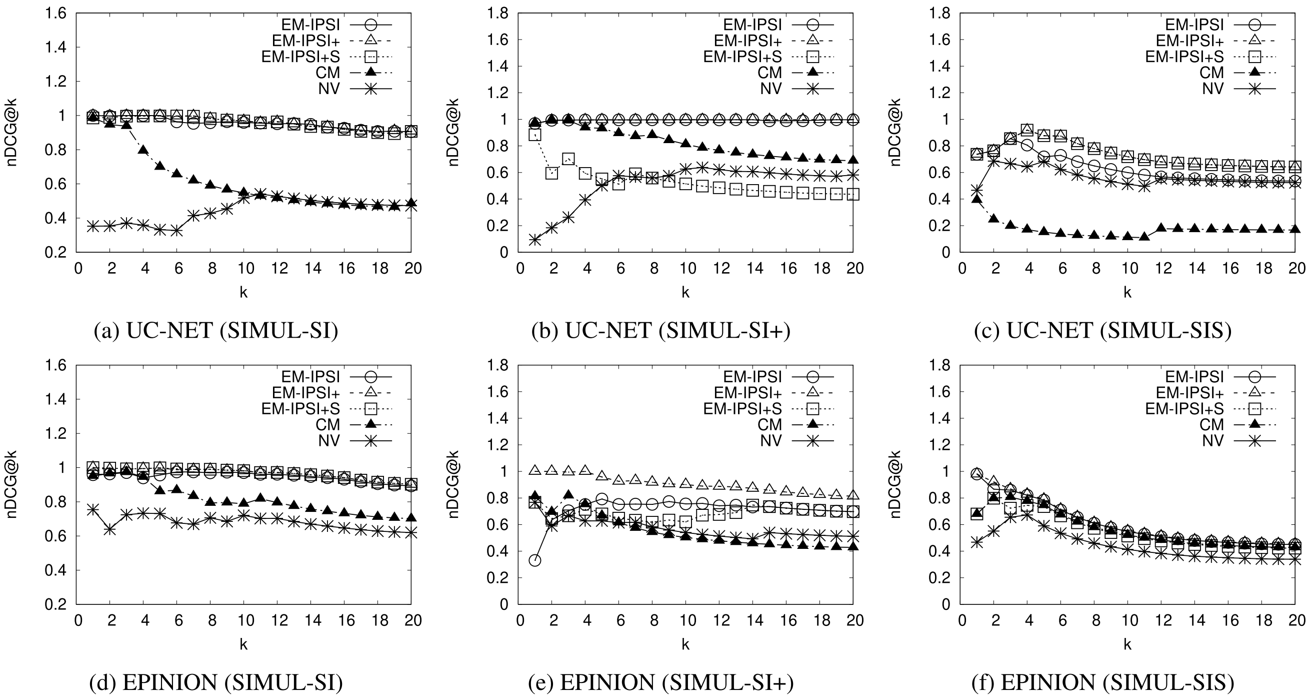
3) INTERNAL AND EXTERNAL INFECTION RATES

In Figure 2a, we plotted those 47 hashtags with respect to their ρ_h and ϵ_h computed by *EM-IPSI+S*. The size of the symbol represents the score by Twitter users. We discover that the hashtags in Group 2 show high contagiousness (ρ) as well as high external infection rates (ϵ); this is probably because people are exposed to the news about the political scandal from external media, but they also discuss the matter actively among their friends in Twitter. In contrast, Group 4 achieves high internal infection rates with low ϵ's since those hashtags are primarily propagated between small fan communities without much influence from external media. In the table in Figure 2b, we also provide some statistics of the hashtags in each group. Clearly, the vertices mentioning the hashtags in Group 4 use them more frequently (P4) and they are also highly connected to each other compared to other groups (P7). This implies that the propagation of hashtags in Group 4 may be due to intentional promotion of celebrities by their fans. Similarly, we can conjecture that the hashtags in Group 5 are also being shared on purpose among the business owners of adult entertainment to advertise their business.

D. PERFORMANCE TESTS WITH SIMULATION

We next present the accuracy of our inference algorithms with simulated diffusion on UC-NET using *SIMUL-SI*, *SIMUL-SI+* and *SIMUL-SIS* in Figures 3(a)-(c). Each dataset includes 20 infectious hashtags and 80 noise hashtags. *SIMUL-SI* generates the infections with a simple assumption that each vertex mentions a hashtag only once, which may infect the followers infinitely after the mention.

All our algorithms detect infectious hashtags in almost exact order of their true ranking. With *SIMUL-SI+*,



(a) UC-NET (SIMUL-SI)

(b) UC-NET (SIMUL-SI+)

(c) UC-NET (SIMUL-SIS)

(d) EPINION (SIMUL-SI)

(e) EPINION (SIMUL-SI+)

(f) EPINION (SIMUL-SIS)

FIGURE 3. **Accuracy test by the simulation with UC-NET and EPINION.**

EM-IPSI and *EM-IPSI+* give high accuracy, but *EM-IPSI+S* does not. It is because while *SIMUL-SIS* considers the influence of all the mentions of the hashtag by followees for infection, *SIMUL-SI+* randomly generates hashtag mentions after being infected without considering the influence from the followees. For *SIMUL-SIS*, *EM-IPSI+S* achieves the best accuracy as is expected. We also obtain a similar trend in the accuracy test with EPINION for all algorithms as shown in Figures 3(d)-(f). When applied to the diffusion data generated by *SIMUL-SIS*, the performance of all our inference algorithms degraded. In our analysis, it is because *SIMUL-SIS* generates too many infected vertices due to the dense connections in EPINION network, which makes the algorithms hard to distinguish infectious hashtags from noises.

Furthermore, in Figures 4a∼4c, we present the accuracy of the implemented algorithms in terms of Precision@k with varying k from 1 to 20 for the three simulation models. The result shows the similar trend to the graphs in Figure 3 measuring the performance in terms of nDCG@k and our algorithms are the best performers. We also find that our algorithms achieve a score 1 for Precision@k with k in the range from 1 to 15 with simulations SIMUL-SI and SIMUL-SI+, which means that they can find top-15 influential hashtags exactly in their correct ranking order among the 20 ground truths.

1) COMPARISON TO GROUND-TRUTH INFECTION RATES

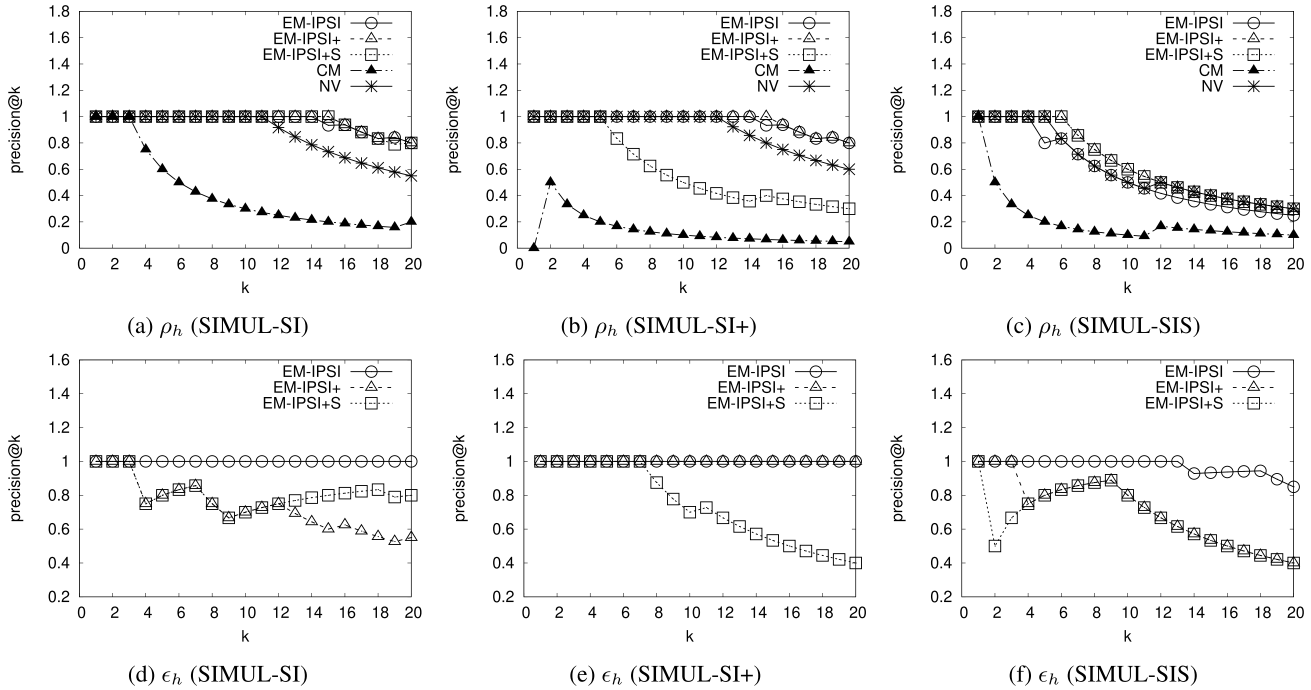
In Figure 5, using UC-NET and EPINION, we plotted the influential hashtags found by the implemented algorithms in the coordinate plane where x and y axises are ground-truth and estimate infection rates respectively. The graphs show that EM-IPSI algorithm calculates the infection rates of hashtags the most similar to the real rates used in the generation of synthetic data sets. However, the other algorithms EM-IPSI+ and EM-IPSI+S also show the linear trends of plots. Considering the discovery that those algorithms detect the top-k hashtags in the exact order of their infection rates in the previous discussion about Precision@k, the plots confirm that our algorithms can assess infection rates proportionally to the real infection rates well even if they may not calculate the rates exactly.

2) EXTERNAL INFECTION RATES

We also tested the accuracy of the estimated external infection rates by our proposed algorithm and show the performance in terms of Precision@k in Figures 4d∼4f with our three simulation models. Note that the goal of our models is to find infectious hashtags that spread over networks rapidly and we don't aim to find the hashtags becoming popular by solely being infected from outside of social networks. Thus, we computed the accuracy of hashtags of top-k external infection rates among the 20 ground truth internally infectious hashtags. The graphs show that our models also reasonably distinguish the hashtags, which are not only external infectious but also internally influential, by about 70% of precision up to top-10 results.

3) SELECTION OF δ

To find a reasonable value of δ, a constant probability which is used when a user determines to access social media or



(a) ρ_h (SIMUL-SI)

(b) ρ_h (SIMUL-SI+)

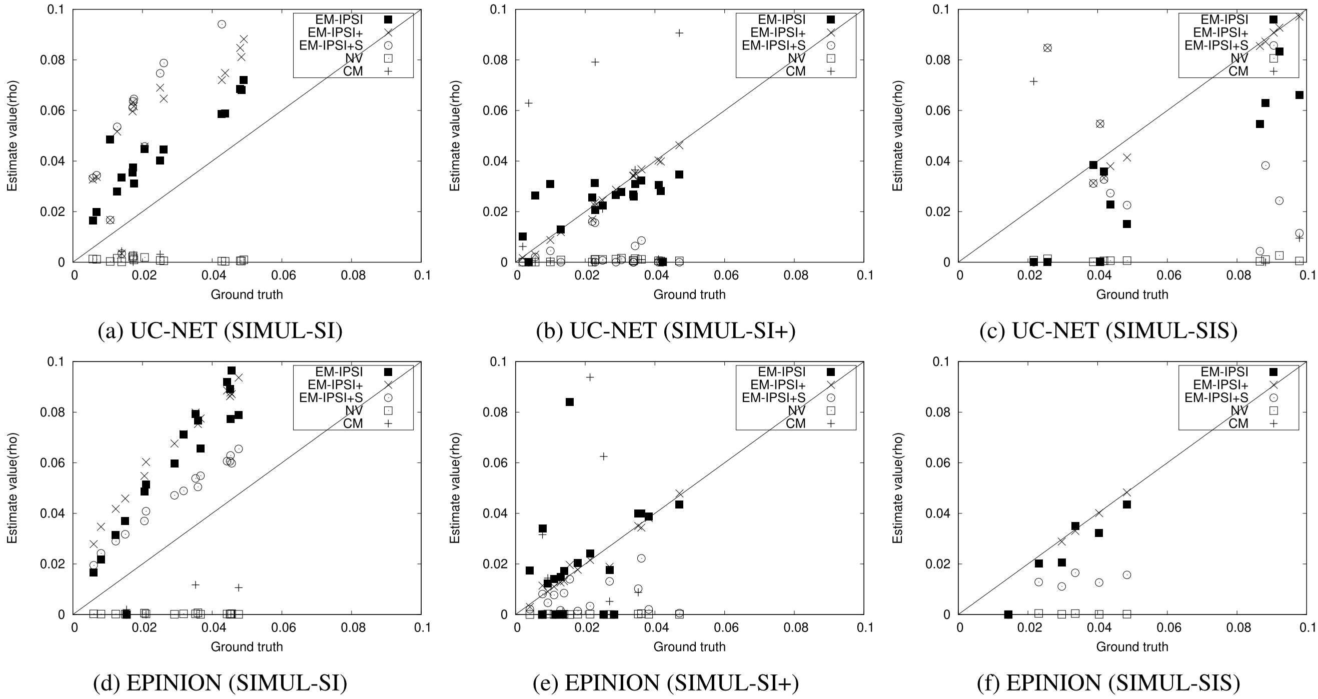
(c) ρ_h (SIMUL-SIS)

(d) ϵ_h (SIMUL-SI)

(e) ϵ_h (SIMUL-SI+)

(f) ϵ_h (SIMUL-SIS)

FIGURE 4. Accuracy test for ρ_h and ϵ_h by the simulation with UC-NET.



(a) UC-NET (SIMUL-SI)

(b) UC-NET (SIMUL-SI+)

(c) UC-NET (SIMUL-SIS)

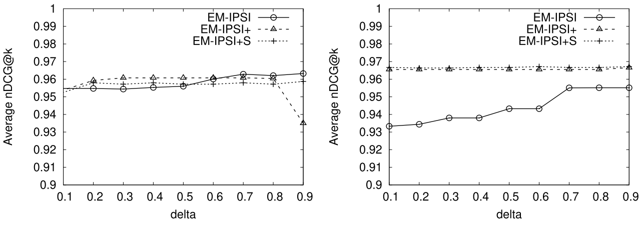
(d) EPINION (SIMUL-SI)

(e) EPINION (SIMUL-SI+)

(f) EPINION (SIMUL-SIS)

FIGURE 5. Plotting estimated and ground truth values of ρ_h in the simulation with UC-NET and EPINION.

any external media for news, we tested EM-IPSI, EM-IPSI+ and EM-IPSI+S with varying δ from 0.1 to 0.9. Figure 6 shows the average nDCG@k over k from 1 to 20 with datasets UC-NET and EPINION. We found that the performance changes slightly with all algorithms. It means that our algorithm can find a hashtag spreading over a social network



(a) UC-NET (b) EPINION

FIGURE 6. Accuracy test with varying δ.

mainly though the network very well regardless of δ and its external infection rate. Based on this results, we set the default δ to 0.7 in our experments.

We also evaluated the performance with varying the length of a *day*. The performance was not affected when we tested with wide ranges of values for the two parameters. However, if a *day* is too lengthy, such as 3 times of the length of a day used in synthetic generation, the accuracy of the inference outcome dropped suddenly.

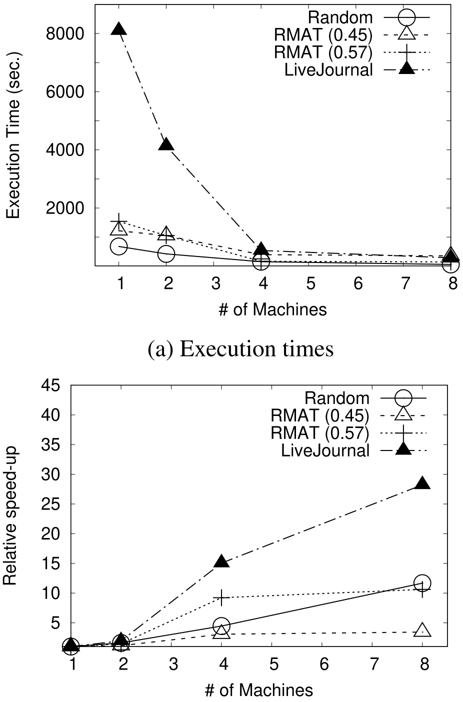
E. SCALABILITY TESTS

To evaluate the scalability of our distributed EM algorithm, we perform two sets of experiments – strong-scaling test and weak-scaling test, both of which are commonly used to evaluate the performance of distributed applications. Because the inference algorithms show similar distributed characteristics, we only report the result of EM-IPSI. In the experiment, we fix the number of iterations of EM-IPSI to be ten, because we evaluate the scalability of the algorithm.

In the strong-scaling test, the infection data is generated with *SIMUL-SI* based on *LJ-NET*; the data contains about 39 million mentions of 100 hashtags in total. With this input data, we run our distributed EM algorithm on 1, 2, 4, 8 machines; we run the test ten times on each machine and show the average execution times in Figure 7a (the variances of the execution times are less than 1–2% of the execution times, thus we do not show them in the figure). The figure demonstrates that the algorithm scales very well; as we increase the number of machines for the computation, the execution times drop accordingly. Notice that on eight machines, it takes less than 300 seconds for the algorithm to finish. In Figure 7b, we plotted the relative speed-up which is the ratio of the execution time to that with 1 machine. The performance scales super-linearly from 2 machines to 4 machines; this is because with less than two machines, there is not enough memory to store messages sent between vertices. The messages sent to a vertex must be all delivered so that its vertex function (the costly part of the algorithm) can be run. With insufficient memory, delivering messages becomes the bottle neck rather than executing the vertex function, which prevents it from fully exploiting the available parallelism.

For the weak-scaling test, using the three types of synthetically generated graphs, the infection data is generated with *SIMUL-SI* to have 100 hashtags; the sizes of generated infection data are from 23 million to 340 million mentions. Note that the size of the infection data is not proportional to the graph size because of the random nature of the generation model. We run our distributed EM algorithm on 1, 2, 4, 8 machines with graphs of 1M, 2M, 4M, 8M vertices respectively. Figure 7a shows the average execution times for three types of input graphs. Since we increase the number of machines in proportion to the graph size, the ideal speed-up should be one for all settings. Still the execution times decrease in weak-scaling test from 2 to 4 machines because of the similar reason that we obtain super-linear speed-up in strong-scaling test.

Overall, the two sets of evaluations demonstrate that our distributed algorithm scales well and runs in practical time. While the cluster size for our evaluation is not large, the evaluation result suggests that our distributed algorithm is CPU-bounded rather than network bounded, hence the algorithm would easily scale to larger clusters.



(a) Execution times

(b) Relative speed-up

FIGURE 7. Scalability of our distributed EM-IPSI algorithm.

VI. RELATED WORK

Early work in information diffusion was studied in the context of spreading diseases. In epidemiology, compartmental models group population into compartments that represent infection status of the population and observe changes of the compartment size to estimate the infection rate [12]. We have tested the performance of a simple modification of compartmental model (CM) compared to our proposed algorithms in this paper. In [19] and [20], the approach in

TABLE 3. Comparison of closely related work in six dimensions (A1 – A6).

	A1	A2	A3	A4	A5	A6
S. A. Myers, et al. [5]	expl.	n	O	O	X	Prob. of *external infection* of information
X. Wu, et al. [16]	impl.	n	X	X	X	Hidden links between infected nodes
M. Gomez-Rodriguez, et al. [17]	impl.	1	X	X	X	Cascade tree (path) of infection
S. A. Myers, et al. [15]	expl.	1	X	X	X	Prob. for a user u of infecting another user v with each $(u, v) \in E$
N. Barbieri, et al. [18]	expl.	1	O	X	X	Prob. for a user u of infecting another user v with each $(u, v) \in E$
K. Saito, et al. [2]	expl.	n	X	X	X	Prob. for a user u of infecting another user v with each $(u, v) \in E$
A. Goyal, et al. [4]	expl.	1	X	X	X	Prob. for a user u of infecting another user v with each $(u, v) \in E$
Our models: IPSI/ IPSI+/IPSI+S	expl.	n	X/O/O	O	X/X/O	Prob. of *internal and external* infection of information

epidemiology, that describes the change of population with differential equations and observes the change of diffusion rate by elapsed time from the first infection, is applied to study information diffusion process in online social networks. But since the goal of these techniques is not to measure the diffusion rate of individual propagations but to compute the overall speed of information flows in a network, it is hard to obtain accurate infection rate of each hashtag using these techniques.

In modeling diffusion over network, the Independent Cascade (IC) model is one of widely-studied models, initially proposed in the context of marketing [7], [21]. In the IC model, a newly infected node u has a single chance to infect its neighbor node v; the infection succeeds with probability $p_{u,v}$ that is given for every edge. The IC model is generalized in network SIR (Susceptible-Infectious-Recovered) model, where an infected node remains infectious until it gets recovered. However, these IC models are based on traditional problem setting whose purpose is to analyze the trend of information propagation, and where the diffusion occurs depending on the networks only.

We compare the IC models in six dimensions, A1 – A6. A1 is whether the models work on networks with explicit links or implicit ones. The models for implicit links generally aim to infer the hidden connections from the investigation of epidemiological data. A2 is whether the models allow repeated exposure to infections; most of the models assume that a node in a network has a single chance to be infected. A3 is whether the models consider the temporal decay of infection (i.e., an infected node has a decaying influence on its neighbors). A4 is whether the models consider external infections – that is, a node may be infected not from its neighbors but from external influence. A5 is whether a node can be recovered and be infected repeatedly. Most importantly, A6 compares the inference result of the models.

The column for A6 in the table shows that none of the existing IC models compute the internal (network) infection rate. The most closely related work [5] focuses on estimating the temporal trend of probability that a node is infected from external source; they assume that the internal (network) infection probability (and its change over time) is given.

The problem of inferring latent social networks from observed diffusion data is studied in [15] and [16]; especially [16] adopted the SI model and formulated the problem of finding a latent social network as a maximum likelihood problem. While their approach may look similar to ours in the problem formulation, they studied a different problem of finding latent networks and influence probabilities.

Topic-aware IC (TIC) model is proposed in [18] for the estimation of infection probability and they provide an EM algorithm for inference. The TIC model is then adopted and extended in many influence maximization techniques such as [22]–[24]. These models focus on finding a chance of maximizing influence and assume that a vertex has a single chance to infect its neighbors while our model assumes the multiple trials of infection through time.

The IC models extended in [2], [4], [18] define the likelihood functions that are similar to ours. However, their goal is to compute the tendency of each node for adopting information from its neighbors; those models do not estimate the diffusion rates of certain information in a network. Consequently the resulting EM algorithms shown in our paper are significantly different to these work.

Furthermore, the rumor propagation dynamics model proposed in [25] considers the anti-rumor information and user's psychological factors together. Although the model's goal is not to estimate the infection rate of individual hashtags as we do in this paper, the model successfully detects the trend of rumor's propagation. In [26] and [27], interesting influence models are developed to capture the behavior of hot topics evolving and spreading over social networks.

VII. CONCLUSION

We introduced realistic information diffusion models that suit well for the propagation of hashtags on social networks. Based on the diffusion models, we proposed inference algorithms to estimate the internal and external infection rates of individual hashtags. By extensive experiments using both real-life and synthetic data, we validated the effectiveness of the proposed algorithms.

REFERENCES

[1] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on Twitter," in *Proc. 7th Workshop Social Netw. Mining Anal. (SNAKDD)*, 2013, pp. 1–9.

[2] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. KES*, 2008, pp. 67–75.

[3] K. Zhu and L. Ying, "Information source detection in the SIR model: A Sample-Path-Based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2016.

[4] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM)*, 2010, pp. 241–250.

[5] 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 Mining (KDD)*, 2012, pp. 33–41.

[6] L. Bennett, *News: The Politics of Illusion* (Classics in Political Science), 7th ed. White Plains, NY, USA: Longman, 2006.

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

[8] D. A. Sprague and T. House, "Evidence for complex contagion models of social contagion from observational data," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0180802.

[9] D. Chandler, *Introduction to Modern Statistical Mechanics*. New York, NY, USA: Oxford Univ. Press, 1987.

[10] G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: A system for large-scale graph processing," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2010, pp. 135–146.

[11] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[12] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," in *Proc. Roy. Soc. London. A, Containing Papers Math. Phys.*, vol. 115, no. 772, pp. 700–721, 1927.

[13] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Netw.*, vol. 31, no. 2, pp. 155–163, May 2009.

[14] M. Richardson, R. Agrawal, and P. M. Domingos, "Trust management for the semantic Web," in *Proc. Int. Semantic Web Conf.*, 2003, pp. 351–368.

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

[16] X. Wu, A. Kumar, D. Sheldon, and S. Zilberstein, "Parameter learning for latent network diffusion," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, vol. 2013, pp. 2923–2930.

[17] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discovery from Data*, vol. 5, no. 4, pp. 1–37, Feb. 2012.

[18] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," *Knowl. Inf. Syst.*, vol. 37, no. 3, pp. 555–584, Apr. 2013.

[19] M. Fang, P. Shi, W. Shang, X. Yu, T. Wu, and Y. Liu, "Locating the source of asynchronous diffusion process in online social networks," *IEEE Access*, vol. 6, pp. 17699–17710, 2018.

[20] Y. Hu, R. J. Song, and M. Chen, "Modeling for information diffusion in online social networks via hydrodynamics," *IEEE Access*, vol. 5, pp. 128–135, 2017.

[21] J. Goldenberg, B. Libai, and E. Muller, "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata," *Acad. Marketing Sci. Rev.*, vol. 9, no. 3, pp. 1–18, 2001.

[22] C. C. Aslay, N. Barbieri, F. Bonchi, and R. A. Baeza-Yates, "Online topic-aware influence maximization queries," in *Proc. EDBT*, 2014, pp. 295–306.

[23] S. Chen, J. Fan, G. Li, J. Feng, K.-L. Tan, and J. Tang, "Online topic-aware influence maximization," *Proc. VLDB Endowment*, vol. 8, no. 6, pp. 666–677, Feb. 2015.

[24] W. Chen, T. Lin, and C. Yang, "Real-time topic-aware influence maximization using preprocessing," in *Proc. CSoNet*, 2015, pp. 1–13.

[25] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, "Rumor propagation dynamic model based on evolutionary game and anti-rumor," *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, Nov. 2018.

[26] Y. Xiao, N. Li, M. Xu, and Y. Liu, "A user behavior influence model of social hotspot under implicit link," *Inf. Sci.*, vol. 396, pp. 114–126, Aug. 2017.

[27] Y. Xiao, C. Song, and Y. Liu, "Social hotspot propagation dynamics model based on multidimensional attributes and evolutionary games," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 67, pp. 13–25, Feb. 2019.

YOUNGHOON KIM received the B.S. degree in computer science and engineering and the Ph.D. degree from Seoul National University, in 2006. He was a Visited Scholar with the University of Illinois at Urbana–Champaign, in 2014, invited by Prof. J. Han. He is currently an Assistant Professor with Hanyang University at ERICA, South Korea. He has presented his work in many top conferences on computer science known for their impact history and their rigorous review process, such as ICDE 2010 and 2012, WWW 2013, SIGMOD 2013, and KDD 2015. His research interests include machine learning with stochastic modeling, substring query processing in database, and social network analysis focusing on text mining.

JIWON SEO received the B.S. degree in electrical engineering from Seoul National University, in 2005, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, in 2008 and 2015, respectively. He is currently an Assistant Professor with the Computer Science Department, Hanyang University, South Korea. His research interests include big data analytic systems and large-scale deep learning systems.