Providing Probabilistic Guarantees on the Time of Information Spread in Opportunistic Networks

Yoora Kim[†], Kyunghan Lee[‡], Ness B. Shroff[†], and Injong Rhee[♯]
[†]{kimy, shroff}@ece.osu.edu, [‡]khlee@unist.ac.kr, [♯]rhee@ncsu.edu

***Abstract*—A variety of mathematical tools have been developed for predicting the spreading patterns in a number of varied environments including infectious diseases, computer viruses, and urgent messages broadcast to mobile agents (e.g., humans, vehicles, and mobile devices). These tools have mainly focused on estimating the average time for the spread to reach a fraction (e.g., α) of the agents, i.e., the so-called average completion time $E(T_\alpha)$. We claim that providing probabilistic guarantee on the time for the spread T_α rather than only its average gives a much better understanding of the spread, and hence could be used to design improved methods to prevent epidemics or devise accelerated methods for distributing data. To demonstrate the benefits, we introduce a new metric $G_{\alpha,\beta}$ that denotes the time required to guarantee α completion with probability β, and develop a new framework to characterize the distribution of T_α for various spread parameters such as number of seeds, level of contact rates, and heterogeneity in contact rates. We apply our technique to an experimental mobility trace of taxies in Shanghai and show that our framework enables us to allocate resources (i.e., to control spread parameters) for acceleration of spread in a far more efficient way than the state-of-the-art.**

I. Introduction

Spreading patterns of pandemics [1], computer viruses [2], and information [3], [4] have been widely studied in various research disciplines including epidemics, biology, physics, sociology, and computer networks. In these disciplines, most studies have been devoted to characterizing spread behaviors toward a network of mobile agents including humans, vehicles, and mobile devices[1] over time. These studies can be classified into two groups based on their objectives. Interestingly, both these objects lie in opposite directions: slowing down or acceleration of spread. For the research that deals with biological and electronic viruses, how to slow down the spread has been the most important question to be answered. On the other hand, another set of research work for computer data and information distribution has pursued designing engineering methods to accelerate the spread.

Whatever the goals are, existing studies have relied on common mathematical frameworks such as the branching process, mean-field approximation, and stochastic differential equations [5]. Due to the characteristics of these frameworks, the spread of virus or information has generally been analyzed in terms of its average behavior under various epidemic models summarized in [6], where epidemic models define whether agents are recoverable[2] or not and whether they become immune after recovery or are still susceptible to infection. Here, average behavior typically indicates $\mathrm{E}[N_t]$ where N_t denotes the number of infected nodes in the network at time t.

Average analysis gives an answer to a question on how many nodes are infected (or informed) *on average* under a specific epidemic model after a time duration t from the emergence of a virus (or generation of information). There have been many extensions to this analysis through aforementioned frameworks. The authors of [7] identified how much a network topology affects the speed of virus spreading and the authors of [8] derived a closed form equation of the critical level of virus infection rate allowing a virus to persist in a network when the virus is recoverable with a certain rate. More realistic average spread behaviors of a virus with the heterogeneity inherent in human mobility patterns have been studied through simulations in [9]. In computer networks, [10] analyzed the average propagation behavior of code red worm in the Internet using measurement data from ISP and an epidemic model. [3] applied understanding on the average behavior of virus spread to information propagation in delay tolerant networks. Similarly, [2] analyzed the average spread behaviors of self-propagating worms on the Internet using branching process.

While there has been a plethora of work on average analysis, the problem of allocating optimal amounts of resource to a network of a set of nodes for slowing down or accelerating spread has been under-explored.[3] Specifically, higher order spread behaviors over time rather than average behaviors to design optimal resource allocation have not been well understood. The right question should be what will be the distribution of the number of infected nodes at time t, which is equivalent to what will be the temporal distribution of the event that n nodes are infected. Characterizing the temporal distribution of spread allows one to guarantee the time for spread with high probability and it leads to control knobs

This work has been supported in part by the National Science Foundation CNS-1065136, CNS-0910868, and CNS-1016216; the Army Research Office MURI grants W911NF-08-1-0238 and W911NF-12-1-0385; and the year of 2012 Research Fund of the UNIST (Ulsan National Institute of Science and Technology).

Y. Kim is with the Department of Electrical and Computer Engineering, The Ohio State University, USA. K. Lee is with the School of Electrical and Computer Engineering, UNIST, Korea. N. B. Shroff is with the Department of Electrical and Computer Engineering and the Department of Computer Science and Engineering, The Ohio State University, USA. I. Rhee is with the Department of Computer Science, North Carolina State University, USA.

[1]We will interchangeably use agents and nodes unless confusion arises.

[2]A virus that cannot be recovered can be considered to be identical to undeletable or unforgettable information.

[3] [11] studied the optimal allocation of wireless channels of a carrier to mobile nodes in a content delivery network, which maximizes the sum utility defined with the content delivery time to the nodes. In the work, a bound on the content delivery time was studied, but its exact distribution was left unsolved.

for allocating resources to a network with its own purpose of spread. However, understanding the temporal distribution involves non-trivial challenges since there is a huge dimension of diversity in contact events among nodes in a network.

In order to address the challenges involved in obtaining deeper understanding of resource allocation, in this paper, we propose a new analytical framework based on CTMC (continuous time Markov chain), which allows us to fully characterize the temporal aspect of spread behaviors. For simplicity, we put our emphasis on information distribution among intermittently meeting mobile nodes forming an opportunistic network, i.e., a mobile social network, but our results are easily applicable to general spread of epidemics. Our framework is capable of answering many intriguing engineering questions such as "what is the distribution of time for a network to have 75% penetration rate?" and "If 75% penetration is aimed, when is the time to guarantee that level of penetration with 99% of confidence?". It can also answer a more fundamental question involving heterogeneity of nodes in a network, "Does heterogeneity help or hurt spreading?" We show the efficacy of our solution in answering these questions with the use of one of the largest experimental GPS (global positioning system) trace of taxies in Shanghai. Our simulation studies on the trace provide added verification that our framework is robust and enables us to engineer the network in a far more efficient way than existing understandings of spread.

The rest of the paper is organized as follows. In Section II, we provide our system model along with definitions of relevant metrics. In Section III, we develop our analytical framework and present major analytical results. Based on our framework, we characterize the temporal distribution of spread behavior and provide their applications in Section IV. We present simulation studies using Shanghai taxi trace and conclude our paper in Sections V and VI, respectively.

II. Model Description

A. Overview of Epidemic Models

In classic epidemiology, an individual (i.e., node) is classified into either susceptible, infected, or removed (or immune) according to its status for a disease [5]. A susceptible individual refers to the one who is not infected yet, but is prone to be infected. An infected individual refers to the one who already got the disease and is capable of spreading it to susceptible individuals. A removed individual indicates the one who was previously infected but became immune to the disease. These three classifications are conventionally denoted by S, I, and R, respectively, and induce SIS, SIR, and SI epidemic models and their variants. In this paper, we focus on the SI model in which once a susceptible individual is infected, it stays infected for the remainder of the epidemic process. The SI model fits particularly well with information spread in opportunistic networks since once a data is delivered to an individual, it is considered that the data is delivered to its upper layer and it is no longer required (i.e., permanently infected).

B. Our System Model

We consider a network (or a population) consisting of N mobile nodes. We assume that all nodes in the network can be classified into K different types according to their mobility patterns and epidemic attributes. We denote the collection of the kth type of nodes as group k $(k = 1, \ldots, K)$. Let N_k be the number of nodes in group k and denote $\boldsymbol{N} \triangleq (N_k)_{1 \le k \le K}$. Then, we have $|\boldsymbol{N}| \triangleq \sum_k N_k = N$ (throughout this paper, we use a bold font symbol for an arbitrary vector or a matrix notation. In addition, for a vector $\boldsymbol{V} = (V_k)$, we define the operation $|\boldsymbol{V}|$ as $|\boldsymbol{V}| \triangleq \sum_k V_k$).

In our model, the mechanism of information (or a packet or a virus) spread is as follows: initially, the information is delivered to some selected nodes, which we call *seeds*.[4] Whenever a seed, say node a, meets a susceptible node not having the information yet, it spreads the information to the susceptible node with probability $\varphi_a \in (0, 1]$. Then, the susceptible node, say node b, receives the information successfully with probability $\psi_b \in (0, 1]$ and becomes infected (or informed). Once the susceptible node becomes infected, it stays infected for the remainder of the spreading process, and is involved in disseminating the information in a similar manner as the seed. The spreading process ends when all nodes in the network obtain the information. In our spreading model, the probabilities φ_a and ψ_b can be interpreted as the infectivity and the susceptibility of nodes a and b, respectively. For instance, in the case of rumor propagation, φ_a quantifies the tendency of person a to gossip, while ψ_b quantifies the receptive nature of a listener b to the rumor. For the case of packet dissemination in an opportunistic network, φ_a represents the probability that node a schedules to transmit a packet, and ψ_b represents the probability of successful packet reception at node b, which depends on, e.g., the contact period, number of contending nodes, and channel conditions.

The stochastic characteristic of a pairwise meeting process is a critical factor that determines the temporal behavior of the spreading process. In the literature, it has been recently shown that the time duration between two consecutive contacts of a pair of nodes, called *pairwise inter-contact time*, can be modeled by an exponential random variable, e.g., [12]–[14]. In [12], exponential inter-contact patterns are validated experimentally using three different mobility data sets. When nodes follow Lévy flight mobility, which is known to closely mimic human mobility patterns [15], the authors in [14] mathematically proved that the inter-contact time distribution is bounded by an exponential distribution. Thus, in this paper we assume that the pairwise inter-contact time between nodes a and b, denoted by $M_{a,b}$, follows an exponential distribution with rate $\lambda_{a,b}$ (> 0), i.e.,

$$\mathrm{P}\{M_{a,b} > t\} = \exp(-\lambda_{a,b} t), \quad t \ge 0. \tag{1}$$

Suppose that node a is infected and node b is susceptible. From the meeting process between them, we can obtain the infection time $M_{a,b}^{\text{eff}}$ by taking the infectivity φ_a and the susceptibility ψ_b

[4]Note that being selected as seeds can be of willing or unwilling. For instance, a seed of a virus gets the virus unwillingly.

into account. From (1), we have:

$$\mathrm{P}\{M^{\mathrm{eff}}_{a,b} > t\} = \exp(-\lambda_{a,b}\varphi_a\psi_b t), \quad t \geq 0. \tag{2}$$

That is, the infection rate $\lambda^{\mathrm{eff}}_{a,b}$ between an infected node a and a susceptible node b becomes $\lambda^{\mathrm{eff}}_{a,b} = \lambda_{a,b}\varphi_a\psi_b$. The detailed proof of (2) is given in Appendix A of our technical report [16]. Since all nodes in the same group are stochastically identical in terms of mobility pattern and epidemic attribute, the rate $\lambda^{\mathrm{eff}}_{a,b}$ should be determined by the group indices of nodes a and b. Thus, we can rewrite the infection rate as $\lambda^{\mathrm{eff}}_{a,b} = \lambda^{\star}_{\mathcal{F}(a),\mathcal{F}(b)}$, where the subscripts $\mathcal{F}(a)$ and $\mathcal{F}(b)$ denote the group indices of node a and b, respectively. For later use, we define a rate matrix $\mathbf{\Lambda}$ as

$$\mathbf{\Lambda} \triangleq \left(\lambda^{\star}_{k_1,k_2}\right)_{1\leq k_1,k_2\leq K}.$$

Our spreading model is general in that it covers a variety of scenarios from homogeneous to totally heterogeneous cases. For instance, when $K = 1$ our model reduces to the homogeneous case where all nodes in the network are identical with the same infection rate $\lambda^{\star}_{\mathcal{F}(a),\mathcal{F}(b)} = \lambda^{\star}_{1,1}$ ($\triangleq \lambda^{\star}$) for any a, b. On the other hand, when $K = N$ our model induces a totally heterogeneous case where each individual node uniquely forms a group. When $K = 2, \ldots, N-1$, our model is able to capture heterogeneity arising from multiple communities. In addition to heterogeneity, our model is capable of characterizing the impact of various spread parameters such as the level of contact rates and population size by varying the values of the rate matrix $\mathbf{\Lambda}$ and group sizes $\boldsymbol{N}$.

C. Performance Metrics

In this subsection, we describe our performance metrics in detail. Let $S_k(t)$ be the number of susceptible nodes in group k at time $t\,(\geq 0)$. Let $I_k(t)$ be the number of infected nodes in group k at time t. Then, we have $S_k(t) + I_k(t) = N_k$ for all k and t. The key performance metric of our interest is the α-completion time as defined below:

Definition 1 (α-completion time)**.** *For $\alpha \in (0, 1]$, the time required for infecting α fraction of the total population, denoted by T_α, is given by:*

$$T_\alpha \triangleq \inf\left\{t : \sum_{k=1}^{K} I_k(t) \geq \alpha N\right\}. \tag{3}$$

We call T_α the α-completion time *throughout this paper.*

T_α has a strong connection with existing studies that have characterized the average number of infected nodes at time t (i.e., $\sum_k \mathrm{E}[I_k(t)]$) using various mathematical tools, because $\mathrm{E}[T_\alpha]$ is a dual of $\sum_k \mathrm{E}[I_k(t)]$. However, to better understand the spread behavior and to better design spread prevention or acceleration methods, characterization of the distribution of T_α beyond simply the mean is needed. To this end, we introduce a new metric, called (α, β)-guaranteed time, as defined next:

Definition 2 ((α, β)-guaranteed time)**.** *For $\alpha \in (0, 1]$ and $\beta \in (0, 1)$, the minimum time required to guarantee spread to α fraction of the total population with probability β, denoted by $G_{\alpha,\beta}$, is defined by:*

$$G_{\alpha,\beta} \triangleq \inf\left\{t : P\{T_\alpha > t\} \leq 1 - \beta\right\}. \tag{4}$$

We call $G_{\alpha,\beta}$ the (α, β)-guaranteed time *throughout this paper.*

Note that the probability $1 - \beta$ in (4) can be interpreted as an outage probability that the actual spread time T_α exceeds the guaranteed time $G_{\alpha,\beta}$. In this sense, $G_{\alpha,\beta}$ can be used to predict the range of spread time and the confidence of the prediction: the higher we set the value of β, the greater the confidence in the prediction. Thus, $G_{\alpha,\beta}$ facilitates avoiding underestimating the required resources for spreading information to a network. The ratio $R_{\alpha,\beta}$ defined below describes just how much $\mathrm{E}[T_\alpha]$ underestimates the spread time compared to the guaranteed time.

Definition 3 ((α, β)-average to guaranteed time ratio)**.** *For $\alpha \in (0, 1]$ and $\beta \in (0, 1)$, the ratio $R_{\alpha,\beta}$ is defined by*

$$R_{\alpha,\beta} \triangleq \frac{G_{\alpha,\beta}}{E[T_\alpha]}. \tag{5}$$

We call $R_{\alpha,\beta}$ the (α, β)-average to guaranteed time ratio *throughout this paper.*

Finally, we define the set of seeds in each group. Let $s_k \triangleq I_k(0)$ be the number of seeds in group k. If $\sum_k s_k \geq \alpha N$, then we have a trivial result of $T_\alpha = 0$, $G_{\alpha,\beta} = 0$, and $R_{\alpha,\beta} = 1$ for any $\beta \in (0, 1)$. Therefore, in the rest of the paper, we only consider the regime of $\sum_k s_k < \alpha N$.

III. Basic Temporal Analysis Framework

In this section, we develop an analytical framework for deriving the performance metrics defined in (3), (4), and (5). We use the following three steps in our analysis: first, we identify the temporal behavior of the total number of infected nodes $\{\sum_k I_k(t); t \geq 0\}$ (See Lemma 1). Using the result in Lemma 1, we are able to obtain the distribution of the α-completion time T_α (See Lemma 2). Finally, we give formulas for our performance metrics (See Lemma 3).

Step 1: According to Definition 1, we need the temporal distribution of the total number of infected nodes $\sum_k I_k(t)$. Directly solving it appears to be intractable (as illustrated in the following Example II for the case $K = 2$). However, we prove that the *joint temporal distribution* of $I_k(t)$ can be derived from the theory of multi-dimensional CTMC, which could be also used to identify the distribution of $\sum_k I_k(t)$. The result is summarized in Lemma 1 and its derivation is explained through the following Examples I and II.

Lemma 1 (CTMC model)**.** *For $K = 1, 2, \ldots$, let*

$$\boldsymbol{I}(t) \triangleq \big(I_1(t), I_2(t), \ldots, I_K(t)\big).$$

Then, the process $\{\boldsymbol{I}(t); t \geq 0\}$ is a K-dimensional CTMC. Further, it has the following properties:

(P1) The state space is given by $\mathcal{E} \triangleq \prod_{k=1}^{K}\{0, \ldots, N_k\} \setminus \mathbf{0}$ and is decomposed into transient state space $\mathcal{E}^\star$ and absorbing state space $\mathcal{E}^o$ as:

$$\mathcal{E}^\star \triangleq \{\boldsymbol{e} \in \mathcal{E} : |\boldsymbol{e}| < |\boldsymbol{N}|\},$$
$$\mathcal{E}^o \triangleq \{\boldsymbol{N} = (N_1, \ldots, N_K)\}.$$

Without loss of generality, we assume that the states in $\mathcal{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\}$ are arranged as $|\boldsymbol{e}_1| \leq |\boldsymbol{e}_2| \leq \cdots$.

(P2) By the property (P1), the infinitesimal generator $\boldsymbol{Q}$ of the Markov chain is of the following form:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{F}^o \\ \boldsymbol{0} & 0 \end{bmatrix},$$

where $\boldsymbol{F} = (F_{i,j})$ is a matrix representing transition rate from $\mathcal{E}^\star$ to $\mathcal{E}^\star$, and $\boldsymbol{F}^o$ is a column vector representing transition rate from $\mathcal{E}^\star$ to $\mathcal{E}^o$. Due to its importance, we call the matrix $\boldsymbol{F}$ the fundamental matrix.

(P3) Assume $P\{\boldsymbol{I}(0) \in \mathcal{E}^\star\} = 1$. For a given time $t > 0$, let $\boldsymbol{\pi}(t) \triangleq (P\{\boldsymbol{I}(t) = \boldsymbol{e}\})_{\boldsymbol{e}\in\mathcal{E}^\star}$ be the distribution of $\boldsymbol{I}(t)$ on $\mathcal{E}^\star$. Then, it is determined from its initial distribution $\boldsymbol{\pi}(0)$ and the fundamental matrix $\boldsymbol{F}$ as [17]:

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\exp(\boldsymbol{F}t).$$

The distribution of $\boldsymbol{I}(t)$ on $\mathcal{E}^o$ is then obtained by $P\{\boldsymbol{I}(t) = \boldsymbol{N}\} = 1 - |\boldsymbol{\pi}(t)|$.

(P4) Let the ith and the jth states in $\mathcal{E}$ be denoted by $\boldsymbol{e}_i = (i_k)_{1\le k\le K}$ and $\boldsymbol{e}_j = (j_k)_{1\le k\le K}$, respectively. Then, $\boldsymbol{Q} = (Q_{i,j})$ is obtained as:

$$Q_{i,j} = \begin{cases} \sum_l I_{(\boldsymbol{e}_i,\boldsymbol{e}_j):l}(N_l - i_l)\sum_k i_k\lambda^\star_{k,l} & \text{if } i \neq j, \\ -\sum_{l\neq i} Q_{i,l} & \text{if } i = j, \end{cases}$$

where $I_{(\boldsymbol{e}_i,\boldsymbol{e}_j):l} \in \{0,1\}$ and takes 1 if and only if $j_l = i_l + 1$ and $j_k = i_k$ for all $k \neq l$. Then, we can obtain $\boldsymbol{F}$ by restricting $\boldsymbol{Q}$ to the space $\mathcal{E}^\star$ as $\boldsymbol{F} = \boldsymbol{Q}|_{\mathcal{E}^\star\times\mathcal{E}^\star}$, i.e., $\boldsymbol{F} = (Q_{i,j})$ for all i, j such that $\boldsymbol{e}_i, \boldsymbol{e}_j \in \mathcal{E}^\star$.

Proof: Refer to Appendix B of our technical report [16]. ■

Example I. (Homogeneous model, Single community model) We start our analysis with the simplest case of $K = 1$ (i.e., homogeneous model), and drop the group index in all notations for simplicity. In this case, we have $\boldsymbol{I}(t) = I(t)$ and $\mathcal{E} = \{1, \dots, N\}$. Then, we can identify the temporal behavior of $I(t)$ as follows: first note that the process $\{I(t); t \ge 0\}$ is a *counting process* in that it counts the number of events that have taken place during $(0, t]$. Hence, state transitions occur only to the adjacent state from $i\,(= 1, \dots, N-1)$ to $i+1$, and then eventually the system is absorbed to state N. Thus, the state space $\mathcal{E}$ is decomposed into transient state space $\mathcal{E}^\star = \{1, \dots, N-1\}$ and absorbing state space $\mathcal{E}^o = \{N\}$. Suppose that the system enters state $i \in \mathcal{E}^\star$ at time t_0. Let X_i be the sojourn time of state i. Note that the sojourn time is equivalent to the time to have one more infected node, which is the same as the minimum infection time from i number of infected nodes to $N - i$ number of susceptible nodes, i.e.,

$$X_i = \min\left\{M^{\text{eff}}_{a,b};\, a \in \mathcal{I}(t_0), b \in \mathcal{S}(t_0)\right\},$$

where $\mathcal{I}(t)$ and $\mathcal{S}(t)$ denote index sets of infected nodes and susceptible nodes at time t, respectively. Since $M^{\text{eff}}_{a,b} \sim \text{Exp}(\lambda^\star)$ from (2) and is independent for all nodes, we have:

$$X_i \sim \text{Exp}(i(N-i)\lambda^\star).$$

Therefore, the process $\{I(t); t \ge 0\}$ is a CTMC with transition diagram depicted in Fig. 1. From the transition diagram, we can easily obtain the matrix $\boldsymbol{F}$. For details, refer to Appendix C of our technical report [16].

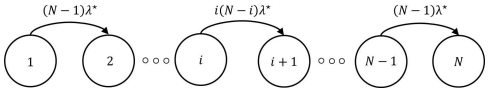


Fig. 1. Transition diagram of the Markov chain $\{I(t); t \ge 0\}$ when $K = 1$.

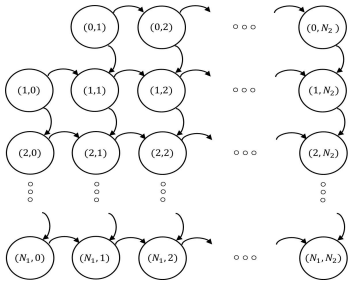


Fig. 2. Transition diagram of the Markov chain $\{(I_1(t), I_2(t)); t \ge 0\}$ when $K = 2$. The rate from (i_1, i_2) to $(i_1 + 1, i_2)$ is $(N_1 - i_1)\sum_{k=1}^2 i_k\lambda^\star_{k,1}$, and the rate from (i_1, i_2) to $(i_1, i_2 + 1)$ is $(N_2 - i_2)\sum_{k=1}^2 i_k\lambda^\star_{k,2}$.

Example II. (Double community model) We next consider the case when $K = 2$. In this case, if we set the state variable as the total number of infected nodes (i.e., $\sum_k I_k(t)$), then it becomes intractable to identify the statistics of sojourn time X_i of state i, unless we know how the overall infected nodes in the network are distributed to each group. For this reason, we set the vector $(I_1(t), I_2(t))$ as the state variable. Suppose that at time t_0, the system enters state (i_1, i_2). Since the process $\{I_1(t) + I_2(t); t \ge 0\}$ is a counting process, the very next state transitions occur only to either (i_1+1, i_2) or (i_1, i_2+1), and then eventually the system is absorbed to state (N_1, N_2). Hence, state space $\mathcal{E} = \{(0,1), (1,0), (1,1), \dots, (N_1, N_2)\}$ is decomposed into transient state space $\mathcal{E}^\star = \{\boldsymbol{e} \in \mathcal{E} : |\boldsymbol{e}| < N_1 + N_2\}$ and absorbing state space $\mathcal{E}^o = \{(N_1, N_2)\}$. For $(i_1, i_2) \in \mathcal{E}^\star$, let $X_{(i_1,i_2):i_1}$ and $X_{(i_1,i_2):i_2}$ be the time required to infect one additional node in groups 1 and 2, respectively. Then, by a similar reason as in Example I, we have

$$X_{(i_1,i_2):i_1} = \min\left\{M^{\text{eff}}_{a,b}\,; a \in \mathcal{I}_1(t_0) \cup \mathcal{I}_2(t_0), b \in \mathcal{S}_1(t_0)\right\},$$

where $\mathcal{I}_k(t)$ and $\mathcal{S}_k(t)\,(k = 1, 2)$ denote index sets of infected nodes and susceptible nodes in group k at time t, respectively. Thus, $X_{(i_1,i_2):i_1}$ follows an exponential distribution as in Example I, but in this case the rate is given by $i_1(N_1 - i_1)\lambda^\star_{1,1} + i_2(N_1 - i_1)\lambda^\star_{2,1} = (N_1 - i_1)\sum_{k=1}^2 i_k\lambda^\star_{k,1}$. Similarly, we have that $X_{(i_1,i_2):i_2}$ follows an exponential distribution as summarized below:

$$X_{(i_1,i_2):a} \sim \begin{cases} \text{Exp}\big((N_1 - i_1)\sum_{k=1}^2 i_k\lambda^\star_{k,1}\big) & \text{if } a = i_1, \\ \text{Exp}\big((N_2 - i_2)\sum_{k=1}^2 i_k\lambda^\star_{k,2}\big) & \text{if } a = i_2. \end{cases} \quad (6)$$

Note that the sojourn time $X_{(i_1,i_2)}$ of state (i_1, i_2) is the minimum value between $X_{(i_1,i_2):i_1}$ and $X_{(i_1,i_2):i_2}$. Hence, from (6), $X_{(i_i,i_2)}$ follows an exponential distribution. Therefore, the process $\{(I_1(t), I_2(t)); t \ge 0\}$ is a 2-dimensional CTMC with transition diagram depicted in Fig. 2. From the transition diagram, we can easily obtain the fundamental matrix $\boldsymbol{F}$. For details, refer to Appendix C of our technical report [16].

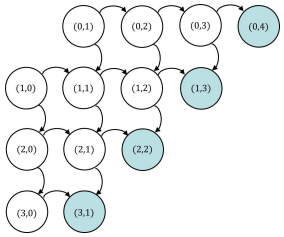


Fig. 3. An example of $\mathcal{E}_\alpha (= \mathcal{E}_\alpha^\star \cup \mathcal{E}_\alpha^o)$ for $(N_1, N_2) = (3, 5)$, $\lceil \alpha N \rceil = 4$: shaded states form $\mathcal{E}_\alpha^o$, and the others form $\mathcal{E}_\alpha^\star$.

Step 2: Using the results in Lemma 1, we can derive the distribution of T_α. We take two steps: (i) first, we truncate the state space $\mathcal{E}$ to $\mathcal{E}_\alpha \triangleq \{\boldsymbol{e} \in \mathcal{E} : |\boldsymbol{e}| \le \lceil \alpha N \rceil\}$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. (ii) Next, we split the truncated state space $\mathcal{E}_\alpha$ into transient state space $\mathcal{E}_\alpha^\star$ and absorbing state space $\mathcal{E}_\alpha^o$ as:

$$\mathcal{E}_\alpha^\star \triangleq \{\boldsymbol{e} \in \mathcal{E}_\alpha : |\boldsymbol{e}| < \lceil \alpha N \rceil\},$$
$$\mathcal{E}_\alpha^o \triangleq \{\boldsymbol{e} \in \mathcal{E}_\alpha : |\boldsymbol{e}| = \lceil \alpha N \rceil\}.$$

On the state space $\mathcal{E}_\alpha^\star \cup \mathcal{E}_\alpha^o$, we define a truncated process $\boldsymbol{I}_\alpha(t)$ from the process $\boldsymbol{I}(t)$ as follows: $\boldsymbol{I}_\alpha(t)$ evolves according to $\boldsymbol{I}(t)$ unless $\boldsymbol{I}(t) \in \mathcal{E}_\alpha^o$. When $\boldsymbol{I}(t)$ enters one of states in $\mathcal{E}_\alpha^o$, say $\boldsymbol{e}$, truncation happens and $\boldsymbol{I}_\alpha(t)$ is absorbed to the state $\boldsymbol{e}$. Then, by Lemma 1 the process $\{\boldsymbol{I}_\alpha(t); t \ge 0\}$ is a K-dimensional CTMC with possibly multiple absorbing states in $\mathcal{E}_\alpha^o$. Moreover, by Definition 1, T_α is the time taken by the truncated Markov chain to be absorbed into $\mathcal{E}_\alpha^o$. An example of transition diagram is shown in Fig. 3.

Similarly to (P2) in Lemma 1, the infinitesimal generator $\boldsymbol{Q}_\alpha$ of the process $\{\boldsymbol{I}_\alpha(t); t \ge 0\}$ is of the following form:

$$\boldsymbol{Q}_\alpha = \begin{bmatrix} \boldsymbol{F}_\alpha & \boldsymbol{F}_\alpha^o \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

Here, $\boldsymbol{F}_\alpha$ is a matrix representing transition rate from $\mathcal{E}_\alpha^\star$ to $\mathcal{E}_\alpha^\star$, and can be obtained from the fundamental matrix $\boldsymbol{F}$ of the original process $\{\boldsymbol{I}(t); t \ge 0\}$ by

$$\boldsymbol{F}_\alpha = \boldsymbol{F}|_{\mathcal{E}_\alpha^\star \times \mathcal{E}_\alpha^\star} (= \boldsymbol{Q}|_{\mathcal{E}_\alpha^\star \times \mathcal{E}_\alpha^\star}). \tag{7}$$

Similarly, $\boldsymbol{F}_\alpha^o$ is a matrix representing transition rate from $\mathcal{E}_\alpha^\star$ to $\mathcal{E}_\alpha^o$, and is obtained by $\boldsymbol{F}_\alpha^o = \boldsymbol{Q}|_{\mathcal{E}_\alpha^\star \times \mathcal{E}_\alpha^o}$. Therefore, the value α determines where to truncate the matrix $\boldsymbol{F}$ or $\boldsymbol{Q}$ in Lemma 1 and how to redefine transient and absorbing state spaces. For α values satisfying $\lceil \alpha N \rceil = N$, we have $\mathcal{E}_\alpha^\star = \mathcal{E}^\star$ and $\mathcal{E}_\alpha^o = \mathcal{E}^o$, which gives $\boldsymbol{F}_\alpha = \boldsymbol{F}$ and $\boldsymbol{F}_\alpha^o = \boldsymbol{F}^o$.

Once we have the truncated fundamental matrix $\boldsymbol{F}_\alpha$ from the original matrix $\boldsymbol{F}$, we can obtain the distribution of T_α as in the following lemma.

Lemma 2 (Distribution of T_α)**.** *The cumulative distribution function (CDF) of the α-completion time is given by*

$$H_\alpha(t) \triangleq P\{T_\alpha \le t\} = 1 - \boldsymbol{h}_\alpha \exp(\boldsymbol{F}_\alpha t)\boldsymbol{1},$$

where $\boldsymbol{h}_\alpha \triangleq (P\{\boldsymbol{I}_\alpha(0) = \boldsymbol{e}\})_{\boldsymbol{e} \in \mathcal{E}_\alpha^\star}$ is a row vector denoting the initial distribution, $\boldsymbol{F}_\alpha$ is given in (7), and $\boldsymbol{1}$ is a column vector of ones. In addition, it can be expressed as in the following form [18]:

$$H_\alpha(t) = 1 - \sum_i \exp(-|\rho_i|t) P_i(t),$$

where ρ_i is the ith eigenvalue of $\boldsymbol{F}_\alpha$ with multiplicity denoted by m_i, and $P_i(t)$ is a $(m_i - 1)$th order polynomial function of t. Since $\boldsymbol{F}_\alpha$ is an upper triangular matrix, the eigenvalues are from the distinct diagonal elements of the matrix, which are all real and negative.

Proof: Refer to Appendix D of our technical report [16]. ■

Step 3: Based on Lemma 2, we can derive formulas for our performance metrics, as shown in Lemma 3.

Lemma 3 (Formulas for $G_{\alpha,\beta}$ and $R_{\alpha,\beta}$)**.** *The inverse function of the distribution function $H_\alpha(\cdot)$ in Lemma 2 exists and yields the (α, β)-guaranteed time in (4) as*

$$G_{\alpha,\beta} = H_\alpha^{-1}(\beta).$$

The fundamental matrix $\boldsymbol{F}_\alpha$ is invertible, and its inverse matrix gives the nth moment of T_α as $E[(T_\alpha)^n] = n! \boldsymbol{h}_\alpha (-\boldsymbol{F}_\alpha)^{-n} \boldsymbol{1}$ $(n = 1, 2, \ldots)$. Therefore, the ratio $R_{\alpha,\beta}$ is obtained by

$$R_{\alpha,\beta} = \frac{H_\alpha^{-1}(\beta)}{\boldsymbol{h}_\alpha (-\boldsymbol{F}_\alpha)^{-1} \boldsymbol{1}}.$$

Proof: Refer to Appendix E of our technical report [16]. ■

Major applications leveraging $G_{\alpha,\beta} (= G_{\alpha,\beta}(\boldsymbol{\Lambda}, \boldsymbol{s}))$ include the followings:

1) For distributing a firmware or a software update to smartphones (and tablets) through opportunistic contacts among nodes when cellular network carriers wish to avoid abusing network resources while guaranteeing the time to deliver the update with more than 99% of confidence, $G_{\alpha,\beta}$ becomes significantly useful to determine the required number of seeds in the network. For instance, to guarantee delivery with probabilityβ for α fraction of nodes within time T_{bound}, the number of seeds $\boldsymbol{s}$ who directly get the update from the carriers can be determined from:
$$\boldsymbol{s} = G_{\alpha,\beta}^{-1}(\boldsymbol{\Lambda}, T_{\text{bound}}).$$
2) For an autonomous disaster broadcasting system, which purely leverages opportunistic contacts without relying on network infrastructures, the target level of infection rates $\boldsymbol{\Lambda}$, which achieves a desirable time bound T_{bound}, can be determined by:
$$\boldsymbol{\Lambda} = G_{\alpha,\beta}^{-1}(T_{\text{bound}}, \boldsymbol{s})$$
for given (α, β). Based on this prediction, we can scale up or down the infection rates $\boldsymbol{\Lambda}$ among nodes by optimally controlling the communication ranges of mobile devices.
3) For a highly contagious disease emerged at a city, if medical facilities in the city have capacity for up to α portion of citizens who typically have $\boldsymbol{\Lambda}$ infection rates, the regional government can estimate the allowed time to execute emergency plans by referring to:
$$T_{\text{bound}} = G_{\alpha,\beta}(\boldsymbol{\Lambda}, \boldsymbol{s}).$$

IV. Analytical Characteristics and Applications

In this section, we present analytical characteristics derived from our framework, and provide how to utilize these characteristics in practical applications.

A. Impact of the level of infection rates

The behavior of information spread is determined by various spreading factors. Using our framework, we first answer the question on how the level of infection rates $\lambda_{a,b}^{\text{eff}}$ affect the distribution of α-completion time.

Theorem 1 (Impact of the level of infection rates)**.** *Suppose that the infection rate $\lambda_{a,b}^{\text{eff}}$ is scaled by $\gamma\,(>0)$ times for all a, b. Let $\hat{T}_\alpha$, $\hat{G}_{\alpha,\beta}$, and $\hat{R}_{\alpha,\beta}$ be the correspondences of T_α, $G_{\alpha,\beta}$, and $R_{\alpha,\beta}$ after the scale, respectively. Then, for any $\alpha \in (0,1]$, we have*

$$\hat{T}_\alpha \overset{d}{=} \gamma^{-1} T_\alpha, \tag{8}$$

where $\overset{d}{=}$ denotes "equal in distribution." The relationship in (8) yields for any $\alpha \in (0,1]$ and $\beta \in (0,1)$ the followings:

$$\hat{G}_{\alpha,\beta} = \gamma^{-1} G_{\alpha,\beta},$$
$$E[(\hat{T}_\alpha)^n] = \gamma^{-n} E[(T_\alpha)^n].$$

Hence, $\hat{R}_{\alpha,\beta}$ becomes $\hat{R}_{\alpha,\beta} = R_{\alpha,\beta}$.

Proof: Refer to Appendix F of our technical report [16]. ■

The result in Theorem 1 shows that the spread becomes faster *proportionally* to the level of infection rates in *distribution sense*. Similarly, we show that the average $\mathcal{M}(t) \triangleq \sum_k \mathrm{E}[I_k(t))]$ and its time derivative $\mathcal{D}(t) \triangleq \frac{d}{dt}\mathcal{M}(t)$ scale respectively as $\hat{\mathcal{M}}(t) = \mathcal{M}(\gamma t)$ and $\hat{\mathcal{D}}(t) = \gamma \mathcal{D}(\gamma t)$ for all $t \geq 0$. The detailed proof is given in our technical report [16].

B. Impact of population size

We next characterize the impact of population size on information spread. In our epidemic model, each non-seed node can be considered as a workload to finish. However, once the node becomes infected, it works in a similar manner as the seed and is involved in spreading the information. Hence, it is not straightforward whether the population size accelerates or slows down the speed of information spread. Our framework gives the answer, as shown in Theorem 2.

Theorem 2 (Impact of population size)**.** *Suppose $\alpha = 1$ (i.e., spread completion), $K = 1$ (i.e., homogeneous model), and $s_1 = 1$ (i.e., one seed). As the population size N increases, we have*

- *$G_{\alpha,\beta}$ is strictly decreasing for sufficiently large β.*
- *$E[T_\alpha]$ is strictly decreasing.*

In addition, it scales respectively as

- $G_{\alpha,\beta} = \Theta\big((\lambda^\star)^{-1} N^{-1} (\log N - \log(\log \frac{1}{\beta}))\big)$.
- $E[T_\alpha] = \Theta\big((\lambda^\star)^{-1} N^{-1} \log N\big)$.

Hence, $R_{\alpha,\beta}$ scales as $\Theta(1)$.

Proof: Refer to Appendix H of our technical report [16]. ■

The results in Theorem 2 indicate that adding a node in the system accelerates the information spread when per-pair infection rates are unchanged.

Remark 1. *To assist understanding of Theorem 2, we consider a non-cooperative spread model, where seed chosen at the beginning only spreads the information. As the population size N increases, we have*

- *$G_{\alpha,\beta}$ is strictly increasing for sufficiently large β.*
- *$E[T_\alpha]$ is strictly increasing.*

In addition, it scales respectively as

- $G_{\alpha,\beta} = \Theta\big((\lambda^\star)^{-1} (\log N - \log(\log \frac{1}{\beta}))\big)$.
- $E[T_\alpha] = \Theta\big((\lambda^\star)^{-1} \log N\big)$.

Hence, $R_{\alpha,\beta}$ scales as $\Theta(1)$.

More properties of our model (namely, cooperative model) and the non-cooperative are compared in the following table:

	Cooperative model	Non-cooperative model
Variance of T_α	Strictly decrease with N and scale as $\Theta((\lambda^\star N)^{-2})$	Strictly increase with N and converge to $(\lambda^\star)^{-2}\zeta(2)$
Skewness of T_α	Strictly decrease with N and scale as $\Theta((\lambda^\star N)^{-3})$	Strictly increase with N and converge to $(\lambda^\star)^{-3}\zeta(3)$
$\mathrm{E}[(T_\alpha)^n]$	$\mathrm{E}[(T_\alpha)^n] < \infty$	$\mathrm{E}[(T_\alpha)^n] = \infty$

In the table, $\zeta(c) \triangleq \sum_{n=1}^{\infty} n^{-c}$ denotes the Riemann zeta function. The proof for the results in Remark 1 is given in Appendix I of our technical report [16]. Our analysis showing that $G_{\alpha,\beta}$ behaves differently for the scaling of N and $\lambda^\star$ tells that resource allocation for information spread should be carefully designed based on the willingness of cooperation in a spread process (i.e., infectivity in a spread process).

C. Impact of multiple community

The impact of heterogeneity in information or virus spreading has been less explored. Using our CTMC-based framework, we analyze and understand the temporal spread behavior under a heterogeneous network with multiple groups compared with a homogeneous network. In particular, we focus on answering "Does heterogeneity persistently expedite the spreading or not?", "Is there an optimal heterogeneity level for information spread?", and "Is there an upper or a lower bound on the gain from the heterogeneity over homogeneity?".

In this subsection, we provide the answers to these questions by studying dual community model ($K = 2$). Note that our framework can be easily extended to study the cases when $K \geq 3$. In order to focus on heterogeneity arising from multiple community, we make assumptions as follows: (i) two groups are of the same size $N_1 = N_2 (= N/2)$. (ii) The inter-group infection rates are the same for both directions, i.e., $\lambda_{1,2}^\star = \lambda_{2,1}^\star$ (iii) There is one seed. Without loss of generality, the seed is chosen arbitrarily from group 1.

Let $\gamma_1 \triangleq \lambda_{1,1}^\star / \lambda_{1,2}^\star$ and $\gamma_2 \triangleq \lambda_{2,2}^\star / \lambda_{1,2}^\star$. The values of γ_1 and γ_2 control the intra-group infection rates, and are chosen freely in the range $0 \leq \gamma_1, \gamma_2 < \infty$. Note that $(\gamma_1, \gamma_2) = (1,1)$ reduces to the homogeneous model and larger deviation from $(1,1)$ induces more heterogeneity. For a fair comparison with a homogeneous model of size N and infection rate $\lambda^\star$, we use the following constraint that represents the same average infection rate:

$$\lambda^\star = \frac{\sum_a \sum_{b \neq a} \lambda_{g(a),g(b)}^\star}{N(N-1)}. \tag{9}$$

With the help of Theorems 1 and 2 showing the scaling of $\lambda^\star$ and N, we can characterize and generalize the impact of heterogeneity by only observing a specific setting of $(\lambda^\star, N)$. For simplicity, we choose $(1, 40)$. We then vary (γ_1, γ_2) in the

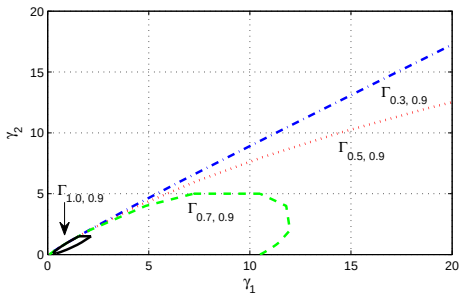


Fig. 4. Comparison of (α,β)-guaranteed time $G_{\alpha,\beta}$ with the homogeneous model for $\beta = 0.9$ and $\alpha = 0.3, 0.5, 0.7, 1.0$: if $(\gamma_1,\gamma_2) \in \Gamma_{\alpha,\beta}$, then heterogeneity in multiple community accelerates the information spread (i.e., reduces the guaranteed time $G_{\alpha,\beta}$). If $(\gamma_1,\gamma_2) \notin \Gamma_{\alpha,\beta}$, then heterogeneity slows down the information spread.

range $0 \le \gamma_1, \gamma_2 \le 20$. From Lemma 3, we obtain the (α,β)-guaranteed time $G_{\alpha,\beta}$ and compare it with the homogeneous counterpart. Fig. 4 shows the result. In the figure, $\Gamma_{\alpha,\beta}$ is the region such that if $(\gamma_1,\gamma_2) \in \Gamma_{\alpha,\beta}$, then heterogeneity yields reduced guaranteed time $G_{\alpha,\beta}$, compared with the homogeneous model, and vice versa. Hence, the region $\Gamma_{\alpha,\beta}$ can be interpreted as the area where heterogeneity accelerates the information spread. From the figure, we can observe the followings: (i) as α increases, the region $\Gamma_{\alpha,\beta}$ shrinks. Hence, for a fixed (γ_1,γ_2), there exists a threshold α_{th} such that $(\gamma_1,\gamma_2) \in \Gamma_{\alpha,\beta}$ if $\alpha \le \alpha_{\text{th}}$ and $(\gamma_1,\gamma_2) \notin \Gamma_{\alpha,\beta}$ if $\alpha > \alpha_{\text{th}}$. In addition, the threshold decreases as (γ_1,γ_2) deviates from (1,1). This implies that heterogeneity accelerates the spread at beginning phase (i.e., $\alpha \le \alpha_{\text{th}}$) while slowing down the spread at ending phase (i.e., $\alpha > \alpha_{\text{th}}$), and the time portion of being accelerated shrinks with more heterogeneity. (ii) For any $\alpha \in \{0.3, 0.5, 0.7, 1.0\}$, there is a non-empty region $\bigcap_\alpha \Gamma_{\alpha,\beta}$, where *heterogeneity always accelerates the information spread* (i.e., $\alpha_{\text{th}} = 1$ for all α). (iii) In the region $\{(\gamma_1,\gamma_2) : \gamma_1 < \gamma_2\}$, heterogeneity always slows down the information spread. That is, if the seed is chosen from a less infective group, then heterogeneity never accelerates the information spread.

As a special case, we consider a system where the inter-group infection rate is determined from intra-group infection rates by $\lambda^\star_{1,2} = (\lambda^\star_{1,1} + \lambda^\star_{2,2})/2$, and the seed is chosen from more infective group. Let $\gamma \triangleq \max\{\lambda^\star_{1,1}, \lambda^\star_{2,2}\}/\min\{\lambda^\star_{1,1}, \lambda^\star_{2,2}\}$. For fixed $(\lambda^\star, N) = (1, 40)$ as above, we vary γ as $\gamma = 1, 2, 4, 8$, and show the (α,β)-guaranteed time $G_{\alpha,\beta}$ in Fig. 5. From the figure, we confirm that heterogeneity indeed accelerates the spread for smaller penetration (i.e., for low α) but slows down it for higher penetration. This observation is proved in Theorem 3.

Theorem 3 (Impact of multiple community)**.** *Let*

$$D_\alpha(\gamma) \triangleq -\lim_{t\to\infty} \frac{\log P\{T_\alpha(\gamma) > t\}}{t},$$

where $T_\alpha(\gamma)$ denotes the α-completion time when γ is used. Then, $D_\alpha(\gamma)$ exists and satisfies the followings:

- *If $\alpha \le 1 - \frac{2}{N}$, then $\frac{d}{d\gamma} D_\alpha(\gamma) > 0$ for all $\gamma \ge 1$.*
- *If $\alpha = 1$, then $\frac{d}{d\gamma} D_\alpha(\gamma) < 0$ for all $\gamma \ge 1$.*

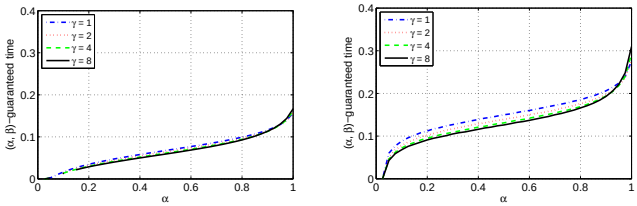


Fig. 5. (α,β)-guaranteed time $G_{\alpha,\beta}$ for $\beta = 0.1$ (left) and $\beta = 0.9$ (right).

- *If $1 - \frac{2}{N} < \alpha < 1$, then $\frac{d}{d\gamma} D_\alpha(\gamma) > 0$ for $\gamma < \frac{5N-16}{N-4}$ and $\frac{d}{d\gamma} D_\alpha(\gamma) < 0$ for $\gamma > \frac{5N-16}{N-4}$.*

Proof: Refer to Appendix J of our technical report [16]. ■

D. Contribution of each node to the information spread

In this section, we provide a method for quantifying the contribution of each individual node to the information spread. The quantification can be useful, e.g., for cellular carriers in incentivizing a node who contributes to alleviate data deluge in cellular networks by distributing packets through opportunistic contacts among nodes. Such an evaluation tool is of importance especially when nodes have heterogeneous attributes in spreading the information. Let C_i denote the degree of contribution of node i to the spread. In this work, we evaluate C_i by using the concept of the *Shapely value* [19], which is known as a good metric measuring the surplus (or the contribution) of a node in the cooperative game theory:

$$C_i \triangleq \frac{G_{\alpha,\beta}(\mathcal{N} \setminus \{i\})}{G_{\alpha(N-1)/N,\beta}(\mathcal{N})}, \tag{10}$$

where $\mathcal{N} \triangleq \{1, \ldots, N\}$ is the index set of nodes, and for an index set $\mathcal{A}$, $G_{\alpha,\beta}(\mathcal{A})$ is the (α,β)-guaranteed time for the network consisting of nodes $i \in \mathcal{A}$. Hence, the numerator and the denominator in (10) denote the guaranteed times in the network *without* and *with* the node i, respectively, and consequently a node i with high contribution to the information spread has a large C_i value. Due to page limitation, we omit detailed application of the metric C_i and its analysis.

E. Applications

How to optimally distribute given resources to nodes in a network to minimize the time for spreading of information to the network is of an important research question. Our results presented in this section provide initial understanding to this question. Theorem 3 proves that when the number of nodes N increases, heterogeneity in $\mathbf{\Lambda}$ expedites the spread of information for most of the time except some time duration at the end of spread, where the duration converges to zero as N goes to infinity. It is important to point out that our understanding implies the existence of a small region of $\mathbf{\Lambda}$ with heterogeneous contact rates, which always make the spread faster than a network with homogeneous $\mathbf{\Lambda}$. By applying these two observations to designing a network, we have the following applications:

1) For a network delivering information to a community using vehicles or message ferries (e.g., DakNet [20], DieselNet [21], and ZebraNet [22]) of which total amount

of fuel is given, the amount of fuel distributed to each vehicle can be asymmetric to guarantee faster spread of information all the time compared to symmetric distribution.

2) When the number of nodes in a network is extremely large (e.g., users in facebook), advertising a product to the network can be expedited by providing incentives to users to forward information to others in a highly skewed manner. Our results support that evenly distributed incentives to the entire population would lead to much slower spreading compared to unfair incentives. This tells that the same speed of spread can be achieved by only providing a smaller amount of total incentive to the network when incentives are optimally distributed with the understanding of skewness.

V. Simulation Study

We study the efficacy of our framework and characterizations using by far the largest vehicular mobility trace obtained from more than a thousand taxies in Shanghai, China [23].[5] The experimental trace tracked GPS coordinates of taxies at every 30 seconds during 28 days in Shanghai. The trace was analyzed in [24] and it was shown that the taxies have exponentially distributed pairwise inter-contact time, which is well aligned with our CTMC-based framework.

Figs. 6 (a), (b), and (c) characterize the statistics of the taxi network with 1000 randomly chosen taxies in the aspect of *number of contacts*, *number of neighbors in a communication range* (50 meter in our analysis), and *contact duration*, respectively. We apply these three factors for evaluating the effective contact rates $\lambda^{\text{eff}}_{a,b} = \lambda_{a,b}\varphi_a\psi_b$ derived in 2, where $\varphi_a = 1$ and ψ_b is 1 over average number of neighbors multiplied by the expected number of contacts to make a successful data transfer. Note that the latter is derived from the contact time distribution and the time required for a data transfer. The results for a homogeneous network (i.e., $\lambda^\star$) and for a heterogeneous network with two groups (i.e., $\lambda^\star_{1,1}$, $\lambda^\star_{2,2}$ and $\lambda^\star_{1,2}$) are summarized in Table I. Note that the infection rates in Table I satisfy the constraint in (9) that was introduced for a fair comparison between a homogeneous model and a heterogeneous model. Based on the statistics in Table I, we can predict the information spread time and examine possible methods to properly allocate resources for the taxi network.

TABLE I
Infection rates for a homogeneous network and for a heterogeneous network with two groups of taxies.

Homogeneous Network	Heterogeneous Network		
$\lambda^\star$	$\lambda^\star_{1,1}$	$\lambda^\star_{2,2}$	$\lambda^\star_{1,2}\ (=\lambda^\star_{2,1})$
$4.14 \cdot 10^{-4}$	$7.17 \cdot 10^{-4}$	$1.93 \cdot 10^{-4}$	$3.72 \cdot 10^{-4}$

Based on Table I, we simulate probabilistic guarantees for the completion time in a homogeneous and a heterogeneous network, each with 100 taxies. We assume a firmware update to be distributed for mobile devices, which will take around 90 seconds demanding 1.15 number of contacts on average. The number of taxi is scaled down to 100 due to computation complexity involved in matrix operations. Figs. 7 (a), (b), and (c) show the (α, β)-guaranteed time for $\alpha \in [0, 1]$ and $\beta \in \{0.5, 0.9, 0.99\}$ with the number of seeds given by 1, 10, and 20, respectively. The figures tell that if we target 90% penetration with 99% confidence (i.e., $(\alpha, \beta) = (0.9, 0.99)$), then the network with a single seed is estimated to take about 11.6 days (i.e., 278 hours) to achieve the target level of information spread. This estimation largely differs from the existing estimation of average time to achieve 90% of penetration, which is close to 7 days. This clarifies that designing plans associated with the successful spread to 90% of nodes should allow about 4.6 days more. If not, a set of planed work may not be executable on time. If shorter time duration needs to be guaranteed to avoid the plan being delayed, our framework is able to suggest to add seeds to the network as shown in Figs. 7 (b) and (c). As the number of seeds increases to 10 or 20, the time for 90% penetration with 99% confidence reduces from 278 hours to 137 and 113 hours, respectively. These predictions guide how to optimally plan the information spread.

Similarly, we can study a heterogeneous network with two groups. Figs. 7 (d), (e), and (f) show the (α, β)-guaranteed time for $\alpha \in [0, 1]$ and $\beta \in \{0.5, 0.9, 0.99\}$ with 1, 10, and 20 seeds, respectively. Direct comparison between Figs. 7 (a), (b), (c) and Figs. 7 (d), (e), (f) confirms our claims from Theorem 3 that the (α, β)-guaranteed time in a heterogeneous network is faster for lower α, but is slower for higher α close to 1. This implies that if it is mandatory to achieve 100% penetration, making the nodes in a network to be more homogeneous (by providing more resources to relatively inactive nodes) can be helpful, when increasing the level of average contact rates is not possible due to resource concern.

VI. Conclusion

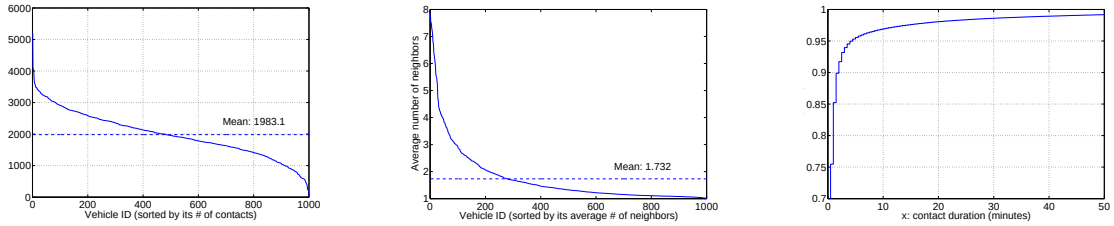
In this paper, we characterize the probabilistic guarantee of the time for information spread in opportunistic networks by developing a CTMC-based analytical framework and introducing the metric $G_{\alpha,\beta}$. We also identify the temporal scaling behavior of information spread for a set of key spread factors. Through various examples of application scenarios and simulations over the Shanghai taxi trace, we show that our framework enables us to estimate proper amount of resource to a network in information spread by providing the detailed statistics of the guaranteed time for given penetration targets. We believe our framework can be viewed as an important first step in the design of highly sophisticated acceleration methods for information spread (or prevention methods for epidemics).

[5] Our framework is applicable to various networks including taxi networks. Due to the availability of data, we limit simulation study to a taxi network.

References

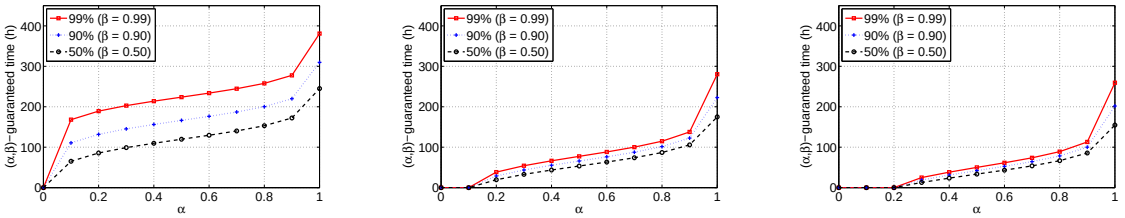
[1] M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell, "Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape," *Science*, vol. 294, no. 5543, pp. 813–817, October 2001.

[2] S. H. Sellke, N. B. Shroff, and S. Bagchi, "Modeling and automated containment of worms," *IEEE Transactions on Dependable and Secure Computing*, vol. 5, pp. 71–86, April 2008.

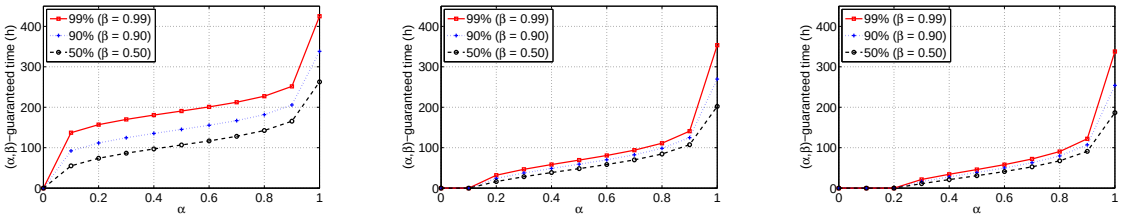


(a) Number of contacts for each vehicle (b) Average number of neighbors (c) CDF of aggregated contact durations

Fig. 6. (a) Number of contacts of a vehicle with all other vehicles during 28 days. (b) Average number of neighbors when a node is in a contact with another node. (c) CDF of aggregated contact durations between all taxi pairs.



(a) (α, β)-guaranteed time with 1 seed (b) (α, β)-guaranteed time with 10 seeds (c) (α, β)-guaranteed time with 20 seeds



(d) (α, β)-guaranteed time with 1 seed (e) (α, β)-guaranteed time with 10 seeds (f) (α, β)-guaranteed time with 20 seeds

Fig. 7. Distribution of the (α, β)-guaranteed time for $\alpha \in [0, 1]$ and $\beta = \{0.5, 0.90, 0.99\}$ with (a) 1 seed, (b) 10 seeds, and (c) 20 seeds in a homogeneous network and with (d) 1 seed, (e) 10 seeds, and (f) 20 seeds in a heterogeneous network with two groups.

[3] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance modeling of epidemic routing," *Elsevier Computer Networks*, vol. 51, no. 10, pp. 2867–2891, July 2007.

[4] P. Jacquet, B. Mans, and G. Rodolakis, "Information propagation speed in mobile and delay tolerant networks," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5001 –5015, October 2010.

[5] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*. Springer, 2000.

[6] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of Royal Society Interface*, vol. 2, no. 4, pp. 295–307, September 2005.

[7] M. Karsai, M. Kivela, R. K. Pan, K. Kaski, J. Kertesz, A.-L. Barabasi, and J. Saramaki, "Small but slow world: How network topology and burstiness slow down spreading," *Physical Review E*, vol. 83, no. 5, p. 025102, February 2011.

[8] P. V. Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 1–14, February 2009.

[9] Z. Yang, A.-X. Cui, and T. Zhou, "Impact of heterogeneous human activities on epidemic spreading," *Physica A*, vol. 390, pp. 4543–4548, 2011.

[10] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis," in *Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS)*, 2002.

[11] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and scalable distribution of content updates over a mobile social network," in *Proceedings of IEEE INFOCOM*, 2009.

[12] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *Proceedings of Intl. Conference on Autonomic Computing and Communication Systems*, 2007.

[13] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proceedings of ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing*, 2009.

[14] Y. Kim, K. Lee, N. B. Shroff, I. Rhee, and S. Chong, "On the generalized delay-capacity tradeoff of mobile networks with Lévy flight mobility," The Ohio State Universi, Tech. Rep., July 2012, available at arXiv: http://arxiv.org/abs/1207.1514.

[15] K. Lee, S. Hong, S. Kim, I. Rhee, and S. Chong, "SLAW: A new human mobility model," in *Proceedings of IEEE INFOCOM*, 2009.

[16] Y. Kim, K. Lee, N. B. Shroff, and I. Rhee, "Providing probabilistic guarantees on the time of information spread in opportunistic networks," The Ohio State Universi, Tech. Rep., January 2013, available at arXiv: http://arxiv.org/abs/1301.2220.

[17] P. Bremaud, *Markov Chains*. Springer, 2008.

[18] O. D. Aalen, "Phase type distributions in survival analysis," *Scandinavian journal of statistics*, vol. 22, no. 4, pp. 447–463, 1995.

[19] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, pp. 307–317, 1953.

[20] A. Pentland, R. Fletcher, and A. Hasson, "Daknet: Rethinking connectivity in developing nations," *Computer*, January 2004.

[21] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Routing for vehicle-based disruption tolerant networks," in *Proceedings of IEEE INFOCOM*, 2006.

[22] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.-S. Peh, and D. Rubenstein, "Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet," in *Proceedings of ASPLOS*, 2002.

[23] S. J. U. Traffic Information Grid Team, Grid Computing Center, "Shanghai taxi trace data," http://wirelesslab.sjtu.edu.cn/.

[24] K. Lee, Y. Yi, J. Jeong, H. Won, I. Rhee, and S. Chong, "Max-contribution: On optimal resource allocation in delay tolerant networks," in *Proceedings of IEEE INFOCOM*, 2010.