Efficient Spread Size Approximation of Opinion Spreading in General Social Networks

Byeongjin Choe,[1] Yishi Lin,[2] Sungsu Lim,[3] John C.S. Lui,[2] and Kyomin Jung[1,*]

[1] *Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea*
[2] *Dept. of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, China*
[3] *Dept. of Computer Science and Engineering, Chungnam National University, Daejeon, Korea*

(Dated: November 6, 2019)

In contemporary society, understanding how information, such as trends and viruses, spreads in various social networks is an important topic in many areas. However, it is difficult to mathematically measure how widespread the information is, especially for a general network structure. There have been studies on opinion spreading, but many studies are limited to specific spreading models such as the susceptible-infected-recovered (SIR) model and the independent cascade model, and it is difficult to apply these studies to various situations. In this paper, we first suggest a general opinion spreading model (GOSM) that generalizes a large class of popular spreading models. In this model, each node has one of several states, and the state changes through interaction with neighboring nodes at discrete time intervals. Next, we show that many GOSMs have a *stable property* that is a GOSM version of a uniform equicontinuity. Then, we provide an approximation method to approximate the expected spread size for stable GOSMs. For the approximation method, we propose a *concentration theorem* that guarantees that a generalized mean-field theorem calculates the expected spreading size within small error bounds for finite time steps for a slightly dense network structure. Furthermore, we prove that a "*single simulation*" of running the Monte-Carlo simulation is sufficient to approximate the expected spreading size. We conduct experiments on both synthetic and real-world networks and show that our generalized approximation method well predicts the state density of the various models, especially in graphs with a large number of nodes. Experimental results show that the generalized mean-field approximation and a single Monte-Carlo simulation converge as shown in the concentration theorem.

I. INTRODUCTION

How political opinion, product adoption, rumors, trends and viruses spread through a society has been of fundamental interest for many years and has been studied in a wide variety of research disciplines. In particular, accurately analyzing the spread size of information such as a virus in a human society is one of the growing concerns in diffusion problems. For example, recent epidemics, such as severe acute respiratory syndrome (SARS), influenza A, and Ebola, show that viruses can easily spread on a global scale. Similarly, rumors and trends can also spread quickly and widely to alter human behavior. Furthermore, analyzing the spread size and density of information can be used to maximize the influence of the information. With the growing usage of online social networks (OSNs) and blog sites, predicting spread size to measure the influence of the information is taking on added significance. Many companies are now performing viral marketing on OSNs, and they rely on the word-of-mouth of adopters to influence other people to increase product sales. Since companies can only perform viral marketing with a limited advertising budget, predicting and evaluating the effectiveness of viral marketing strategies (in terms of number of final adopters) is of the utmost importance.

For the study of information spread size approximation problems, there have been a number of works on building spreading models through mathematical assumptions about social network structures and how information spreads in social networks. In these models, a person is represented by a node, and a relationship between people is represented by a graph structure of the nodes. Information such as whether or not a person is an iPhone user or is infected by an epidemic are assigned on each node as a state. For example, in the epidemic model, each person (i.e., node) can be assigned one of three states: infected, not infected, or recovered.

The *voter model* is a prime example of a spreading model [1, 2]. In this model, each node updates its state by following the state of a randomly chosen neighbor. The original voter model has only two possible states: 0 or 1. The general voter model [3, 4] generalizes this original model. One of the variants of the voter model is the Naming Game [5], in which each node has a set of states that evolves conditioned on its own state as well as its neighbors.

The *susceptible-infected-recovered* (SIR) model [6] is a well-known model for predicting the diffusion of epidemics. The SIS model is one of the variations of the SIR model. In this setting, each infected (or activated) node attempts to infect its neighbors independently and succeed with a fixed probability or rate. The *independent cascade model* [7] is also one of the models that aims to explain the opinion spreading process and can

* Corresponding author: kjung@snu.ac.kr;
K. Jung is with Automation and Systems Research Institute (ASRI), Seoul National University.

be regarded as an SIR model. In the independent cascade model, every influenced node has a single chance to influence each of its uninfluenced neighbors. Recently, the Hawkes point process model was used to analyze user activities in social networks [8, 9]. It is known that the rate of events in an extended model is identical to the rate of new infections in the SIR model [10].

Another well-known spreading model is the *linear threshold model* [11, 12], where each node possesses a real-valued threshold, to which the sum of incoming influences from neighboring nodes is compared at each iteration. If this combined influence is greater than the threshold, then the node becomes influenced (or activated). Kempe et al. [7] integrated both the threshold model and the cascade model. They revealed that the general threshold model is equivalent to the *general cascade model*, which is the generalized version of the independent cascade model.

However, evaluating the expected density of nodes with each state in the network is often difficult [11, 13, 14], especially for a general network structure and a general spreading model. Many analyses on these spreading models have been performed on networks satisfying certain properties that are very strong. For the linear threshold model, most of previous studies have focused on special graphs, such as complete graphs [11] and locally tree-like graphs [14–16]. Additionally, there is a vast literature on the SIR models for analyzing the spread on networks, including locally tree-like graphs [6, 17] and graphs with a clustered structure [18]. Schneider-Mizell and Sander [3] explored a general voter model on bipartite networks and random scale-free networks.

Therefore, there have been some studies to predict the state density in complex networks. Moretti et al. [4] analyzed the general voter model by the mean-field approximation on networks that disregard the specific connection pattern. Sahneh et al. [19] constructed a generalized epidemic spreading model for multistate and multilayer networks and provides a mean-field approximation for the overall density. In recent studies, a pairwise mean-field approximation [20–22] that improves the accuracy of the mean-field approximation has been studied. In the pairwise approximation, the approximation is performed through the probability of the state combination of two connected nodes to consider the dynamical correlation between neighbors [23, 24]. However, these studies focus only on the specific spreading model, so they have limitations in scalability. Moreover, the theoretical accuracy of the approximation is not properly shown.

In this paper, we first provide a general opinion spreading model (GOSM) that is a generalization of various discrete-time spreading models. The GOSM represents models that update each node's state with some probability that is dependent on the present node's state and its neighbors' states. Then, we focus on GOSMs that have the *stable property*, which can be seen as a GOSM version of a uniform equicontinuity. In essence, we define a GOSM as stable when slight changes in the state density of a set of nodes do not cause a dramatic change in the state transition probability in the next step. We prove that stable GOSM includes well-known spreading models, such as the voter model and the SIR model.

Next, we provide a concentration theorem showing that our generalized mean-field approximation converges to exact solutions with high probability for any stable GOSM. Specifically, the theorem proves that under the stable GOSM, for any initial node states, our generalized mean-field approximation is close to the true expected state density with probability $1 - o(n^{-\delta})$ for a certain δ, where n is the number of nodes. The concentration theorem also provides a theoretical background for applicability of the mean-field approximation. Moreover, surprisingly, we prove that just a *single simulation* of the Monte-Carlo simulation can efficiently approximate the true expected state density with any finite number of states and network structures that have degree $\omega(\log n)$. Previous mean-field approximation studies [4, 20–22] addressed only a limited set of GOSMs that cannot be approximated in other GOSMs, but our theorem is applicable to all stable GOSM.

To show that our concentration theorem is also proven experimentally, we demonstrate the effectiveness of generalized approximation and the single Monte-Carlo run via extensive experiments on both synthetic networks and real-world networks. Experimental results show that our generalized mean-field approximation is sufficiently accurate compared to the state-of-the-art approximation in many stable GOSMs such as the general voter model, the epidemic spreading model, and the daily active user (DAU) model [25]. We also experimentally show that a single simulation of the Monte-Carlo simulation can approximate the true expected state density sufficiently well.

Paper Organization. The rest of the paper is structured as follows. In Section II, we present the GOSM that can represent many state-change spreading models on networks. In Section III, we suggest the stable property for the GOSM and show many classical spreading models have the stable property. In Section IV, we present the concentration theorem, which shows a new framework of calculating a spreading model's state density, and we formally prove its validity. Our experimental results on both synthetic and real data are given in Section V. Finally, our conclusion is given in Section VI.

II. GENERAL OPINION SPREADING MODELS

In this section, we describe our class of *GOSMs* as a general framework for opinion spreading in social networks. Such a model is a special case of a discrete-time Markov process in which the future states of each node depend upon only the present states of that node and its neighbors, not on the sequence of present and past states. After we present the expression of GOSM, we show how several popular spreading models (e.g., the voter model

and the SIR and SIS models) can be easily represented under our GOSM.

In our model, we consider a given directed graph $G = (V, E)$, where V is a set of nodes (with $n = |V|$) and E is a set of directed edges. A directed edge e_{uv} from node u to node v implies that node u can influence node v, and we say that node u is an in-neighbor of node v. At each discrete time step $t = 0, 1, 2, \ldots$, each node $v \in V$ exists in a state drawn from a finite state space $S = \{0, 1, \ldots s-1\}$. Let $s_v(t) \in S$ be a state variable of node v at time t, and let $I_v^i(t) \in \{0, 1\}$ be an indicator variable such that $I_v^i(t) = 1$ if $s_v(t) = i$ and 0 otherwise. If $s = 2$, then the state of a node may correspond to an indicator of adoption, i.e., whether or not the node has adopted a certain opinion. If node v's state variable satisfies $s_v(t) = 0$, node v has not yet adopted the opinion, whereas if $s_v(t) = 1$, it has. The state of node v at time 0 is assigned based on the initial probability $\{a_v^i(0)\}$. Specifically, $Pr[s_v(0) = i] = a_v^i(0)$, independently of all other nodes. If $\{a_v^i(0)\}$ takes its value only from $\{0, 1\}$, then the initial state of node v is fixed. The state of node v at time $t+1$ is updated probabilistically, where the probability is determined only by the state of v and the states of its set of in-neighbors ("neighbors", for brevity), which we denote by $N(v)$, at time t. The indegree of v, which we denote by d_v, is simply $d_v = |N(v)|$.

Let us define the function f_v^k as the probability that v is in state k at time $t+1$ if the states of its neighboring nodes at time t are given, that is,

$$Pr[s_v(t+1) = k | \{s_u(t)\}_{u \in N(v)}] = \mathbb{E}[I_v^k(t+1)] = f_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}).$$

Then, $\sum_{k \in S} f_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}) = 1$. The inputs of the function f_v^k are the state indicators of the current node v and its neighbors u.

Next, we define $\bar{f}_v^k(\mathbf{x}_v, \{\mathbf{x}_u\}_{u \in N(v)})$, which is a natural extension of $f_v^k(\cdot)$ that extends the input variables from indicators $\{I_v^i(t)\}_{i \in S} \in \{0, 1\}^s$ to real numbers $\mathbf{x}_v = \{x_v^i\}_{i \in S} \in [0, 1]^s$. The indicator variables can also be the input of $\bar{f}_v^k$. Then,

$$\bar{f}_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}) = f_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}).$$

Examples are given in II A, and f_v^k and $\bar{f}_v^k$ can be derived for other models as well.

The goal of this paper is to estimate the expected state density:

$$\mathbb{E}\Big[\frac{1}{|W|} \sum_{u \in W} I_u^i(t)\Big] \quad \text{for a given } t,\ i \in S, \text{ and } W \subseteq V.$$

For example, suppose that we are considering the influence regarding the adoption of a product, where $S = \{0, 1\}$, with 1 meaning that the person adopts the product and 0 meaning otherwise. When we take $W = V$, we are interested in the fraction of all people in network G who adopt the product. If we consider W to be all female users in network G, then we are interested in the fraction of women who adopt the product. We assume that each node v has a given initial state probability $a_v^i(0)$ (i.e., $\mathbb{E}[I_v^i(0)] = Pr[I_v^i(0) = 1] = a_v^i(0)$) such that $\sum_{i \in S} a_v^i(0) = 1$.

Let us now show how several well-known spreading models can be represented by our model. Specifically, we will give the details of the mappings and how to analyze these models under our framework.

A. Simple Voter Model

A voter model provides a set of rules for contact-based spreading in a network. In statistical physics, such a model has also been used to study the phase transition phenomenon of a certain type of Ising model [26]. Let us first describe the *simple voter model* [1]. A node v updates its state by copying that of a randomly chosen neighbor. In each time step, node v adopts the state of its neighbor u with probability $1/d_v$, where d_v is the indegree of v. We can use the GOSM to describe the state-change rule for the simple voter model as follows:

$$f_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}) = \frac{1}{d_v} \sum_{u \in N(v)} I_u^k(t) \quad (1)$$

for all $k \in S$. The extended function $\bar{f}_v^k$ is given by

$$\bar{f}_v^k(\mathbf{x}_v, \{\mathbf{x}_u\}_{u \in N(v)}) = \frac{1}{d_v} \sum_{u \in N(v)} x_u^k. \quad (2)$$

B. General Voter Model

In addition to the simple voter model discussed above, there is also the general voter model, in which each edge e_{uv} has a *weight* w_{uv} and $\sum_{u \in N(v)} \omega_{uv} = d_v$. In the general voter model, node v selects its neighbor u as its reference neighbor with a probability proportional to w_{uv} in each time step. Let $p_{i,j,k}$ be the probability that node v's state will change from i to k if the state of the reference node u is j. Using the notation of our model, we have the following:

$$f_v^k(\{I_v^i(t)\}_{i \in S}, \{I_u^i(t)\}_{i \in S, u \in N(v)}) = \sum_{i \in S} I_v^i(t)[\sum_{j \in S} [p_{i,j,k} \frac{1}{d_v} \sum_{u \in N(v)} \omega_{uv} I_u^j(t)]]. \quad (3)$$

C. SIR Model and Independent Cascade Model

In the SIR model, each node is in one of three states: *susceptible*, *infected*, or *recovered*. Using our model, we

can label these states as states 0, 1, and 2, respectively. In each time step, a susceptible node v has an opportunity to be infected by each of its infected neighbors. Each infected neighbor u will succeed in infecting v with probability $\beta_{uv} \in [0,1]$. Furthermore, infected nodes will recover with probability γ; recovered nodes will not be infected again and lose the ability to infect others. The state-change rule for this model is as follows:

$$\begin{aligned}
&\text{(i) } f_v^0(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t) \prod_{u\in N(v)} (1 - I_u^1(t)\beta_{uv}), \\
&\text{(ii) } f_v^1(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t)(1 - \prod_{u\in N(v)} (1 - I_u^1(t)\beta_{uv})) + I_v^1(t)(1-\gamma), \\
&\text{(iii) } f_v^2(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) = I_v^1(t)\gamma + I_v^2(t).
\end{aligned} \tag{4}$$

In the *independent cascade* model [7], each node is in one of three states: *inactivated*, *activated*, or *already activated*, where a node in the last state has lost the ability to influence others. These three states of the independent cascade model correspond to the three states of the SIR model, and the dynamics of the independent cascade model are similar to stochastic SIR dynamics. In the original independent cascade model, $\gamma = 1$, meaning that an activated node always deactivates after it tries to influence its neighbors.

D. SIS Model and Its Generalized Form

In the *susceptible-infected-susceptible* (SIS) model, each node has two possible states: the *susceptible* state and the *infected* state, which we label as states 0 and 1, respectively. The SIS model is similar to the SIR model. However, in the SIS model, infected nodes have a chance to spontaneously revert to the susceptible state. Using our GOSM, we can specify the state-change rule for the SIS model as follows:

$$\begin{aligned}
&\text{(i) } f_v^0(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t) \prod_{u\in N(v)} (1 - I_u^1(t)\beta_{uv}) + I_v^1(t)\gamma, \\
&\text{(ii) } f_v^1(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t)(1 - \prod_{u\in N(v)} (1 - I_u^1(t)\beta_{uv})) + I_v^1(t)(1-\gamma).
\end{aligned}$$

A generalized SIS model [27] can also be expressed as a GOSM. Such a model considers a state space $S = \{0, 1, \ldots, s-1\}$; the infection rate between nodes in the states $\ell - 1$ and ℓ is β_ℓ, and each node in a nonzero state can revert to the zero state with probability γ. As an extension to the SIS model, this model can also be mapped to the GOSM. For instance, the state-change rule for a ternary SIS model is expressed by the following set of equations:

$$\begin{aligned}
&\text{(i) } f_v^0(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t) \prod_{u\in N(v)} (1 - I_u^1(t)\beta_1) + (1 - I_v^0(t))\gamma, \\
&\text{(ii) } f_v^1(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t)(1 - \prod_{u\in N(v)} (1 - I_u^1(t)\beta_1)) \\
&\quad + I_v^1(t)(\prod_{u\in N(v)} (1 - I_u^2(t)\beta_2))(1-\gamma), \\
&\text{(iii) } f_v^2(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = \Big(I_v^1(t)(1 - \prod_{u\in N(v)} (1 - I_u^2(t)\beta_2)) + I_v^2(t)\Big)(1-\gamma).
\end{aligned}$$

E. DAU Model

The *daily active user (DAU)* model [25] was proposed to capture the growth and decline of users in OSNs. In this model, each user can be in one of the following three states: *nonmember* (U), *inactive* (I), or *active* (A). We label them as states 0, 1, and 2, respectively. There are four state-change rules in the DAU model:

1. *Reaction*: If an inactive (I) user comes in contact with an active (A) user, the inactive (I) user will become active (A) with probability α.
2. *Decay*: An active (A) user can spontaneously become inactive (I) with probability β.
3. *Word-of-mouth reaction*: If a nonmember (U) user comes in contact with an active (A) user, the nonmember (U) user will become active (A) with probability γ.
4. *Media and marketing diffusion*: A nonmember (U) user can spontaneously become active (A) with probability λ.

The DAU model is a deterministic model and is based on the assumption that the underlying network is complete. However, one can easily extend the DAU model to a stochastic model on a noncomplete graph as follows. Suppose that in a given time step, node v comes in contact with its neighbor u. The state-change probabilities in the stochastic DAU model are as follows, as expressed for nodes v and $u \in N(v)$:

- Node v is a nonmember user: If node u is active, then node v will be activated with probability $\lambda + \gamma(1-\lambda)$. Otherwise, node v will become active with probability λ.
- Node v is inactive: If node u is active, then node v will be activated with probability α. Otherwise, the state of node v will not change.

- Node v is active: Node v will become inactive with probability β.

The state-change rule for the stochastic DAU model is expressed in our model by the following set of equations:

$$\begin{aligned}
&\text{(i) } f_v^0(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t)(1-\lambda-\frac{1}{d_v}\sum_{u\in N(v)} I_u^2(t)(1-\lambda)\gamma), \\
&\text{(ii) } f_v^1(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^1(t)(1-\frac{1}{d_v}\sum_{u\in N(v)} I_u^2(t)\alpha) + I_v^2(t)\beta, \\
&\text{(iii) } f_v^2(\{I_v^i(t)\}_{i\in S}, \{I_u^i(t)\}_{i\in S, u\in N(v)}) \\
&\quad = I_v^0(t)\lambda + I_v^0(t)(\frac{1}{d_v}\sum_{u\in N(v)} I_u^2(t)(1-\lambda)\gamma) \\
&\quad + I_v^1(t)(\frac{1}{d_v}\sum_{u\in N(v)} I_u^2(t)\alpha) + I_v^2(t)(1-\beta).
\end{aligned}$$

Therefore, the GOSM provides a way to express and analyze the behavior of the DAU model. The validity of this modeling is verified by the experiments reported in Section V.

III. STABLE PROPERTY

In this section, we define the concept of *stable property*, which is a variation of the uniform equicontinuity for the GOSM. This property is satisfied in well-known GOSMs such as the general voter model and the SIR model, as we prove in subsection VI.

In general, the family F of functions is called uniformly equicontinuous if for all $\epsilon > 0$, there exists $\delta > 0$ such that $d(f(x_1), f(x_2)) < \epsilon$, for all $f \in F$ and all $x_1, x_2 \in X$ such that $d(x_1, x_2) < \delta$. However, the stable property targets a weighted average of probability function $\{f(x_u)\}_{x_u\in X}$ in the GOSM instead of the general function $f(x_1)$. The definition of the stable property is as follows.

Definition 1 *(Stable Property) We say a model that has functions $\{\bar{f}_v^k(\mathbf{x}_v, \{\mathbf{x}_u\}_{u\in N(v)})\}_{v\in V, k\in S}$ is stable if for any $\epsilon > 0$, there exists a $\delta > 0$ depending only on ϵ and the set of spreading model functions $f_v^k(\cdot)_{v\in V}$, so that the following condition holds: for any vector sets $\{\mathbf{x}_v\}_{v\in V}$ and $\{\mathbf{y}_v\}_{v\in V}$ where $\mathbf{x}_v = \{x_v^i\}_{i\in S} \in [0,1]^s$ and $\mathbf{y}_v = \{y_v^i\}_{i\in S} \in [0,1]^s$, if*

$$\Big|\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv} x_u^k - \frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv} y_u^k\Big| \le \delta \qquad (5)$$

holds for all $k \in S$, all $v \in V$, and all $\{\bar{\omega}_{uv}\}_{u\in N(v)} \in K_v$. Then, we have

$$\begin{aligned}
&\Big|\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_v^k(\mathbf{x}_v, \{\mathbf{x}_u\}_{u\in N(v)}) \\
&\qquad - \frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_v^k(\mathbf{y}_v, \{\mathbf{y}_u\}_{u\in N(v)})\Big| \le \epsilon \qquad (6)
\end{aligned}$$

holds for all $v \in V$, all $k \in S$ and all $\{\bar{\omega}_{uv}\}_{u\in N(v)} \in K_v$ where $K_v = \{(\bar{\omega}_{uv})_{u\in N(v)} | \sum_{u\in N(v)} \bar{\omega}_{uv}^2 \le 4d_v, \bar{\omega}_{uv} \in \mathbb{R}_{++}\}$.

This stable property represents that a sufficiently small difference between weighted averages of inputs $\mathbf{x}_v$ and $\mathbf{y}_v$ induces a small difference in weighted average outputs.

Many popular GOSMs whose state change probability function according to the state ratio of neighboring nodes is equicontinuous have the stable property. For example, the general voter model satisfies the stable property because if we determine δ as $(\sum_{i,j\in S} 3p_{i,j,k})\delta = \epsilon$, then the stable property holds under the condition of $\sum_{u\in N(v)} {\omega_{uv}}^2 \le 4d_v$. Additionally, the SIR model also has the stable property where $d_v \max(\beta_{uv}, u \in N(v)) \le 2$, $\beta_{uv} < 0.98$. Detailed proofs are described in the following subsection. One of the spreading models, the majority rule model [28] which has an extremely skewed state change rule, does not have a stable property. In the majority rule, assuming there are only two states, the majority rule's state change probability function is not continuous when the state ratio of neighboring nodes is 0.5. For another example, let us assume that there is a state change rule that randomly selects one of the states of neighboring nodes in general, but when 80% of neighboring nodes becomes one state, follows the state. This rule also does not satisfy the stable property.

Proofs of stability for popular GOSMs such as the general voter model and the SIR model are included in appendix A.

IV. CONCENTRATION THEOREM FOR GENERAL OPINION SPREADING MODELS

In this section, we introduce a formula for approximating the expected state density of nodes under a stable GOSM. Moreover, we provide a concentration theorem that guarantees the error bound of approximations.

According to the stability property of a GOSM, if the current state density and its approximation are sufficiently similar, we can compute a sufficiently similar approximate value of the state density from the approximation of the current state density in the next time step. On this basis, we present a state density approximation $a_v^k(t)$ as follows.

Definition 2 *(Our approximation of the state density) We define $a_v^k(t+1) \in [0,1]$, our approximation of $\mathbb{E}[I_u^i(t+1)]$ (i.e., $Pr[I_v^k(t+1) = 1]$), as follows:*

$$a_v^k(t+1) = \bar{f}_v^k(\{a_v^i(t)\}_{i\in S}, \{a_u^i(t)\}_{i\in S, u\in N(v)}), \qquad (7)$$

where $\{a_v^k(0)\}$ are the initial state probabilities.

The following theorem shows that this approximation for the next time step can be extended over multiple time steps. In other words, if a model has the stability property, then with high probability, our approximation of the state density is very close to the expected state density for a finite number of time steps for an arbitrary network structure of node degree $\omega(\log n)$.

Theorem 1 *(The concentration theorem) Let $G = (V, E)$ be a given directed graph, and let S be the set of state vectors, with $n = |V|$ and $s = |S|$. Consider a subset of nodes $W \subseteq V$, with $m = |W|$. Suppose that $m = \omega(\log n)$ and $d_v = \omega(\log n)$ for all $v \in V$. If $\{\bar{f}_v^k(\cdot)\}_{v\in V,k\in S}$ has the stability property, then for any initial state probabilities $\{a_v^i(0)\}$, any constant $T \in \mathbb{N}$, any $\epsilon > 0$, and a certain $\delta > 0$, we have*

$$Pr\left[0 \le t \le T, \forall i \in S, \left|\frac{1}{m}\sum_{u\in W} a_u^i(t) - \mathbb{E}\left[\frac{1}{m}\sum_{u\in W} I_u^i(t)\right]\right| \le \epsilon\right] = 1 - o(n^{-\delta}).$$

Here, $\frac{1}{m}\sum_{u\in W} a_u^i(t)$ is our approximation of the state density of W, and $\frac{1}{m}\sum_{u\in W}\mathbb{E}[I_u^i(t)]$ is the expected state density that we want to compute.

By a simple modification of the proof, Theorem 1 can also be applied to compute a weighted state density in the case that there is a weight associated with each neighboring node.

Outline of the proof. To prove the above theorem, we prove Lemma 4, which states that for any node $v \in V$ and any time step t, the observed state density of all neighboring nodes' states, $\frac{1}{d_v}\sum_{u\in N(v)} I_u^i(t)$, is close to the expected state density, $\frac{1}{d_v}\sum_{u\in N(v)}\mathbb{E}[I_u^i(t)]$, for all nodes $v \in V$ with high probability.

To prove Lemma 4, we consider the approximated state density $\frac{1}{d_v}\sum_{u\in N(v)} a_u^i(t)$, which is close to the expected state density, and prove that it is close to $\frac{1}{d_v}\sum_{u\in N(v)} I_u^i(t)$ with high probability.

Lemma 4 can be proven through mathematical induction. Lemma 2 is the first step in the inductive method; in Lemma 2, we derive an error bound for the approximated state density of weighted neighboring nodes at $t = 0$. Lemma 3 is an inductive step that proves that the one-time observed state density is close to the approximated state density at time $t + 1$ under the given condition at time t.

To prove Theorem 1, we consider the following graph G' and apply Lemma 4. Suppose that a new node $v' \notin V$, which has only inward edges from $u \in W$, is added to G. We call this graph G'. Node v' can be influenced by all nodes in W but does not influence any nodes. Formally, we define a directed graph $G' = (V', E')$, where $V' = V \cup \{v'\}$ and $E' = E \cup \{e_{uv'}\}_{\forall u\in W}$. We will apply Lemma 4 to G' to prove Theorem 1.

Formula (7) and the algorithm that computes $a_v^i(t)$ based on (7) can be understood as a generalized mean-field approximation for a nonsymmetric network structure and arbitrary initial states. For example, if we consider the simple voter model and assume that for each $i \in S$, the $\{a_v^i(0)\}$ are the same for all $v \in V$, then for all times t and $i \in S$, the $a_v^i(t)$ become the same for all v, as in the usual mean-field approximation.

Proof of Theorem 1. We define

$$r_v^i(t) = \frac{1}{d_v}\sum_{u\in N(v)} \omega_{uv} I_u^i(t)$$

to denote the observed state density of v's neighboring nodes with weights $\omega_{uv} \ge 0$, $(\omega_{uv})_{\forall u\in N(v)} \in K_v$. By the linearity of expectation, we have $\mathbb{E}[r_v^i(t)] = \frac{1}{d_v}\sum_{u\in N(v)} \omega_{uv}\mathbb{E}[I_u^i(t)]$. We define

$$b_v^i(t) = \frac{1}{d_v}\sum_{u\in N(v)} \omega_{uv} a_u^i(t)$$

as the approximation of $r_v^i(t)$. For the case in which all weights ω_{uv} are equal to 1, we denote the corresponding quantities by $\bar{r}_v^i(t)$ and $\bar{b}_v^i(t)$. Let $\min(d_v, v \in V) = d_{v,min}$.

Lemma 2 *For any initial state probabilities $\{a_v^i(0)\}$ and any $\{\omega_{uv}\}_{u\in N(v)} \in K_v$,*

$$Pr\left[\forall v \in V, \forall i \in S, \left|r_v^i(0) - b_v^i(0)\right| \le \epsilon_0\right] \ge 1 - 2sn\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right). \quad (8)$$

Lemma 2 holds for $t = 0$ and serves as the basis for induction. We will now prove Ineq. (8), which shows that the probability of satisfying $\left|r_v^i(0) - b_v^i(0)\right| \le \epsilon_0$ is sufficiently high.

The initial state $s_v(0)$ is determined only by the given initial probabilities $\{a_v^i(0)\}_{i\in S}$. The initial probabilities for each node are given values and do not depend on the initial probabilities for any other node. Therefore, the initial states of all nodes are mutually independent. We apply Hoeffding's inequality since $r_v^i(0)$ can be described as a linear combination of independent indicator variables $I_v^i(0)$ for a given $i \in S$. Therefore, for all $v \in V$ and all $i \in S$, $r_v^i(0) = \sum_{u\in N(v)} \frac{1}{d_v}\omega_{uv} I_u^i(0)$ and $Pr(\frac{\omega_{uv}}{d_v} I_v^i(0) \in [0, \frac{\omega_{uv}}{d_v}]) = 1$ are satisfied. Then, we have

the following inequality at $t = 0$ for any $i \in S$:

$$\begin{aligned} &Pr\left[\left|\frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}I_u^i(0) - \frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}a_u^i(0)\right| \geq \epsilon_0\right] \\ &= Pr\left[\left|\frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}I_u^i(0) - \frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}\mathbb{E}[I_u^i(0)]\right| \geq \epsilon_0\right] \\ &\leq 2\exp\left(-\frac{2{\epsilon_0}^2}{\sum_{u\in N(v)}\left(\omega_{uv}/d_v\right)^2}\right) \leq 2\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right). \end{aligned} \tag{9}$$

Hoeffding's inequality shows that the observed state density among v's neighbors rarely deviates from the expected state density among those neighbors. By the union bound, we obtain

$$\begin{aligned} &Pr\left[\forall v \in V, \forall i \in S, \left|r_v^i(0) - b_v^i(0)\right| \geq \epsilon_0\right] \\ &\qquad \leq 2sn\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right) \end{aligned} \tag{10}$$

and

$$\begin{aligned} &Pr\left[\forall v \in V, \forall i \in S, \left|r_v^i(0) - \mathbb{E}[r_v^i(0)]\right| \leq \epsilon_0\right] \\ &\quad = Pr\left[\forall v \in V, \forall i \in S, \left|r_v^i(0) - b_v^i(0)\right| \leq \epsilon_0\right] \\ &\quad \geq 1 - 2sn\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right). \end{aligned} \tag{11}$$

■

The next lemma is an inductive step that shows that if the presented statement holds for some natural number t, then it also holds for $t+1$.

Lemma 3 *For $t = 0, 1, 2, \ldots, T$, if $\left|r_v^i(t) - b_v^i(t)\right| \leq \epsilon_t$ is satisfied for all $v \in V$, all $i \in S$, and all $\{\omega_{uv}\}_{u\in N(v)} \in K_v$, then for any $\{\omega_{uv}\}_{u\in V} \in K_v$,*

$$\begin{aligned} &Pr\Big[\forall v \in V, \forall i \in S, \left|r_v^i(t+1) - b_v^i(t+1)\right| \leq \epsilon_{t+1} \\ &\qquad \left|\forall v \in V, \forall i \in S, \left|r_v^i(t) - b_v^i(t)\right| \leq \epsilon_t\right] \\ &\qquad \geq 1 - 2sn\exp\left(-\frac{2{\epsilon_t}^2 d_{v,min}}{4}\right). \end{aligned} \tag{12}$$

Proof of this lemma is in appendix B.

Lemma 4 *Let $G = (V, E)$ be a given directed graph, and let S be the set of state vectors, with $n = |V|$ and $s = |S|$. Suppose that $d_v = \omega(\log n)$ for all $v \in V$. If $\{\bar{f}_v^k(\cdot)\}_{v\in V}$ has the stability property, then for any initial state probabilities $\{a_v^i(0)\}$, any constant $T \in \mathbb{N}$, any $\epsilon > 0$, and a certain $\delta > 0$, we have*

$$\begin{aligned} &Pr\Big[0 \leq t \leq T, \forall v \in V, \forall i \in S, \left|\bar{r_v^i}(t) - \mathbb{E}[\bar{r_v^i}(t)]\right| \leq \epsilon\Big] \\ &\qquad \geq 1 - o(n^{-\delta}). \end{aligned} \tag{13}$$

We prove this lemma by applying mathematical induction to Lemmas 2 and 3. The remaining proof of this lemma is in appendix C.

Remaining proof of Theorem 1. Recall the graph $G' = G \cup v'$ and $v' \notin V$, which has only inward edges from $u \in W$. By applying Lemma 4 to graph G', we obtain, for any $v \in V'$,

$$\begin{aligned} &Pr\big[0 \leq t \leq T, \forall i \in S \left|\bar{a_v^i}(t) - \mathbb{E}[\bar{r_v^i}(t)]\right| \leq \epsilon\big] \\ &\geq Pr\big[0 \leq t \leq T, \forall v \in V', \forall i \in S, \left|\bar{a_v^i}(t) - \mathbb{E}[\bar{r_v^i}(t)]\right| \leq \epsilon\big] \\ &= 1 - o((n+1)^{-\delta}). \end{aligned}$$

Therefore, for node v',

$$\begin{aligned} &Pr\left[0 \leq t \leq T, \forall i \in S, \left|\frac{1}{m}\sum_{u\in W}a_u^i(t) - \mathbb{E}\left[\frac{1}{m}\sum_{u\in W}I_u^i(t)\right]\right| \leq \epsilon\right] \\ &\qquad \geq 1 - o(n^{-\delta}). \end{aligned} \tag{14}$$

■

In the previous arguments, we have assumed that each node has an independent initial state distribution $a_v^i(0)$. Note, however, that Theorem 1 is also applicable to a model in which the initial states of the nodes are deterministic. Since deterministic initial node states are expressed as $a_v^i(0) = I_v^i(0)$, Lemma 2 still holds when the initial states of the nodes are given deterministically.

Corollary 5 *(Single Monte Carlo simulation) Let $G = (V, E)$ be a given directed graph, and let S be the set of state vectors, with $n = |V|$ and $s = |S|$. Consider a subset of nodes $W \subseteq V$, with $m = |W|$. Suppose that $m = \omega(\log n)$ and $d_v = \omega(\log n)$ for all $v \in V$. If $\{\bar{f}_v^k(\cdot)\}_{v\in V, k\in S}$ has the stability property, then for any initial state probabilities $\{a_v^i(0)\}$, any constant $T \in \mathbb{N}$, any $\epsilon > 0$, and a certain $\delta > 0$, we have*

$$\begin{aligned} &Pr\left[0 \leq t \leq T, \forall i \in S, \left|\frac{1}{m}\sum_{u\in W}I_u^i(t) - \mathbb{E}\left[\frac{1}{m}\sum_{u\in W}I_u^i(t)\right]\right| \leq \epsilon\right] \\ &\qquad = 1 - o(n^{-\delta}). \end{aligned}$$

Here, $\frac{1}{m}\sum_{u\in W}I_u^i(t)$ is the state density of W in one Monte Carlo simulation. The above theorem states that we can approximate the state density through just one simulation of a Monte Carlo simulation because each observed density $\frac{1}{m}\sum_{u\in W}I_u^i(t)$ will have a similar value to the expected state density with high probability.

V. EXPERIMENTAL RESULTS

In this section, we present the results of empirical verification of our theorems and algorithms. The models we examine are the *SIS model*, the simple voter model, and the DAU model described in previous sections. We

will show that our approximation can well predict the expected state density of a set of nodes.

Datasets. The datasets used in the experiments include four synthetic random networks and two real networks. The synthetic random networks were generated using the *Barabasi-Albert* (BA) model and the *Watts-Strogatz* (WS) model. We downloaded the real-world networks from the Stanford Large Network Dataset Collection [29]. Slashdot is a technology-related news website known for its specific user community. In 2002, Slashdot introduced the Slashdot Zoo feature, which allows users to tag each other as friends or foes. The corresponding network consists of friend/foe links between Slashdot users. Gowalla is a location-based social networking website where users share their locations by checking in. Table I summarizes the basic statistics of the networks used in our experiments.

dataset	type	# of nodes	# of edges
BA_{1000}	undirected	1,000	3,990
BA_{10000}	undirected	10,000	39,990
$WS_{1000,10}$	undirected	1,000	10,000
$WS_{10000,100}$	undirected	10,000	1,000,000
Gowalla	undirected	196,591	950,327
Slashdot	directed	77,360	905,468

TABLE I: Datasets

Experimental setup. For each experiment, we first predicted the state density based on our approximation $a_u^s(t)$. Next, we ran one Monte Carlo simulation. $I_u^s(t)$ is an indicator function indicating whether node u is in state s at time t in a simulation run. Then, we ran $1,000$ additional Monte Carlo simulations to estimate the probability of each node being in each state in each time step. More specifically, we used the relative frequency of node u being in state s at time t as an estimate of $\mathbb{E}[I_u^s(t)]$. According to the *Chernoff bound*, the relative error of the above estimation method is insignificant; therefore, we ignored it in our experimental analysis. For any given subset W of nodes in the network, we compared our approximation, $\frac{1}{|W|}\sum_{u\in W} a_u^s(t)$, and the result of a single Monte Carlo simulation, $\frac{1}{|W|}\sum_{u\in W} I_u^s(t)$, with the true expected state density, $\frac{1}{|W|}\sum_{u\in W} \mathbb{E}[I_u^s(t)]$, for each state s in each time step t. According to Theorem 1, there should be no significant differences between these three values with high probability.

A. Results for the SIS Model

In this subsection, we compare the results of our generalized mean-field approximation (GMF), a single Monte Carlo simulation run and pairwise(pair-quenched) approximation (PWA) for the *SIS model*. The model we consider here has two states: a *susceptible* state, denoted by S, and an *infected* state, denoted by I. Initially, all

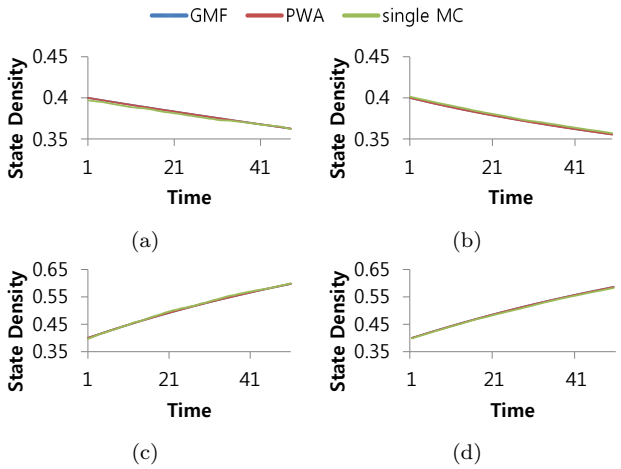

FIG. 1: "S" state density for the SIS model with the parametric value sets P_1 and P_2 and initial state distribution $D_1 = (0.4, 0.6)$. (a) P_1, BA_{10000}, (b) P_1, *Slashdot*, (c) P_2, BA_{10000}, (d) P_2, *Slashdot*.

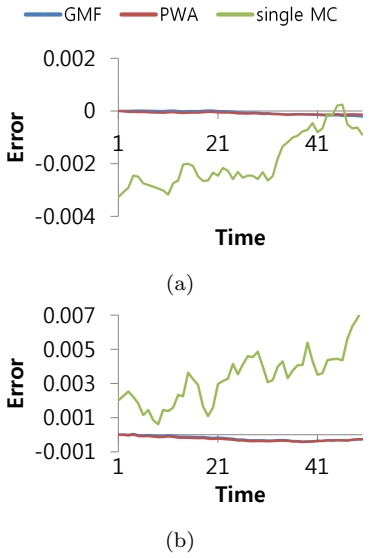

FIG. 2: Error with respect to the true expected density of state S for the SIS model with parametric value sets (a) P_1 and (b) P_2 and initial state distribution $D_1 = (0.4, 0.6)$ on the BA_{10000} dataset.

nodes in the network have a uniform initial state distribution of $D_1 = (S = 0.4, I = 0.6)$.

Figure 1 shows the results obtained on the synthetic undirected network BA_{10000} and on the real directed network *Slashdot*. We use two sets of parametric values for this model, denoted by P_1 and P_2. We set $P_1 = (\beta = 0.0005, \gamma = 0.01)$ and $P_2 = (\beta = 0.0005, \gamma = 0.0001)$. Figure 2 show the differences in value between the true state density and the results of the three approximation methods with the parametric value sets P_1 and P_2.

For both sets of parametric values, there is no significant difference between the true expected state density and the density predicted by generalized mean-field ap-

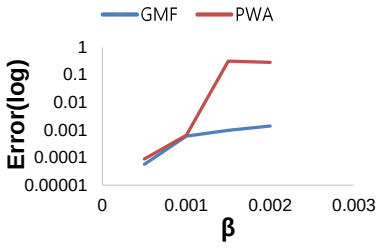

FIG. 3: Error with respect to the true expected density of state S at time $t = 50$ for the SIS model with a fixed value of $\gamma = 0.001$, various values of β and initial state distribution $D_1 = (0.4, 0.6)$.

proximation. The error of generalized mean-field approximation is less than 0.1% for various experimental environments. Pairwise approximation also approximates the true expected state density with an error of less than 0.1% in most cases, similar to the generalized mean-field approximation. Therefore, generalized mean-field approximation method approximates the state density under various SIS model parameters and able to perform as good as the pairwise approximation. In addition, in some cases, error of generalized mean-field approximation is smaller than the pairwise approximation. If the infection probability β is large, as shown in Figures 3, the error is drastically larger than in other cases. We believe that the reason for this phenomenon is because the experiment was conducted in a discrete environment. The pairwise approximation assumes a continuous time step, but larger β represents that the experiment move away from continuous time step environment.

As shown in Figure 2, running a Monte Carlo simulation once also yields an approximation of the true expected state density but with a larger error and greater instability over time than generalized mean-field approximation. Therefore running one Monte Carlo simulation is one of the effective approximation although the error is larger than that of generalized mean-field approximation.

B. Results for the Simple Voter Model

In this subsection, we present the results for a simple voter model with two states, the *positive* state and the *negative* state. We also refer to this model as the *binary voter model*. We focus on a special initial condition such that each node initially has a probability of 0.5 of being *positive* and a probability of 0.5 of being *negative*. In this case, we have $\mathbb{E}\left[\frac{1}{|W|}\sum_{u\in W} I_u^i(t)\right] = \mathbb{E}\left[\frac{1}{|W|}\sum_{u\in W} I_u^i(0)\right]$ for all $W \subseteq V$, $t > 0$ and $s \in \{\text{positive}, \text{negative}\}$.

Figure 4 shows the results for the binary voter model. There is little difference between generalized mean-field approximation and the expected state ratio on the various networks. Therefore, generalized mean-field approximation is very effective in predicting the expected state

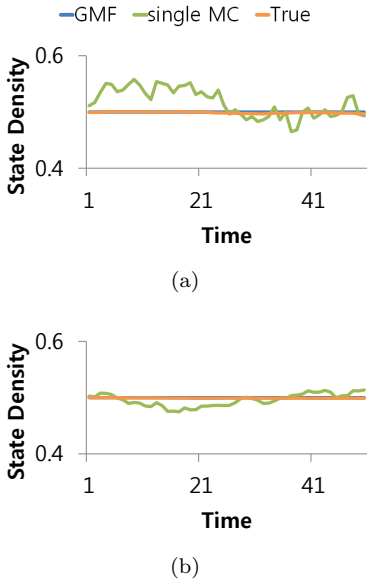

FIG. 4: Positive state density for the binary voter model with datasets (a) $WS_{1000,10}$ and (b) $WS_{10000,100}$.

density in this experimental setting. In addition, these experiments show how the accuracy of the prediction obtained through one Monte Carlo simulation run depends on the size of the graph. Recall that Corollary 5 provides a lower bound on the probability such that we can predict the state density with a small error by running one Monte Carlo simulation. For a given T, the lower bound on this probability approaches 1 as the number of nodes in the graph becomes sufficiently large. Figure 4 demonstrates the correctness of the corollary by showing that the accuracy of the prediction obtained from a single Monte Carlo simulation run does increase as the graph size increases.

C. Results for the DAU Model

We show the results for the stochastic version of the *DAU model* in this subsection. We examine how the results of a single Monte Carlo simulation behave for networks of different sizes and structures. In the experimental results, the states U, I and A represent the *nonmember*, *inactive* and *active* states, respectively.

Parametric values for the DAU model. We use two sets of parametric values, denoted by P_1 and P_2. We set $P_1 = (\alpha = 0.2, \beta = 0.05, \gamma = 0.08, \lambda = 0.001)$ and $P_2 = (\alpha = 0.02, \beta = 0.05, \gamma = 0.001, \lambda = 0.08)$. In accordance with the analysis of the DAU model, the model with parametric value set P_1 is characterized as a model with *self-sustainability and gradual word-of-mouth growth*, which means that the density of the active state will converge to a positive value. By contrast, the DAU model with parametric value set P_2 is characterized as a model with *unsustainable and intense media-and-*

marketing-driven growth, which means that the number of active users will initially increase but converge to zero in the steady state.

We use a *uniform initial state distribution* to define the initial conditions for the nodes. In a *uniform initial state distribution*, all nodes have the same probability of being in any given state. We use two *uniform initial state distributions*, one for each set of parametric values.

Figures 5 and 6 show the results for all six datasets with parametric value set P_1 and a uniform initial state distribution of $D_1 = (nonmember = 0.5, inactive = 0.4, active = 0.1)$. The initial state of each node is independently chosen from the distribution D_1. For example, each node initially has a probability 0.4 of being inactive and a probability 0.1 of being active. For simplicity, we also write D_1 as $D_1 = (0.5, 0.4, 0.1)$.

Figure 5 shows that for $BA10000$ and *Gowalla* tested, we can accurately *predict* the state density in any time step by running only one Monte Carlo simulation. The results obtained on the $WS_{1000,10}$, $WS_{10000,100}$, $BA1000$ and *Slashdot* datasets are similar as shown in Figure 10 in appendix. One should note that for real networks, even if a network is weakly connected, there still exists a significantly large state density with very few incoming edges because of the power-law degree distributions in real networks. For example, approximately 25% of the nodes in the undirected *Gowalla* network have only one neighbor. Figure 5 (b) show that even if the degree distribution of a real network does not satisfy the minimum degree requirement of Theorem 1, a single run of a Monte Carlo simulation can still yield a highly accurate prediction of the state density.

Figure 6 shows the results obtained when W consists of the top 1%, top 5% and top 10% of the nodes ranked by (in)degree.

Note that the state density of the *nonmember* state converges to zero, while the state densities of the other two states converge to one. Moreover, for the DAU model, we are mainly interested in the fraction of users who are active. Therefore, we show the fraction of active users as a function of t. From Figure 6, we can observe that when W contains at least the top 5% of the nodes as ranked by degree, then a single Monte Carlo simulation run can accurately predict the expected density of the active state. If W contains only the top 1% of the nodes as ranked by degree, then the predicted density of active nodes may not always be accurate. This inaccuracy arises because there is an implicit constraint on the cardinality of W in our main theorem. However, in this case, a single Monte Carlo simulation run can still be used to predict the trend of the change in density as well as the approximate density of active nodes in the steady state.

We now consider another set of experimental configurations. Figure 7 shows the results obtained on $BA10000$ and *Gowalla* with parametric value set P_2 and a *uniform initial state distribution* of $D_2 = (nonmember = 0.7, inactive = 0.1, active = 0.2)$. The results obtained

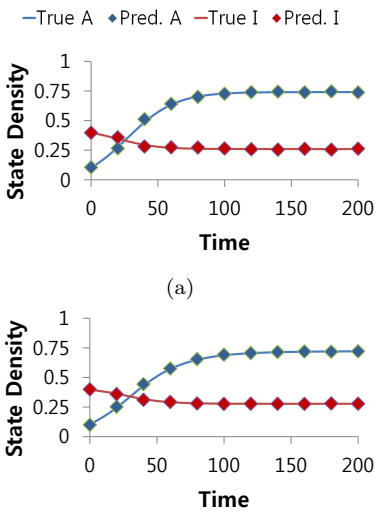

FIG. 5: State density for the DAU model with the *self-sustainable* parametric value set P_1 and the initial state distribution $D_1 = (0.5, 0.4, 0.1)$. The set W contains all nodes with nonzero (in)degrees. (a) BA_{10000}, (b) *Gowalla*.

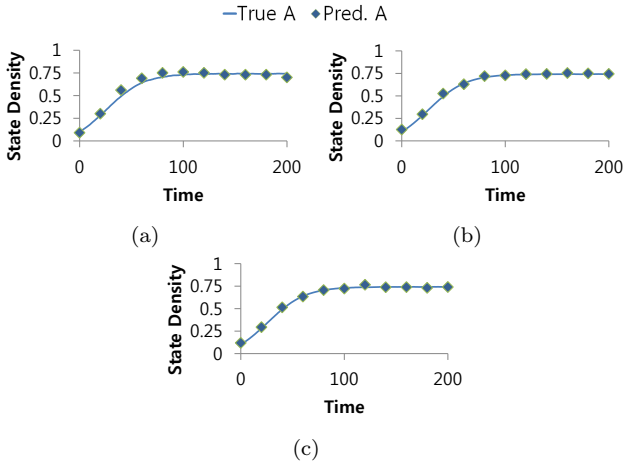

FIG. 6: State density for the DAU model with the *self-sustainable* parametric value set P_1 and the initial state distribution $D_1 = (0.5, 0.4, 0.1)$. The set W contains the (a) top 1%, (b) 5% or (c) 10% of the nodes as ranked by (in)degree.

on the $WS_{1000,10}$, $WS_{10000,100}$, $BA1000$ and *Slashdot* datasets are similar as shown in Figure 11 in appendix. From Figure 7, we can see that for parametric value set P_2 and initial state distribution D_2, we can accurately predict the state density by running a Monte Carlo simulation only once. Moreover, the prediction is very accurate for any arbitrary time step on all datasets tested.

Effect of initial conditions. We have previously considered two different *uniform initial state distributions*. Now, we consider a *nonuniform initial state distribution* as follows. For each network, the nodes with the top

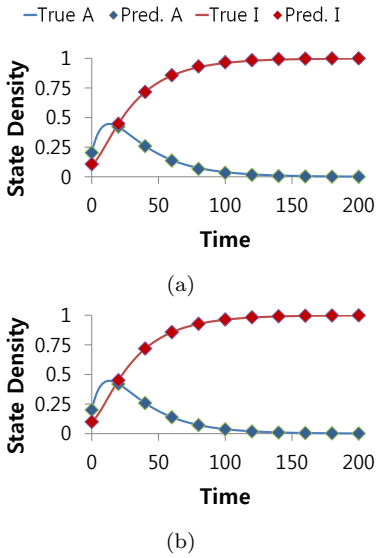

FIG. 7: State density for the DAU model with the *unsustainable* parametric value set P_2 and the initial state distribution $D_2 = (0.7, 0.1, 0.2)$. The set W contains all nodes with nonzero (in)degrees. (a) BA_{10000}, (b) *Gowalla*.

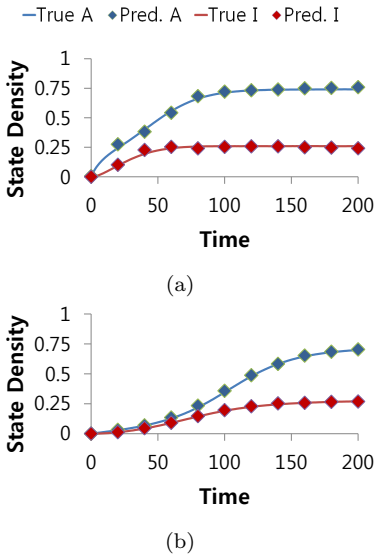

FIG. 8: State density for the DAU model with the *unsustainable* parametric value set P_1 and datasets (a) BA_{10000} and (b) *Gowalla*. Initially, the nodes with the top 50 highest (in)degrees are selected as active nodes, while all other nodes are in the "nonmember" state. The set W contains all nodes with nonzero (in)degrees.

50 highest degrees are initially selected as active nodes, while all other nodes are initially in the "nonmember" state. Figures 8 and 9 show the results obtained on the BA_{10000} and *Gowalla* datasets with this initial condition. Here, the DAU model has parametric value set P_1. Figures 8 and 9 show that for all datasets, a single Monte Carlo simulation run can provide an accurate prediction of the density of any given state in the steady state for the case in which W contains all nodes with nonzero (in)degrees and for the case in which W contains at least the top 10% of the nodes as ranked by (in)degree. Figure 9 shows that if W contains a small fraction of the nodes with the highest (in)degrees, e.g., the top 1% of the nodes, then a single Monte Carlo simulation run can accurately predict the expected density of the active state for most time steps. The results obtained on the $WS_{10000,100}$, *Gowalla* and *Slashdot* datasets are similar.

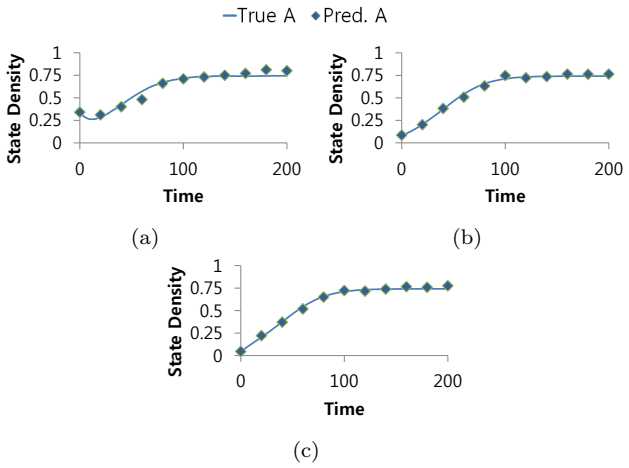

FIG. 9: State density for the DAU model with the *unsustainable* parametric value set P_1 and dataset $BA10000$. Initially, the nodes with the top 50 highest (in)degree are selected as active nodes, while the other nodes are in the "nonmember" state. The set W contains the (a) top 1%, (b) 5% or (c) 10% of the nodes as ranked by (in)degree.

Consider a given network $G = (V, E)$. If the minimum degree among all nodes in G is $\omega(\log |V|)$, then the degree requirement in Theorem 1 is satisfied. Then, for any node set W containing a large number of the nodes in V, a single Monte Carlo simulation run is sufficient to accurately predict the expected state density in any time step. Otherwise, suppose that the minimum degree requirement in Theorem 1 is not satisfied for the given network G. For example, the dataset may be very sparse, or it may contain a large fraction of nodes with very low (in)degrees. The experimental results demonstrate that even in this case, for a given W containing a large state density or nodes with high (in)degrees, a single Monte Carlo simulation run is still sufficient to predict the state density to a certain accuracy. Moreover, for the case in which W contains only a small fraction of the nodes in the network, e.g., only 1%, one Monte Carlo simulation run is still sufficient to effectively predict the trend of the change in the expected state density.

VI. CONCLUSION

In this paper, we have presented a GOSM and analyzed the opinion spread size of GOSMs for general network structures. Many discrete-time spreading models, such as the voter model, the DAU model, the independent cascade model, and the SIS model, can be formulated as GOSMs. In addition, we have shown that many well-known GOSMs have the stability property, which is the GOSM equivalent of uniform continuity. We prove that for a slightly dense network, our generalized mean-field approximation method can successfully compute the state density. Our approximation is applicable to all stable GOSMs, unlike the conventional mean-field approximation. Extensive experiments confirm that generalized mean-field approximation method indeed computes the state densities for the given nodes very well in practice. We have shown that a single Monte Carlo simulation run can also be used as a cost-efficient means of approximating the state density both theoretically and practically. Our results can be applied to various practical problems arising in social networks, including epidemic spreading analysis, the influence maximization problem [30], and viral marketing [31]. Two possible avenues for future work include relaxing the structure condition that requires a slightly dense network and generalizing our results to continuous-time Markovian opinion spreading models.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No.10073144) and by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (No.2016M3C4A7952632).

Appendix A: Proof of Stability for GOSMs

1. General Voter Model

For example, we consider two cases of OSM, the general voter model and the cascade model. First, we formally prove that the general voter model is stable. Since the DAU model is a special case of the general voter model, the stability of the DAU model follows directly.

Lemma 6 *If $f_v^k(\cdot)$ is given as (3) and for all $v \in V$, $\sum_{u\in N(v)} {\omega_{uv}}^2 \le 4d_v$, then $\{\bar{f}_v^k(\cdot)\}_{v\in V}$ is stable.*

Proof. With the Definition 1's notation, let δ be

$$\delta = \frac{\epsilon}{\sum_{i\in S}\sum_{j\in S} 3p_{i,j,k}}.$$

From the condition of stable property (5), we have for all $k \in S$, all $v \in V$, and all $\{\bar{\omega}_{uv}\}_{u\in N(v)} \in K_v$,

$$\left|\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}x_u^k - \frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}y_u^k\right| < \delta.$$

The objective of this proof is to show Ineq. (6). That is, for any $v \in V$, $k \in S$, and $\{\bar{\omega}_{uv}\}_{u\in N(v)} \in K_v$, we have

$$\left|\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_v^k(\mathbf{x}_v, \{\mathbf{x}_u\}_{u\in N(v)}) - \frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_v^k(\mathbf{y}_v, \{\mathbf{y}_u\}_{u\in N(v)})\right| \le \epsilon.$$

We can rewrite $\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u, \{\mathbf{x}_w\}_{w\in N(u)})$ as follows:

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u, \{\mathbf{x}_w\}_{w\in N(u)}) \\
&= \frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\sum_{i\in S} x_u^i \left[\sum_{j\in S}\left[p_{i,j,k}\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j\right]\right] \\
&= \sum_{i\in S}\sum_{j\in S} p_{i,j,k}\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}x_u^i\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j.
\end{aligned} \tag{15}$$

We show the bound of (15) by using the condition of the stable property. First, we show the upper bound of (15), $\frac{1}{d_v}\sum_{u\in N(v)} \bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u, \{\mathbf{x}_w\}_{w\in N(u)})$.

From the condition of the general voter model in this lemma, $\{\omega_{uv}\}_{u\in N(v)} \in K_v$ for all $v \in V$, i.e. $\sum_{u\in N(v)} {\omega_{uv}}^2 \le 4d_v$ for all $v \in V$. Then, by the stable condition Ineq. (5), we have

$$\frac{1}{d_u}\sum_{w\in N(u)} \omega_{wu}x_w^j \le \frac{1}{d_u}\sum_{w\in N(u)} \omega_{wu}y_w^j + \delta. \tag{16}$$

We know for all $x_w^j \le 1$ and all $u \in V$, $\sum_{w\in N(u)} \omega_{wu} = 1$. Then, we obtain

$$\frac{1}{d_u}\sum_{w\in N(u)} \omega_{wu}x_w^j \le 1. \tag{17}$$

With $\{\bar{\omega}_{uv}\}_{u\in N(v)} \in K_v$ and Ineq. (17),

$$\sum_{u\in N(v)}\left(\bar{\omega}_{uv}\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j\right)^2 \le 4d_v,$$

i.e., $\{\bar{\omega}_{uv}\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j\}_{u\in N(v)} \in K_v$. Hence, by Ineq. (5),

$$\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}x_u^i\left[\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j\right] \leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}y_u^i\left[\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}x_w^j\right]+\delta. \quad (18)$$

We apply Ineqs. (16) and (18) on the Eq. (15) as follows:

$$\frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}\bar{f}_u^k(\mathbf{x}_u,\{\mathbf{x}_w\}_{w\in N(u)}) \leq \sum_{i\in S}\sum_{j\in S}p_{i,j,k}\left[\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}y_u^i\frac{1}{d_u}\sum_{w\in N(u)}\omega_{wu}y_w^j\right] + \sum_{i\in S}\sum_{j\in S}p_{i,j,k}\left[\frac{1}{d_u}\sum_{w\in N(u)}\bar{\omega}_{uv}y_u^i+1\right]\delta. \quad (19)$$

Since for all $y_u^i \leq 1$ and $\sum_{u\in N(v)}\bar{\omega}_{uv} \leq 2d_v$ for all $v \in V$,

$$\sum_{i\in S}\sum_{j\in S}p_{i,j,k}\left[\frac{1}{d_u}\sum_{w\in N(u)}\bar{\omega}_{uv}y_u^i+1\right]\delta \leq \sum_{i\in S}\sum_{j\in S}3p_{i,j,k}\delta. \quad (20)$$

By the definition of $\bar{f}_u^k(\mathbf{y}_u,\{\mathbf{y}_w\}_{w\in N(u)})$ and δ, we have

$$\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u,\{\mathbf{x}_w\}_{w\in N(u)}) \leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{y}_u,\{\mathbf{y}_w\}_{w\in N(u)})+\epsilon. \quad (21)$$

Using a similar method as above, the lower bound is given as

$$\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u,\{\mathbf{x}_w\}_{w\in N(u)}) \geq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{y}_u,\{\mathbf{y}_w\}_{w\in N(u)})-\epsilon. \quad (22)$$

Then, we obtain the boundary equation

$$\left|\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{x}_u,\{\mathbf{x}_w\}_{w\in N(u)}) - \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_u^k(\mathbf{y}_u,\{\mathbf{y}_w\}_{w\in N(u)})\right| \leq \epsilon. \quad (23)$$

This shows that Lemma 6 follows.

■

2. SIR Model

Next, we provide a proof that shows the SIR model is stable. Since the SIS model and the multistate SIS models are variations of the SIR model, a similar proof of the stability can be applied to those models.

Lemma 7 *If $f_v^k(\cdot)$ is given as Eq. (4) and $d_v\max(\beta_{uv}, u \in N(v)) \leq 2$, $\beta_{uv} < 0.98$ are satisfied for all $v \in V$, then $\{\bar{f}_v^k(\cdot)\}_{v\in V}$ is stable.*

Proof. We use the same notations as in Definition 1. If Ineq. (6) is derived from a given condition Ineq. (5), Lemma 7 is proved. We prove this model has a stable property by using Lemma 8.

Let us define

$$\begin{aligned}
&\beta_{uv,max} = \max(\beta_{uv}, u \in N(v)),\\
&\kappa_v = |\log(1-\beta_{uv,max})|/2,\\
&\mu_{uv} = \log(1-\beta_{uv}y_u^1(t)) - \log(1-\beta_{uv})^{y_u^1(t)},\\
&\mu_v = \frac{1}{d_v}\sum_{u\in N(v)}\mu_{uv}.
\end{aligned}$$

Lemma 8 *With the condition of stable property (5), we have for all $v \in V$,*

$$\left|\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^1) - \frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}y_u^1)\right| \leq \mu_v+\kappa_v\delta. \quad (24)$$

We prove Lemma 8 by showing the bound of

$$\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^i). \quad (25)$$

Let us find the lower bound of (25) first. It is easy to show

$$\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^i) \geq \frac{1}{d_v}\sum_{u\in N(v)}x_u^i\log(1-\beta_{uv}) \quad (26)$$

by using $0 \leq \beta_{uv}, x_u^i \leq 1$. The right-hand side of this inequality can be rewritten as

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}x_u^1\log(1-\beta_{uv})\\
&= \frac{-|\log(1-\beta_{uv,max})|}{2d_v}\sum_{u\in N(v)}x_u^1\frac{2\log(1-\beta_{uv})}{-|\log(1-\beta_{uv,max})|}\\
&= \frac{-\kappa_v}{d_v}\sum_{u\in N(v)}x_u^1\frac{\log(1-\beta_{uv})}{-\kappa_v}.
\end{aligned} \quad (27)$$

Because $\frac{2\log(1-\beta_{uv})}{-|\log(1-\beta_{uv,max})|} \leq 2$, $\{\frac{2\log(1-\beta_{uv})}{-|\log(1-\beta_{uv,max})|}\}_{u\in N(v)}$ is in K_v. By the stable condition, we have

$$\begin{aligned}
&\frac{-\kappa_v}{d_v}\sum_{u\in N(v)} x_u^i \frac{\log(1-\beta_{uv})}{-\kappa_v} \\
&\quad\geq -\kappa_v\left(\frac{1}{d_v}\sum_{u\in N(v)} y_u^i\frac{\log(1-\beta_{uv})}{-\kappa_v}+\delta\right) \\
&\quad= \frac{1}{d_v}\sum_{u\in N(v)} y_u^1\log(1-\beta_{uv})-\kappa_v\delta \qquad (28)\\
&\quad= \left(\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv})^{y_u^1}\right)-\kappa_v\delta.
\end{aligned}$$

To prove this lemma, we want to estimate $\log(1-\beta_{uv})^{y_u^i}$ to $\log(1-\beta_{uv}y_u^i)$, so we use the μ_{uv}. The value of μ_{uv} goes to 0 when β_{uv} goes to 0 or $y_u^i(t)$ goes to 0 or 1. For example, for any $y_u^i \in [0,1]$, if $\beta_{uv}=0.5$, then $\mu_{uv} < 0.06$, and if $\beta_{uv}=0.1$, then $\mu_{uv}<0.0015$. Therefore,

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv})^{y_u^1}-\kappa_v\delta \\
&\quad=\left(\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}y_u^1)\right)-\mu_v-\kappa_v\delta. \qquad (29)
\end{aligned}$$

Hence, we obtain the lower bound

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^1) \\
&\qquad\geq \frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}y_u^1)-\mu_v-\kappa_v\delta. \qquad (30)
\end{aligned}$$

We obtain the upper bound by reversing the method derived from the lower bound. The upper bound of (25) is given as

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^1) \\
&\quad\leq \frac{1}{d_v}\sum_{u\in N(v)} x_u^1\log(1-\beta_{uv})+\mu_v \\
&\quad\leq \frac{1}{d_v}\sum_{u\in N(v)} y_u^1(t)\log(1-\beta_{uv})+\mu_v+\kappa_v\delta \qquad (31)\\
&\quad\leq \frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}y_u^1)+\mu_v+\kappa_v\delta.
\end{aligned}$$

Thus, we obtain the following bound inequality:

$$\begin{aligned}
&\left|\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}x_u^1)-\frac{1}{d_v}\sum_{u\in N(v)}\log(1-\beta_{uv}y_u^1)\right| \\
&\qquad\leq \mu_v+\kappa_v\delta. \qquad (32)
\end{aligned}$$

■

Next, we prove that (4) has a stable property by Lemma 8. Let us define $\bar{\epsilon}_v = \exp\left(d_v(\mu_v+\kappa_v\delta)\right)-1$. Then,

$$\left|\frac{\prod_{u\in N(v)}(1-x_u^1\beta_{uv})}{\prod_{u\in N(v)}(1-y_u^1\beta_{uv})}-1\right|\leq\bar{\epsilon}_v. \qquad (33)$$

In the SIR model, the probability function f_v^k is different for each state k. Hence, we need to show the stable property for each state. Let us start with the case of $k=0$. With $0\leq\beta_{wu},x_w^1\leq 1$ for all β_{wu},x_w^1, $\prod_{w\in N(u)}(1-x_w^1\beta_{wu})$ is equal to or less than 1. From the definition of stable property, $\{\bar{\omega}_{uv}\}_{u\in N(v)}\in K_v$. Using methods similar to those we applied from Ineq. (17) to Ineq. (18), we have $\{\bar{\omega}_{uv}\prod_{w\in N(u)}(1-x_w^1\beta_{wu})\}_{u\in N(v)}\in K_v$ and

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}x_u^0\prod_{w\in N(u)}(1-x_w^1\beta_{wu}) \\
&\qquad\leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}y_u^0\prod_{w\in N(u)}(1-x_w^1\beta_{wu})+\delta. \qquad (34)
\end{aligned}$$

By Ineqs. (33) and (34), we have

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^0(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)}) \\
&\leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}y_u^0(1+\bar{\epsilon}_u)\prod_{w\in N(u)}(1-y_w^1\beta_{wu})+\delta \\
&= \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}y_u^0\prod_{w\in N(u)}(1-y_w^1\beta_{wu}) \\
&\quad+\delta+\frac{1}{d_v}\sum_{u\in N(v)}\bar{\epsilon}_u\left(\bar{\omega}_{uv}y_u^0\prod_{w\in N(u)}(1-y_w^1\beta_{wu})\right). \qquad (35)
\end{aligned}$$

Similar to Ineq. (21), $\bar{\epsilon}_u(\cdot)$ is equal to or less than $2\bar{\epsilon}_u$. Let us define $\bar{\epsilon}=\max(\frac{1}{d_v}\sum_{u\in N(v)}\bar{\epsilon}_u, v\in V)$ and $\epsilon_0=\delta+2\bar{\epsilon}$. Then, we have

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^0(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)}) \\
&\qquad\leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^k(\mathbf{y}_v,\{\mathbf{y}_u\}_{u\in N(v)})+\epsilon_0. \qquad (36)
\end{aligned}$$

The upper bound of the case of $k=1$ can also be derived easily by Ineq. (33), Ineq. (34) and $\gamma\leq 1$. Define $\epsilon_1=3\delta+2\bar{\epsilon}$. Then,

$$\begin{aligned}
&\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^1(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)})= \\
&\qquad\leq \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^1(\mathbf{y}_v,\{\mathbf{y}_u\}_{u\in N(v)})+\epsilon_1. \qquad (37)
\end{aligned}$$

Define $\epsilon_2 = (1+\gamma)\delta$. For the case of $k=2$, the upper bound is given as:

$$\begin{aligned}\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^2(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)}) &\le \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\left(y_u^1\gamma+y_u^2\right)+\delta\left(\gamma+1\right)\\ &= \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^2(\mathbf{y}_v,\{\mathbf{y}_u\}_{u\in N(v)})+\epsilon_2.\end{aligned} \tag{38}$$

Using a method similar to the above, the lower bound is given as

$$\begin{aligned}\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^k(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)}) &\ge \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^k(\mathbf{y}_v,\{\mathbf{y}_u\}_{u\in N(v)})-\epsilon_k,\end{aligned} \tag{39}$$

for all states k in S. Let $\epsilon=\max(\epsilon_k, k\in S)=\epsilon_1$. We finally obtain

$$\left|\frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^k(\mathbf{x}_v,\{\mathbf{x}_u\}_{u\in N(v)}) - \frac{1}{d_v}\sum_{u\in N(v)}\bar{\omega}_{uv}\bar{f}_v^k(\mathbf{y}_v,\{\mathbf{y}_u\}_{u\in N(v)})\right|\le\epsilon. \tag{40}$$

From the definition of $\bar{\epsilon_v}$,

$$\begin{aligned}\bar{\epsilon} &= \max\left(\frac{1}{d_v}\sum_{u\in N(v)}\left(\exp\left(d_u\mu_u\right)\exp\left(d_u\kappa_u\delta\right)-1\right), v\in V\right)\\ &\le \max\left(\exp\left(d_u\mu_u\right)\exp\left(d_u\kappa_u\delta\right)-1\right).\end{aligned}$$

From the condition of Lemma 8, $d_v\cdot\beta_{uv,max}\le 1$ for all $v\in V$. Then, $\exp(d_u\mu_u)$ is close to 1 and $\bar{\epsilon}\simeq\exp(d_u\kappa_u\delta)-1$. Because $\max(d_u\kappa_u)<4$ where $\beta<0.98$ and $\delta<1$, $\exp(d_u\kappa_u\delta)=1+d_u\kappa_u\delta+O((d_u\kappa_u\delta)^2)\le 1+16d_u\kappa_u\delta$. Then, from the definition of ϵ, for some constant C, $\epsilon=3\delta+2\bar{\epsilon}\le C\delta$ where $d_v\cdot\beta_{uv,max}\le 1$ and $\beta<0.98$. Hence, the SIR model also possesses the stable property.

■

Appendix B: Proof of Lemma 3

If $I_v^i(t)$ and $a_v^i(t)$ are given values, then with the notation of Definition 1, we can map $I_v^i(t)$ to x_v^i, $a_v^i(t)$ to y_v^i and ϵ_t to δ. In addition, $\{\omega_{uv}\}_{u\in N(v)}\in K_v$ for all $v\in V$. Let κ_{t+1} be ϵ in the notation of Definition 1. Then, the condition of Lemma 3 can be applied to the condition for the stability property given in (5). Hence, by the stability property, for any deterministic values $I_v^i(t)$ and $a_v^i(t)$ and any $\{\omega_{uv}\}_{u\in N(v)}\in K_v$, we have

$$\begin{aligned}&\Big|\frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}\bar{f}_v^k(\{I_v^i(t)\}_{i\in S},\{I_u^i(t)\}_{i\in S,u\in N(v)})\\ &-\frac{1}{d_v}\sum_{u\in N(v)}\omega_{uv}\bar{f}_v^k(\{a_v^i(t)\}_{i\in S},\{a_u^i(t)\}_{i\in S,u\in N(v)})\Big|\le\kappa_{t+1}.\end{aligned} \tag{41}$$

From the definitions of $f_v^k(\cdot)$ and $\bar{f}_v^k(\cdot)$, we have

$$\mathbb{E}[I_v^k(t+1)]=\bar{f}_v^k(\{I_v^i(t)\}_{i\in S},\{I_u^i(t)\}_{i\in S,u\in N(v)}).$$

Moreover, $\bar{f}_v^k(\{a_v^i(t)\}_{i\in S},\{a_u^i(t)\}_{i\in S,u\in N(v)})$ is defined as $a_v^k(t+1)$ in (7). Hence, we have

$$\begin{aligned}&\left|\mathbb{E}[r_v^k(t+1)]-b_v^k(t+1)\right|\\ &=\left|\frac{1}{d_v}\sum\omega_{uv}\mathbb{E}[I_u^k(t+1)]-\frac{1}{d_v}\sum\omega_{uv}a_u^k(t+1)\right|\le\kappa_{t+1},\end{aligned} \tag{42}$$

which gives the bound on the difference between the approximated and expected state densities of v's neighbors, where for all $v\in V$ and all $i\in S$, $I_v^i(t)$ is given and $\left|r_v^i(t)-b_v^i(t)\right|\le\epsilon_t$.

Since the $s_v(t)$ for all $v\in V$ are given, i.e., $I_v^i(t)$ is given for all $v\in V$ and all $i\in S$, the $s_v(t+1)$ are mutually independent. Hence, from Hoeffding's inequality, for any $v\in V$ and any $i\in S$, we have

$$\begin{aligned}&Pr\left[\left|r_v^i(t+1)-\mathbb{E}[r_v^i(t+1)]\right|\ge\epsilon_t|\{s_v(t)\}_{v\in V}\right]\\ &\le 2\exp\left(-\frac{2{\epsilon_t}^2}{\sum_{u\in N(v)}\left(\omega_{uv}/d_v\right)^2}\right)\le 2\exp\left(-\frac{2{\epsilon_t}^2 d_{v,min}}{4}\right).\end{aligned} \tag{43}$$

We know the approximated probability $a_v^i(t+1)$, which is a deterministic variable. Then, by Ineq. (42), the

$$\left|\frac{1}{d_v}\sum Pr[I_u^k(t+1)=1]-\frac{1}{d_v}\sum a_u^k(t+1)\right|$$

are small, where the $s_v(t)$ for all $v\in V$ are given.

Ineq. (43) is valid for all given $I_v^i(t)$; therefore, Ineq. (43) is also valid for all given $I_v^i(t)$ that satisfy $\left|r_v^i(t)-b_v^i(t)\right|\le\epsilon_t, \forall v\in V$. If the probability of $r_v^i(t+1)$ lying outside the bound is smaller than some specific value for all cases of $s_v^i(t)$, then we can say that the probability of $r_v^i(t+1)$ lying outside the bound is smaller than that specific value without knowing $s_v^i(t)$. Therefore, by

applying Ineq. (42) to Ineq. (43), we obtain

$$\begin{aligned}
&Pr\Big[\big|r_v^i(t+1)-b_v^i(t+1)\big|\geq \epsilon_t+\kappa_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|\leq\epsilon_t\Big]\\
&\leq Pr\Big[\big|r_v^i(t+1)-\mathbb{E}[r_v^i(t+1)]\big|\geq\epsilon_t\\
&\qquad |\{s_v(t)\}_{v\in V},\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|\leq\epsilon_t\Big]\\
&\qquad\leq 2\exp\left(-\frac{2{\epsilon_t}^2 d_{v,min}}{4}\right). \quad (44)
\end{aligned}$$

Let $\epsilon_{t+1}=\epsilon_t+\kappa_{t+1}$. By the union bound, we have

$$\begin{aligned}
&Pr\Big[\forall v\in V,\forall i\in S,\big|r_v^i(t+1)-b_v^i(t+1)\big|\leq\epsilon_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|\leq\epsilon_t\Big]\\
&\qquad\geq 1-2sn\exp\left(-\frac{2{\epsilon_t}^2 d_{v,min}}{4}\right). \quad (45)
\end{aligned}$$

■

Appendix C: Proof of Lemma 4

Lemma 2 gives the proof at $t=0$. Since $\mathbb{E}[r_v^i(0)]=b_v^i(0)$, if $\big|r_v^i(0)-b_v^i(0)\big|\leq\epsilon_0$ is satisfied, then $\big|r_v^i(0)-\mathbb{E}[r_v^i(0)]\big|\leq\epsilon_0$ is also satisfied. Thus,

$$\begin{aligned}
&Pr\Big[\forall v\in V,\forall i\in S,\big|r_v^i(0)-\mathbb{E}[r_v^i(0)]\big|\leq\epsilon_0\Big]\\
&\qquad\geq 1-2sn\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right). \quad (46)
\end{aligned}$$

To prove that this inequality holds for $t=1,\ldots,T$, let us start with Lemma 3. By Lemma 3, $r_v^i(t+1)$ approximates $b_v^i(t+1)$ with high probability if the differential between the densities $r_u^i(t)$ and $b_u^i(t)$ at the previous time is within a small error bound. By applying the inductive approach to Lemma 3, we find that if $\big|r_v^i(0)-b_v^i(0)\big|\leq\epsilon_0$ is satisfied, then we have

$$\begin{aligned}
&Pr\Big[1\leq t\leq T,\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|\leq\epsilon_t\Big]\\
&\qquad\geq\prod_{t=1}^{T}\left(1-2sn\exp(-\frac{2{\epsilon_{t-1}}^2 d_{v,min}}{4})\right). \quad (47)
\end{aligned}$$

For any $\{\omega_{uv}\}_{u\in N(v)}\in K_v$, Ineq. (12) holds. This statement implies that Ineq. (12) holds when $\omega_{uv}=1$ for all $v\in V$ and all $u\in N(v)$. In other words,

$$\begin{aligned}
&Pr\Big[\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t+1)-\bar{b_v^i}(t+1)\big|\leq\epsilon_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|<\epsilon_t\Big]\\
&=Pr\Big[\forall v\in V,\forall i\in S,|r_v^i(t+1)-b_v^i(t+1)|\leq\epsilon_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|r_v^i(t)-b_v^i(t)\big|<\epsilon_t\Big]. \quad (48)
\end{aligned}$$

Under the condition that $\big|r_v^i(t)-b_v^i(t)\big|\leq\epsilon_t$ for all $v\in V$ and all $i\in S$, $\mathbb{E}[\bar{r_v^i}(t+1)]$ without a given $s_v(t)$ lies between the maximum and minimum cases of $\mathbb{E}[\bar{r_v^i}(t+1)]$, where $s_v(t)$ is given for all $v\in V$, i.e., is bounded within $\bar{b_v^i}(t+1)\pm\kappa_{t+1}$ according to Ineq. (42). Hence,

$$\begin{aligned}
&Pr\big[\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t+1)-\mathbb{E}[\bar{r_v^i}(t+1)]\big|\leq\epsilon_{t+1}+\kappa_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t)-\bar{b_v^i}(t)\big|\leq\epsilon_t\big]\\
&\geq Pr\big[\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t+1)-\bar{b_v^i}(t+1)\big|\leq\epsilon_{t+1}\\
&\qquad |\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t)-\bar{b_v^i}(t)\big|\leq\epsilon_t\big] \quad (49)
\end{aligned}$$

follows. By Ineqs. (47) and (49) and Eq. (48), if the condition $\big|r_v^i(0)-b_v^i(0)\big|\leq\epsilon_0$ is satisfied for all $v\in V$ and all $i\in S$, we have

$$\begin{aligned}
&Pr\big[1\leq t\leq T,\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t)-\mathbb{E}[\bar{r_v^i}(t)]\big|\leq\epsilon_t+\kappa_t\big]\\
&\qquad\geq\prod_{t=1}^{T}\left(1-2sn\exp\left(-\frac{2{\epsilon_{t-1}}^2 d_{v,min}}{4}\right)\right). \quad (50)
\end{aligned}$$

Let us define $\epsilon=\max(\epsilon_t+\kappa_t)=\epsilon_T+\kappa_T$. The minimum of all ϵ_t is ϵ_0. Combining Ineqs. (46) and (50) yields

$$\begin{aligned}
&Pr\big[0\leq t\leq T-1,\forall v\in V,\forall i\in S,\big|\bar{r_v^i}(t)-\mathbb{E}[\bar{r_v^i}(t)]\big|\leq\epsilon\big]\\
&\qquad\geq\prod_{t=0}^{T-1}\left(1-2sn\exp\left(-\frac{2{\epsilon_0}^2 d_{v,min}}{4}\right)\right)\\
&\qquad\geq 1-2Tsn\exp\left(\frac{-2{\epsilon_0}^2 d_{v,min}}{4}\right) \quad (51)
\end{aligned}$$

and

$$\begin{aligned}
&Pr\big[0\leq t\leq T,\forall v\in V,\forall i\in S,\big|\bar{a_v^i}(t)-\mathbb{E}[\bar{r_v^i}(t)]\big|\leq\epsilon\big]\\
&\qquad\geq 1-2(T)sn\exp\left(\frac{-2{\epsilon_0}^2 d_{v,min}}{4}\right). \quad (52)
\end{aligned}$$

If $d_{v,min}=\tau\log n$ for some constant $\tau>0$, then we have

$$\begin{aligned}
&Pr\big[0\leq t\leq T,\forall v\in V,\forall i\in S,\big|\bar{a_v^i}(t)-\mathbb{E}[\bar{r_v^i}(t)]\big|\leq\epsilon\big]\\
&\qquad\geq 1-2Tsn^{(1-2{\epsilon_0}^2\tau/4)}. \quad (53)
\end{aligned}$$

Based on the condition of Theorem 1, $d_{v,min}\geq\tau\log n$ for all $\tau>0$ as n goes to ∞. Let $\delta=-1+2{\epsilon_0}^2\tau/4$. Then, δ goes to ∞ as τ goes to ∞. Thus, we have

$$\begin{aligned}
&Pr\big[0\leq t\leq T,\forall v\in V,\forall i\in S,\big|\bar{a_v^i}(t)-\mathbb{E}[\bar{r_v^i}(t)]\big|\leq\epsilon\big]\\
&\qquad=1-o(n^{-\delta}). \quad (54)
\end{aligned}$$

■

Appendix D: Additional Experimental Results

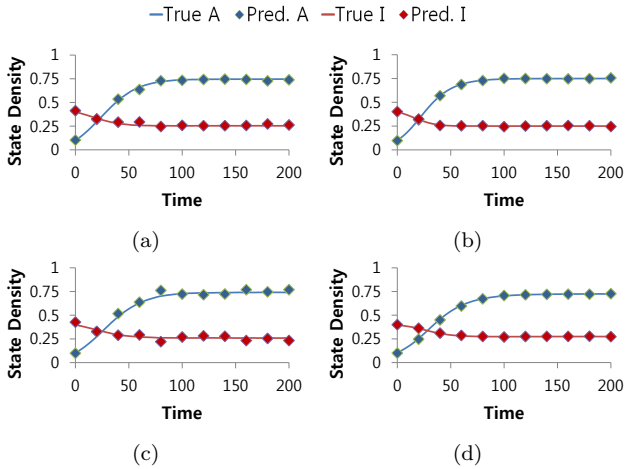

FIG. 10: Results for the DAU model with the *self-sustainable* parametric value set P_1 and the initial state distribution $D_1 = (0.5, 0.4, 0.1)$. The set W contains all nodes with nonzero (in)degrees. (a) $WS_{1000,10}$, (b) $WS_{10000,100}$, (c) $BA1000$, (d) *Slashdot*.

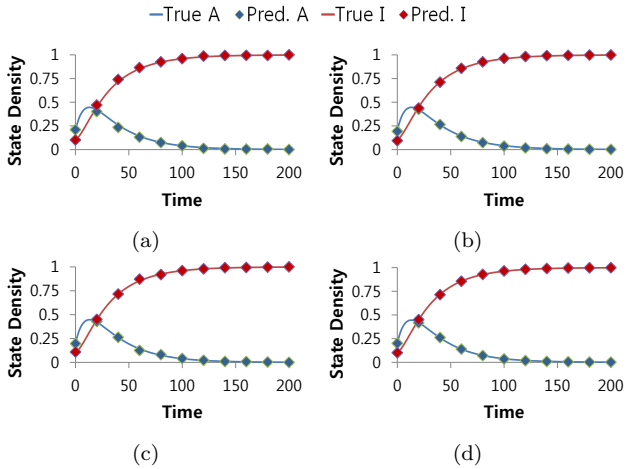

FIG. 11: Results for the DAU model with the *unsustainable* parametric value set P_2 and the initial state distribution $D_2 = (0.7, 0.1, 0.2)$. The set W contains all nodes with nonzero (in)degrees. (a) $WS_{1000,10}$, (b) $WS_{10000,100}$, (c) $BA1000$, (d) *Slashdot*.

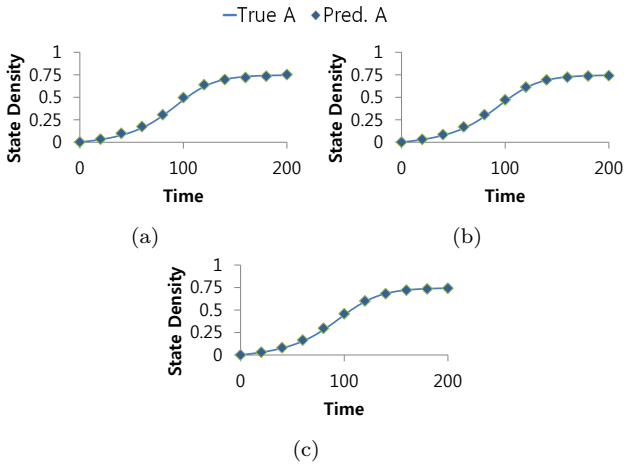

FIG. 12: Results for the DAU model with the *unsustainable* parametric value set P_1 and *gowalla* dataset. Initially, the nodes with the top 50 highest (in)degree are selected as active nodes, while the other nodes are in the "nonmember" state. The set W contains the (a) top 1%, (b) 5% or (c) 10% of the nodes as ranked by (in)degree.

[1] M. E. Yildiz, R. Pagliari, A. Ozdaglar, and A. Scaglione, in *Proc. 2010 Inf. Theory and Applications Workshop (ITA)* (2010) pp. 1–7.

[2] A. S. Peter Clifford, Biometrika **60**, 581 (1973).

[3] C. M. Schneider-Mizell and L. M. Sander, J. Statist. Phys. **136**, 59 (2009).

[4] P. Moretti, A. Baronchelli, M. Starnini, and R. Pastor-Satorras, Dynamics On and Of Complex Netw. , 285 (2013).

[5] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, Phys. Rev. E **84**, 011130 (2011).

[6] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).

[7] D. Kempe, J. Kleinberg, and E. Tardos, in *Proc. 9th ACM SIGKDD int. conf. on Knowledge discovery and data mining* (2003) pp. 137–146.

[8] Y. Wang, X. Ye, H. Zha, and L. Song, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA* (2017) pp. 1644–1654.

[9] Y. Wang, E. Theodorou, A. Verma, and L. Song, in *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain* (2018) pp. 1077–1086.

[10] M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie, in *Proceedings of the 2018 World Wide Web Conference*, WWW '18 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2018) pp. 419–428.

[11] M. Granovetter, Am. J. Soc. **83**, 1420 (1978).

[12] T. W. Valente, Soc. Netw. **18**, 69 (1996).

[13] L. Blume, D. Easley, J. Kleinberg, R. Kleinberg, and E. Tardos, in *Proc. 52nd IEEE Symp. Foundations* (2011) pp. 393–402.

[14] D. J. Watts, in *Proc. Nat. Acad. Sci. United States Am.*, Vol. 99 (2002) pp. 5766–5771.

[15] D. E. Whitney, Phys. Rev. E **82**, 066110 (2010).

[16] P. S. Dodds and D. J. Watts, Phys. Rev. Letters **92**, 218701 (2004).

[17] M. Lelarge and J. Bolot, in *Proc. 2008 ACM SIGMETRICS Int. Conf. on Measurement and Modeling of Computer Systems*, SIGMETRICS '08 (2008) pp. 37–48.

[18] M. E. J. Newman, Phys. Rev. Lett. **103**, 058701 (2009).

[19] F. D. Sahneh, C. Scoglio, and P. Van Mieghem, IEEE/ACM Trans. Netw. **21**, 1609 (2013).

[20] Mata, Anglica S. and Ferreira, Silvio C., EPL **103**, 48003 (2013).

[21] C. Kamp, M. Moslonka-Lefebvre, and S. Alizon, PLOS Computational Biology **9**, 1 (2013).

[22] Q. Wu, S. Chen, and L. Zha, Chaos, Solitons and Fractals **96**, 17 (2017).

[23] M. Boguñá, C. Castellano, and R. Pastor-Satorras, Phys. Rev. Lett. **111**, 068701 (2013).

[24] C.-R. Cai, Z.-X. Wu, M. Z. Q. Chen, P. Holme, and J.-Y. Guan, Phys. Rev. Lett. **116**, 258301 (2016).

[25] B. Ribeiro, in *Proc. 23rd Int. Conf. World Wide Web*, WWW '14 (2014) pp. 653–664.

[26] S. M. Krause, P. Bottcher, and S. Bornholdt, Phys. Rev. E **85**, 031126 (2012).

[27] Y. Lin, J. C. S. Lui, K. Jung, and S. Lim, J. Complex Netw. **2**, 431 (2014).

[28] P. L. Krapivsky and S. Redner, Phys. Rev. Lett. **90**, 238701 (2003).

[29] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," `http://snap.stanford.edu/data` (2014).

[30] Y. Li, J. Fan, Y. Wang, and K. Tan, IEEE Transactions on Knowledge and Data Engineering **30**, 1852 (2018).

[31] H. T. Nguyen, M. T. Thai, and T. N. Dinh, IEEE/ACM Transactions on Networking **25**, 2419 (2017).