This paper was presented as part of the main technical program at IEEE INFOCOM 2011

User-Centric Data Dissemination in Disruption Tolerant Networks

Wei Gao and Guohong Cao
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA 16802
{*weigao, gcao*}*@cse.psu.edu*

***Abstract*—Data dissemination is useful for many applications of Disruption Tolerant Networks (DTNs). Current data dissemination schemes are generally network-centric ignoring user interests. In this paper, we propose a novel approach for user-centric data dissemination in DTNs, which considers satisfying user interests and maximizes the cost-effectiveness of data dissemination. Our approach is based on a social centrality metric, which considers the social contact patterns and interests of mobile users simultaneously, and thus ensures effective relay selection. The performance of our approach is evaluated from both theoretical and experimental perspectives. By formal analysis, we show the lower bound on the cost-effectiveness of data dissemination, and analytically investigate the tradeoff between the effectiveness of relay selection and the overhead of maintaining network information. By trace-driven simulations, we show that our approach achieves better cost-effectiveness than existing data dissemination schemes.**

I. Introduction

Disruption Tolerant Networks (DTNs) [10] consist of mobile nodes which contact each other opportunistically. Due to the low node density and unpredictable node mobility, only intermittent network connectivity exists in DTNs, and the subsequent difficulty of maintaining end-to-end communication links advances "carry-and-forward" approaches for data delivery. More specifically, node mobility is exploited to let mobile nodes physically carry data as relays, and forward data opportunistically upon contact with others. The key problem is hence how to design appropriate relay selection strategy.

Data dissemination is useful in many applications in DTNs, including event notification, network status updates and content publishing. In most of the existing schemes, data is disseminated to all the nodes in the network. These schemes are essentially "network-centric" and ignore the satisfaction of user interest. Data is forwarded to many nodes not interested in the data, and a lot of network resources are therefore wasted. To deal with this problem, data recipients should be appropriately identified based on their interests in the data.

In this paper, we propose the concept of *user-centric* data dissemination in DTNs, which considers satisfying user interests and forwards data only to the nodes that are interested in the data. Such nodes are called "*interesters*" in the rest of this paper. We aim at maximizing the cumulative dissemination cost-effectiveness over all the data items in the network, by designing appropriate relay selection strategy.

This work was supported in part by the US National Science Foundation (NSF) under grant number CNS-0721479, and by Network Science CTA under grant W911NF-09-2-0053.

The major difficulty of user-centric data dissemination in DTNs is that the interesters of a data item are generally unknown *a priori* at the data source, because it is difficult for the data source to have knowledge about the interests of other nodes in the network. Such uncertainty of data recipients is different from unicast [4], [15], [13] or multicast [14] in which the destinations are fixed and pre-known, and makes relay selection for user-centric data dissemination challenging.

Our main idea to overcome the aforementioned difficulty is to let a node estimate the interest of another node in a data item as probability, based on which we propose user-centric data dissemination from the social network perspective. We exploit node *centrality* in DTNs to consider the social contact patterns and interests of mobile nodes simultaneously for effective relay selection. While centrality in Social Network Analysis (SNA) generally represents the capability of a node facilitating the social communication among other nodes [11], we expand the centrality concept to analytically represent the capability of a node to forward data to its interesters. Our detailed contributions are as follows:

- We propose a general probabilistic framework for user-centric data dissemination in DTNs.
- We propose a novel approach to relay selection based on the node centrality values, and ensure that data items are effectively disseminated based on their popularity.
- We provide theoretical insight on the cost-effectiveness of data dissemination.

In our approach, the effectiveness of relay selection depends on the scope of network information maintained at individual nodes. By theoretical analysis, we provide lower bound to the cost-effectiveness of data dissemination, and investigate the tradeoff between this cost-effectiveness and the overhead of maintaining network information. We analytically show that, when such information is maintained in larger scopes, the maintenance overhead and the effectiveness of relay selection increase at similar rates. Hence, network designer has full flexibility to determine the appropriate relay selection strategy to balance the two aspects based on the application requirements.

The rest of this paper is organized as follows. Section II reviews the existing work. Section III provides an overview about problem formulation and the basic idea. Sections IV and V describe our probabilistic framework and our user-centric data dissemination in detail. Theoretical analysis is provided in Section VI. Section VII conducts performance evaluations based on realistic traces, and Section VIII concludes the paper.

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

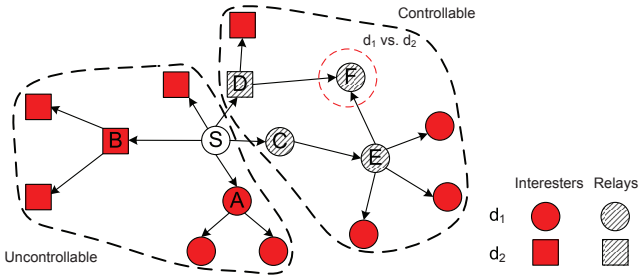


Fig. 1. User-Centric Data Dissemination

II. Related Work

The research on relay selection strategy in DTNs originates from Epidemic routing [26], and some later work [23], [4] studied this problem based on the prediction of node mobility [12]. Recently, social-based data forwarding schemes have also been proposed [8], [15], [14], [22], based on various social network concepts including centrality and communities.

Flooding-based data dissemination is implemented in [19], and theoretical analysis has been conducted on its stochastic regimes [3] or aging rules [17]. Later data dissemination schemes are closely related to publish-subscribe systems [27], [21] with simplified models of user interest. In [20], [2], data items are grouped into pre-defined channels, and data dissemination is based on users' subscriptions to channels. This model implicitly assumes the consistency of user interests over all the data items in the same channel, and simplifies relay selection by using data dissemination history in the past as prior knowledge. Comparatively, we propose a general framework for data dissemination, based on a probabilistic model of user interest without assuming any data inter-dependency.

Social-based data dissemination in DTNs has also been studied. [2] disseminates data by defining community-based relay selection policies. SocialCast [7] investigates the "homophily" phenomenon [24], and assumes that users with common interests contact each other more often. Being orthogonal with the existing work, our approach investigates the social contact pattern of nodes as more accurate and predictable abstraction of node mobility, and exploits centrality which analytically represents such contact pattern for relay selection.

III. Overview

A. Problem Formulation

We formulate user-centric data dissemination as follows:

Problem 1: **User-Centric Data Dissemination**
For n data items originated at source nodes $S_1, S_2, ..., S_n$ with time constraints $T_1, T_2, ..., T_n$, how to disseminate them to maximize the cumulative cost-effectiveness $\sum_{i=1}^{n} \frac{N_I^i(T_i)}{N_R^i(T_i)}$?

In this formulation, $N_R^i(t)$ is the number of selected relays for data d_i at time t, and $N_I^i(t)$ is the estimation at time t on the number of interesters that will receive d_i by time T_i. Each relay estimates this cost-effectiveness ratio based on its own knowledge at time t, and such estimation may vary at different relay. Each relay only has limited buffer space.

B. The Big Picture

Figure 1 illustrates the big picture of user-centric data dissemination. Two data items d_1 and d_2 are disseminated by node S, which is the initial relay. Each node decides whether to be interested in the data when it contacts another node carrying the data, and hence data dissemination is split into two parts, i.e., the uncontrollable part and the controllable part, according to where an interester receives the data from.

In the uncontrollable part, data is disseminated among the interesters autonomically without help of additional relays. In Figure 1, interesters A and B, after having received data from S, carry and forward the data to other interesters. Since the interest of a node is less related with its capability of contacting other interesters, the cost-effectiveness of uncontrollable data dissemination is opportunistic and unreliable.

Comparatively, in the controllable part, relays C and D are intentionally selected among the non-interester nodes, according to their capabilities of forwarding data to interesters. Each relay is selected by another existing relay rather than an interester, so as to ensure that each selected relay is aware of other existing relays, and hence has a local estimate of the cost-effectiveness ratio of data dissemination. The cost-effectiveness of controllable data dissemination can be ensured when relays are appropriately selected.

In this paper, we focus on maximizing the cost-effectiveness of controllable data dissemination. Our approach consists of two parts: (i) *relay selection* for each data d_i to maximize the cost-effectiveness $\frac{N_I^i(t)}{N_R^i(t)}$ of disseminating d_i, (ii) *data item selection* on a relay if its buffer size is not enough to carry all the data items, to maximize the cumulative dissemination cost-effectiveness. In Figure 1, when node F is selected as the relay for data d_1 and d_2 simultaneously, F decides which data to carry if its buffer is only enough to carry one of them.

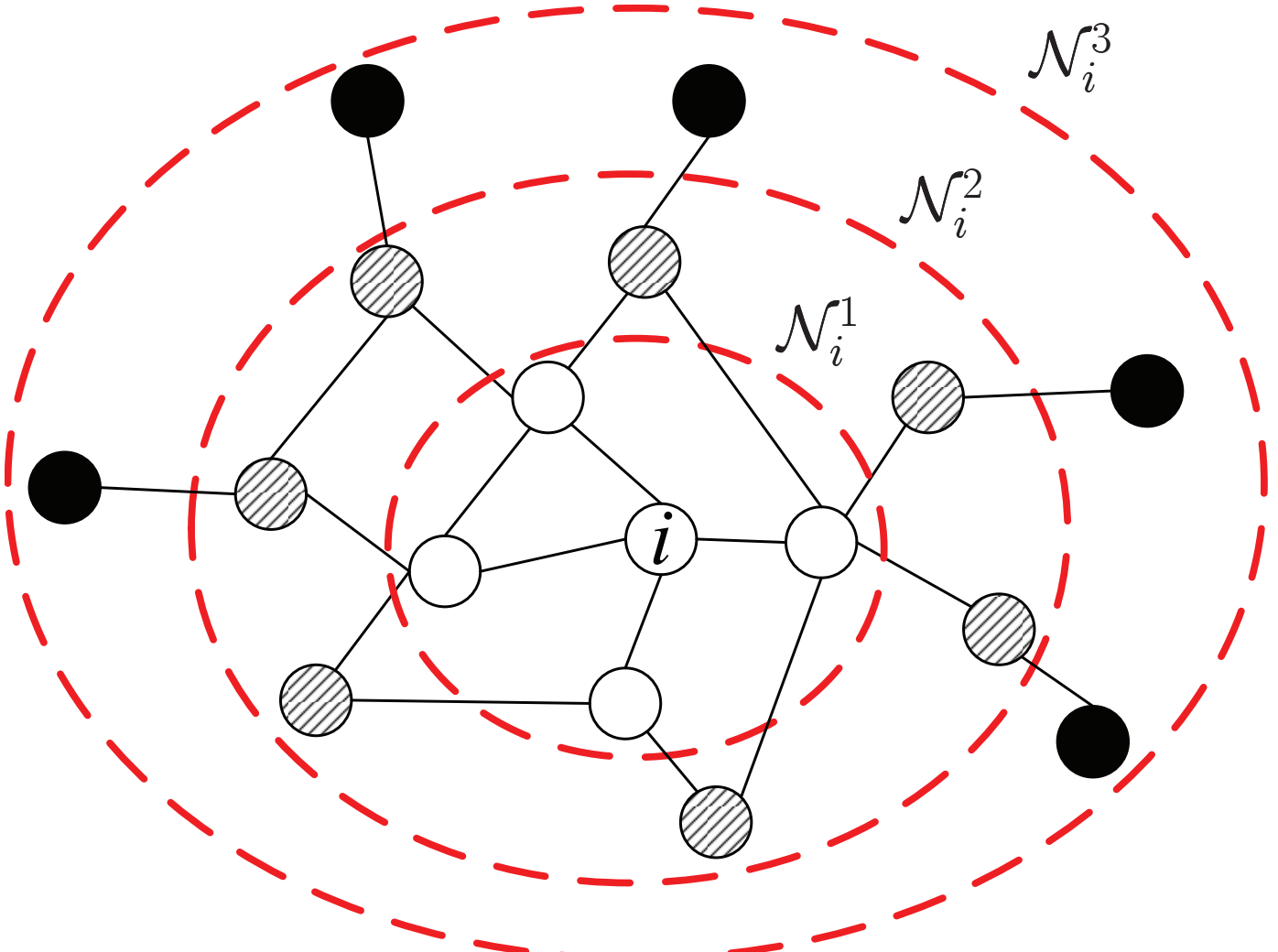
IV. Models

A. Network Modeling

Node contacts are described by the network *contact graph* $G(V, E)$, where the stochastic contact process between a node pair $i, j \in V$ is modeled as an edge $e_{ij} \in E$. Being similar with [1], [17], we consider the pairwise node inter-contact time as exponentially distributed. The contacts between nodes i and j then form a homogeneous Poisson process with the contact rate λ_{ij}, which is calculated in a time-average manner at real-time. Currently, although [18] suggested the aggregate distribution of node inter-contact time to be a mixture of power-law and exponential distributions, there is still no agreement on the pairwise distribution of node inter-contact time, and our modeling has been experimentally validated by [6], [14], [28] to fit well to realistic DTN traces.

B. User Interest & Data Modeling

Our model estimates the interest of a node in a data item as probability, which is calculated from user interest profile and data description. Thus, our model allows a user to have various interests in different data items in the same category.


$\mathcal{N}_i^3$
$\mathcal{N}_i^2$
$\mathcal{N}_i^1$
i

as relays. Therefore, we normalize $C_{ij}(T_k - t)$ with the path length L_{ij} in terms of hop count

$$C_{ij}(T_k - t) = p_{ij}(T_k - t)/L_{ij}. \tag{6}$$

Node i maintains the best opportunistic path with the largest $C_{ij}(T_k - t)$ for each node $j \in \mathcal{N}_i^r$. The information about opportunistic path is disseminated and updated in a per-hop manner among nodes in $\mathcal{N}_i^r$ via their mutual contacts.

B. Relay Selection

*1) **Basic Rule**:* A node i is only selected as the relay by another relay j for data d_k at time t, if selecting i increases the cost-effectiveness ratio $\frac{N_I^k(t)}{N_R^k(t)}$ estimated at relay j. That is,

$$\frac{N_I^k(t) + C_i^{(k)}(t)}{N_R^k(t) + 1} \geq \frac{N_I^k(t)}{N_R^k(t)}, \tag{7}$$

which can be equivalently written as $C_i^{(k)}(t) \geq \frac{N_I^k(t)}{N_R^k(t)}$.

In this case, a new relay always has better capability of disseminating data to interesters than the existing relays. Similar methodology has also been used in [9] for effective data forwarding in DTNs. While selecting more relays always facilitates data dissemination in DTNs due to its opportunistic nature, our approach maximizes the dissemination cost-effectiveness by only selecting the best nodes as relays.

The data source S_k, as the initial relay, sets $N_I^k(0) = C_{S_k}^{(k)}(0)$ and $N_R^k(0) = 1$. Whenever S_k contacts another node i, it determines whether node i should be a relay according to Eq. (7). If so, $N_I^k(t)$ and $N_R^k(t)$ at both S_k and i are updated according to Eq. (8). Note that $N_I^k(t)$ estimates the number of interesters that receive data d_k by the time constraint T_k. Hence, it will only be updated when a new relay is selected, and will not be updated when a relay contacts an interester.

$$\begin{cases} N_I^k(t) \leftarrow N_I^k(t) + C_i^{(k)}(t) \\ N_R^k(t) \leftarrow N_R^k(t) + 1 \end{cases} \tag{8}$$

*2) **Using Multi-hop Centrality**:* Relay selection only based on the local network knowledge may not be optimal. This can be illustrated in Figure 4, where the value besides a node j indicates $p_j^{(k)}$ in Eq. (2), and the value on the dashed edge between nodes i and j indicates $C_{ij}(T_k - t)$ in Eq. (3). The local and multi-hop node centrality values are also listed in the figure. If local centrality is used, A will not select B as the relay due to Eq. (7). However, such decision will fail to select D with high centrality as the relay.

To ensure optimal relay selection, multi-hop centrality should be used instead. As shown in Figure 4, the 2-hop centrality of node B increases a lot due to the high value of local centrality of node D. Hence, B will be selected as relay. Afterwards, $\frac{N_I^{(k)}(t)}{N_R^{(k)}(t)}$ at B will be updated to $\frac{1.2+1.286}{2} = 1.243$, and furthermore enables selecting D as relay. Note that, due to the normalization on $C_{ij}(T_k - t)$ in Eq. (6), we can safely update $N_I^k(t)$ and $N_R^k(t)$ by applying Eq. (8).

The exploitation of multi-hop centrality for relay selection inevitably requires nodes to maintain network information in a larger scope, and leads to higher maintenance overhead. Such tradeoff will be analyzed in Section VI-B.

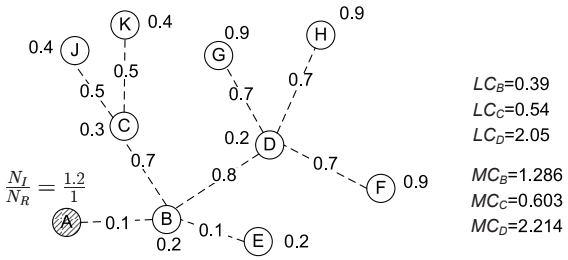


Fig. 4. Illustration of relay selections

C. Data Item Selection

A node i determines which data items to carry, if it is selected as the relay for data items $d_1, d_2, ..., d_m$ but its buffer B_i is not large enough to carry all of them. Data items are selected according to their contribution to increase the cumulative dissemination cost-effectiveness $\sum_{k=1}^{m} \frac{N_I^k(t)}{N_R^k(t)}$. Such selection is formulated as a knapsack problem as follows.

$$\max \sum_{k=1}^{m} w_k x_k \quad \text{s.t.} \sum_{k=1}^{m} s_k x_k \leq B_i, \tag{9}$$

where $x_k \in [0, 1]$ is the indicator variable indicating whether data d_k is carried by node i, and s_k is the size of data d_k. w_k is the contribution of data d_k which is defined as

$$w_k = \frac{N_I^{(k)}(t) + C_i^{(k)}(t)}{N_R^{(k)}(t) + 1} - \frac{N_I^{(k)}(t)}{N_R^{(k)}(t)} = \frac{C_i^{(k)}(t) - \frac{N_I^k(t)}{N_R^k(t)}}{N_R^k(t) + 1}. \tag{10}$$

The solution to Eq. (9) prefers data items with higher popularity, because $C_i^{(k)}(t)$ for popular data items are generally higher. Nevertheless, such preference diminishes when $N_R^k(t)$ increases. Hence, we also ensure that data items with lower popularity can be fairly disseminated, when the popular data items have already been carried by a number of relays.

VI. Analysis

In this section, we provide theoretical insight on our user-centric data dissemination. More specifically, we provide the lower bound of the data dissemination cost-effectiveness, and analytically investigate the tradeoff between the effectiveness of relay selection and the overhead of maintaining network information.

We analyze the process of disseminating a single data item, and the cumulative cost-effectiveness of data dissemination is maximized by the data item selection in Section V-C. For simplicity, we omit the data item index k in the relevant notations. We use the notation $\mathcal{I}(t)$ to indicate the global set of interesters having received the data at time t, and $\mathcal{R}(t)$ to indicate the global set of selected relays at time t. We assume that there are N nodes in the network described by the contact graph $G = (V, E)$, and that the data is generated at time 0.

A. Lower Bound of Dissemination Cost-Effectiveness

We first analyze the lower bounds of $|\mathcal{I}(t)|$ and $|\mathcal{R}(t)|$ when local centrality is used for relay selection, and obviously these bounds also hold when multi-hop centrality is used. Note

that $|\mathcal{I}(t)|$ and $|\mathcal{R}(t)|$ calculated at the global scope may be different from $N_I(t)$ and $N_R(t)$ estimated at individual relays.

As a prerequisite, Lemma 2 first provides a lower bound on the ratio $\frac{N_I(t)}{N_R(t)}$ maintained at an arbitrary relay.

***Lemma** 2: At any time $t \leq T$, we have*

$$\frac{N_I(t)}{N_R(t)} \geq (1 - e^{-s_G(T-t)}) \cdot p_{\min}, \tag{11}$$

where $p_{\min} = \min_{i \in V} p_i$, and

$$s_G = \min_{A \subset V} \frac{\sum_{i \in A, j \in \mathcal{N}_i^1} \lambda_{ij}}{|A|}. \tag{12}$$

The proof of Lemma 2 can be found in Appendix A, and Lemma 3 shows the lower bound on the renewal intervals of $|\mathcal{R}(t)|$.

***Lemma** 3: Let $|\mathcal{R}(t_0)| = k$ for $t_0 \leq T$, and $T_R^{(k+1)}$ be the time needed to select the $(k+1)$-th relay, then we have*

$$\mathbb{P}(T_R^{(k+1)} \leq t) \leq (1 - e^{-(N-k)c_G t}) \cdot (1 - (1 - e^{-s_G t}) \cdot p_{min}), \tag{13}$$

where $1 \leq k \leq N/2$, and c_G is defined as

$$c_G = \max_{A \subset V} \frac{\sum_{i \in A, j \in V \setminus A} \lambda_{ij}}{\max\{|A|, |V \setminus A|\}}. \tag{14}$$

The proof of Lemma 3 can be found in Appendix B. From Lemma 3, we can see that our approach selects more relays for popular data items, which are disseminated to more interesters. Such property is analytically described in the following lemma:

***Lemma** 4: Let T_I be the time needed for the selected relays in $\mathcal{R}(t_0)$ to contact an interester, we have*

$$\mathbb{P}(T_I \leq t) \geq (1 - e^{-k h_G t}) \cdot p_{\min}, \tag{15}$$

where $k = |\mathcal{R}(t_0)| \in [1, N/2]$, and h_G is defined as

$$h_G = \min_{A \subset V} \frac{\sum_{i \in A, j \in V \setminus A} \lambda_{ij}}{\min\{A, V \setminus A\}}.$$

Proof: Considering that interesters can only receive the data from the selected relays, the proof of Lemma 4 is similar with the first part of the proof of Lemma 3, such that a lower bound on the cumulative contact rate λ is given as $\lambda \leq k h_G$ for $1 \leq k \leq N/2$. ∎

From Lemmas 3 and 4, the lower bound on data dissemination cost-effectiveness is described in Theorem 1.

***Theorem** 1: The probability for the dissemination cost-effectiveness $\frac{|\mathcal{I}(t_0)|}{|\mathcal{R}(t_0)|}$ at time $t_0 \leq T$ to increase after time $t \leq T - t_0$ is bounded as*

$$\mathbb{P}\left(\frac{|\mathcal{I}(t_0+t)|}{|\mathcal{R}(t_0+t)|} \geq \frac{|\mathcal{I}(t_0)|}{|\mathcal{R}(t_0)|}\right) \geq (1 - e^{-k h_G t}) \cdot e^{-(N-k)c_G t} \cdot p_{\min}, \tag{16}$$

where $k = |\mathcal{R}(t_0)| \in [1, N/2]$.

Proof: The dissemination cost-effectiveness increases if the existing relays contact new interesters within time t and no new relay is selected. As a result,

$$\mathbb{P}\left(\frac{|\mathcal{I}(t_0+t)|}{|\mathcal{R}(t_0+t)|} \geq \frac{|\mathcal{I}(t_0)|}{|\mathcal{R}(t_0)|}\right) \geq \mathbb{P}(t \geq T_I) \cdot \mathbb{P}(t \leq T_R^{(k+1)}).$$

Theorem 1 is therefore an immediate result from Eqs. (13) and (15). ∎

Particularly, $p_{\min}$ in Theorem 1 can be derived from the cumulative distribution function (CDF) $F_p(x)$ of user interest probability as in Eq. (17).

$$\mathbb{P}(p_{\min} \leq x) = F_{\min}(x) = \prod_{i=1}^{N} \mathbb{P}(p_i \leq x) = (F_p(x))^N. \tag{17}$$

Theorem 1 has the following implications:

1) The cost-effectiveness of disseminating a data item is proportional to the contact capability of relays and the data popularity. It is generally more cost-effective to disseminated popular data items in the network.
2) Eq. (16) shows that, the lower bound of dissemination cost-effectiveness increases exponentially with t. This indicates that our data dissemination approach is sensitive to short time constraints, and will perform much better when the time constraint increases.
3) The lower bound in Eq. (16) varies at different t_0 because $k = |\mathcal{R}(t_0)|$. The bound becomes higher when t_0 increases, which means that our approach tends to achieve higher cost-effectiveness when the time elapses.

B. Tradeoff

As described in Section V-B, maintaining the network information in a larger scope increases the effectiveness of relay selection, at the cost of higher maintenance overhead. In this section, we show analytical results on such tradeoff.

***Lemma** 5: When node centrality is calculated in the R-hop range, where R is the network diameter, the relay selection following Eq. (7) is always optimal. For any relay s with locally estimated $\frac{N_I(t)}{N_R(t)}$, when it contacts node i at time t,*

- *If $C_i(t) < \frac{N_I(t)}{N_R(t)}$, selecting any node $j \in \mathcal{N}_i^1$ as the relay will decrease $\frac{N_I(t)}{N_R(t)}$.*
- *If $C_i(t) \geq \frac{N_I(t)}{N_R(t)}$, there exists $j \in \mathcal{N}_i^1$, such that selecting node j as the relay increases $\frac{N_I(t)}{N_R(t)}$.*

The proof of Lemma 5 can be found in Appendix C. Comparatively, when the network information is maintained in a r-hop range ($r < R$), Theorem 2 gives a upper bound on the probability of non-optimal relay selection.

***Theorem** 2: For a relay s with estimated $\frac{N_I(t)}{N_R(t)}$ at time t, $\forall i \in \mathcal{N}_s^1$ and $j \in \mathcal{N}_i^{r+1} \setminus \mathcal{N}_i^r$, if $C_i(t) < \frac{N_I(t)}{N_R(t)}$ we have*

$$\mathbb{P}\left(C_i(t) + p_j C_{ij}(T-t) > \frac{N_I(t)}{N_R(t)}\right) \leq \frac{(C_{\max})^{r+1} \cdot \mathbb{E}p_j}{(r+1) \cdot (\frac{N_I(t)}{N_R(t)} - C_i(t))}, \tag{18}$$

where $C_{\max} = 1 - e^{-\lambda_{\max}(T-t)}$, and $\lambda_{\max} = \max_{i,j \in V} \lambda_{ij}$.

Proof: According to the calculation of multi-hop centrality defined in Eq. (6), for a $(r+1)$-hop opportunistic path between node i and j, we have

$$C_{ij}(T-t) \leq \frac{(1 - e^{-\lambda_{\max}(T-t)})^{r+1}}{r+1}.$$

As a result, Eq. (18) can be proved by applying Markov's inequality, such that

$$\mathbb{P}\left(p_j C_{ij}(T-t) > \frac{N_I(t)}{N_R(t)} - C_i(t)\right) \leq \frac{\mathbb{E}(p_j C_{ij}(T-t))}{\frac{N_I(t)}{N_R(t)} - C_i(t)}.$$

■

According to Theorem 2, the probability of non-optimal relay selection is negatively proportional to the scope r of network information being maintained, and is also determined by the node contact frequency and data popularity.

However, the maintenance of network information in DTNs is expensive. Lemma 6 shows that the opportunistic path in DTNs cannot be maintained in an iterative manner. Instead, to calculate its r-hop centrality value, a node i has to maintain the complete opportunistic paths to all the nodes in $\mathcal{N}_i^r$.

***Lemma** 6: There does not exist a function $f(\lambda, T)$, such that for any opportunistic path $P_{AB} = (A, N_1, ..., N_{r-1}, B)$ with edge weights $\{\lambda_1, \lambda_2, ..., \lambda_r\}$,*

$$p_{AB}(T) = p_{AN_{r-1}}(T) \otimes f(\lambda_r, T),$$

where $\otimes$ can be any arbitrary arithmetic operation.

The proof of Lemma 6 can be found in Appendix D. Theorem 3 then gives an upper bound on the increasing rate of the overhead of maintaining network information when r increases.

***Theorem** 3: For any node i, the overhead of maintaining network information within the r-hop range is $\Omega(r \cdot c^r)$, where c is a graph-dependent invariant.*

Proof: Let k be the minimum node degree in $G = (V, E)$. Without loss of generality, we assume that G is connected, so with large probability we have $k \geq 2$. For $\forall i \in V$, since $\mathcal{N}_i^r$ is a subgraph of G, we have $|\mathcal{N}_i^1| \leq k$ and $|\mathcal{N}_i^r \setminus \mathcal{N}_i^{r-1}| \leq k(k-1)^{r-1}$. Therefore, let $n_r = |\mathcal{N}_i^r|$, we have

$$n_r \geq \sum_{s=0}^{r-1} k(k-1)^s = \frac{k}{k-2} \cdot ((k-1)^r - 1).$$

The theorem therefore holds because k is an invariant only depending on G, and for $\forall j \in \mathcal{N}_i^r \setminus \mathcal{N}_i^{r-1}$, the length of the opportunistic path between node i and j is at least r. ■

From Theorems 2 and 3 we conclude that, when r increases, the optimal probability of relay selection and the overhead of maintaining network information increase at similar rates. Hence, the network designers have full flexibility to balance between the relay selection effectiveness and maintenance overhead according to the specific application requirements.

C. Global Optimality

In data dissemination, an individual node only estimates the dissemination cost-effectiveness ratio $\frac{N_I(t)}{N_R(t)}$ based on its own network information, and consequently such estimated ratio may be different at nodes. For example, when relay selections happen in the temporal sequence shown in Figure 5(b), the update process and the final values of $\frac{N_I(t)}{N_R(t)}$ at different nodes are shown in Figure 5(b) and 5(c), respectively.

Due to the lack of end-to-end network connectivity in DTNs, such difference essentially makes it difficult to guarantee the global optimality for relay selection, which means that every relay selection increases the global value of $\frac{N_I(t)}{N_R(t)}$. More specifically, a relay selection which increases the local cost-effectiveness ratio may not necessarily increase the global ratio. For example in Figure 5, when node B contacts E and selects E as the relay because $1.7 > \frac{5.0}{3}$, it actually

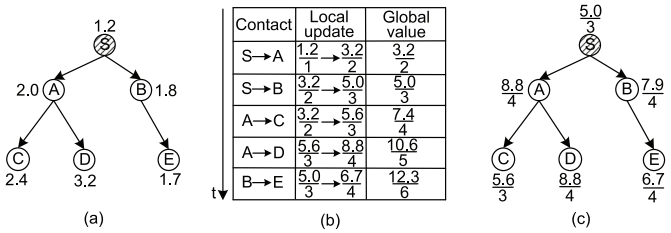


Fig. 5. Maintenance of $\frac{N_I(t)}{N_R(t)}$: (a) node centrality values, (b) update process of $\frac{N_I(t)}{N_R(t)}$, (c) final values of $\frac{N_I(t)}{N_R(t)}$ at different nodes

reduces the global cost-effectiveness ratio from $\frac{10.6}{5} = 2.12$ to $\frac{12.3}{6} = 2.05$. The main reason is that B may be out of contact with S and A when it contacts E, and therefore B is not aware of which relays A has selected. Nevertheless, although it is generally hard to achieve the global optimality for relay selection, our approach ensures its local optimality by exploiting multi-hop centrality, as stated in Lemma 5.

VII. Performance Evaluation

In this section, we compare the performance of our scheme with other data dissemination schemes listed below.

1) **Flooding**, in which all the non-interesters are relays.
2) **Random flooding**, in which each non-interester has a fixed probability to be randomly selected as a relay.
3) **ContentPlace** [2], which uses distributed k-clique method [16] to detect social community structures, and uses the Most Frequently Visited (MFV) policy for determining data utilities.
4) **SocialCast** [7], in which we implemented the Kalman-filter method for co-location prediction.

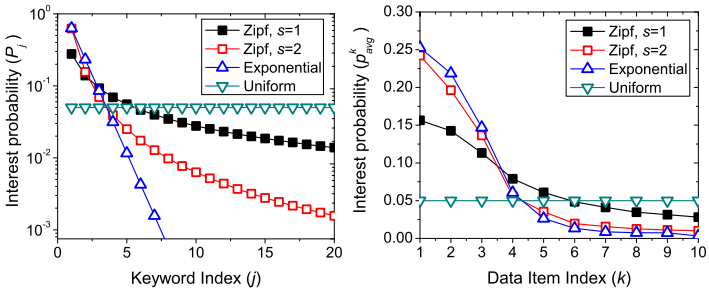
In this section, N_I indicates the average number of interesters having received the data, and N_R indicates the average number of selected relays. Both quantities are kept globally up-to-date. In flooding-based schemes, data items are randomly selected at a relay with limited buffer, and we keep N_R in random flooding at the same level as in our approach to evaluate the effectiveness of random relay selection. All the experiments are run multiple times with randomly generated data for statistical convergence.

TABLE I
Trace summary

Trace	*Infocom06*	*MIT Reality*
Duration (days)	4	246
Granularity (secs)	120	300
No. of devices	78	97
No. of internal contacts	182,951	54,667
Pairwise contact rate (per day)	1.70	0.024

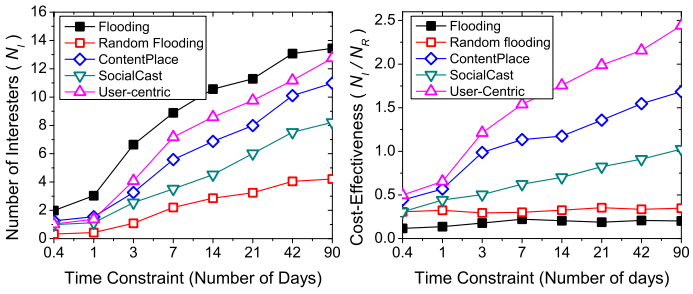
A. Simulation Setup

Our evaluations are conducted on two realistic DTN traces, which record contacts among users carrying Bluetooth-enabled mobile devices. These devices record contacts by periodically detecting their peers nearby. The traces cover various types of corporate environments and have various experiment periods. They are summarized in Table I.



(a) (b)

Fig. 6. User interest probabilities: (a) in keywords, (b) in data items



(a) (b)

Fig. 7. Data dissemination with different time constraints: (a) The number of interesters (N_I), (b) The dissemination cost-effectiveness (N_I/N_R)

In all the experiments, a node updates its contact rates with other nodes in real-time, based on the up-to-date contact counts since the network starts. The first half of the trace is used as the warm-up period for the nodes to accumulate necessary network information. All the data items are generated and disseminated during the second half of the trace.

*1) **User interest**:* We generate user interest profiles based on a keyword space $\mathcal{K}$ with size $M = 20$, and assume that keyword $k_j \in \mathcal{K}$ is the j-th popular keyword in the network.

According to Definition 1, the user interest probability in each keyword k_j is randomly drawn from a normal distribution with P_j as the mean value. We exploit various distributions for generating P_j of different keywords:

- Zipf distribution with exponent s: $P_j = \frac{1/j^s}{\sum_{i=1}^{M} 1/i^s}$.
- Exponential distribution: $P_j = \frac{e^{-j}}{\sum_{i=1}^{M} e^{-i}}$.
- Uniform distribution: $P_j = 1/M$.

P_j with different distributions are shown in Figure 6(a).

*2) **Data item**:* There are 5 data items to be disseminated in the network. The data resources and time of data origination are randomly generated. The sizes of data items are uniformly generated in the range $[100kB, 200kB]$, and the node buffer sizes are uniformly generated in $[200kB, B_{\max}]$, where the value of $B_{\max}$ varies to achieve different buffer constraints.

Each data item d_k is described by 5 keywords with equal weights. To ensure that the data items have different popularity, the keyword indices of data item d_k are $\{k, ..., k+4\}$. The user interest probability in data d_k can hence be calculated according to Eq. (1), and the average interest probability p^k_{avg} in data d_k over all the nodes in the network is illustrated in Figure 6(b). When the mean value P_j is generated exponentially, most of the user interests concentrate on the popular data items. For Zipf distributions, such concentration increases with exponent s. Therefore, Figure 6(b) actually represents different data interest patterns of mobile users in DTNs. In our simulations, a node determines whether it is interested in an encountered data item, by locally performing a Bernoulli trial with its interest probability in the data.

B. Dissemination Cost-effectiveness

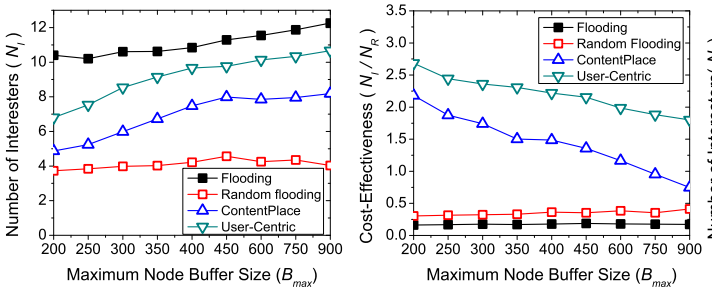
In this section, we evaluate the data dissemination performance and the cost-effectiveness of our approach on the *MIT Reality* trace. The simulation settings in different experiments vary from the basic setting, where the time constraint $T = 21$ days, and $B_{\max} = 500$kB. We generate user interest probabilities following the Zipf distribution with exponent $s = 2$, and maintain network information within the 3-hop range for centrality calculation. For ContentPlace, we randomly group the 5 data items into 2 channels, and let each node have a fixed probability $p = 0.2$ to be interested in each channel.

*1) **Different Time Constraints**:* The experiment results are shown in Figure 7. In general, data items are disseminated to more interesters when the time constraint T is larger. In Figure 7(a), when T is long (> 42 days), N_I of flooding approaches the maximum value $N \cdot p_{avg}$, and N_I of random flooding is much lower due to random relay selection. Because of the centrality-based relay selection, N_I of our approach only degrades 15%-20% from flooding, and performs much better than ContentPlace and SocialCast. SocialCast delivers data to fewer interesters than ContentPlace because the homophily phenomenon may not hold in the traces we use.

Comparatively, Figure 7(b) shows that our approach achieves the highest cost-effectiveness of data dissemination indicated by the ratio N_I/N_R. This ratio is also proportional to T, because a relay has higher chance to contact more interesters when T is large. As shown in Figure 7(b), our approach achieves 30% higher cost-effectiveness than ContentPlace, and 50% higher than SocialCast. Note that the cost-effectiveness of flooding-based schemes remains stable at all cases.

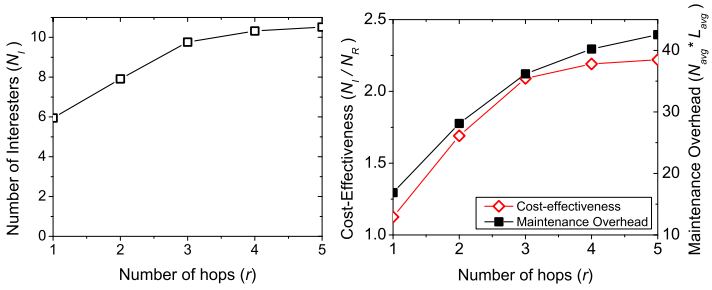
*2) **Different Buffer Constraints**:* The experiment results are shown in Figure 8. We did not include SocialCast because it assumes infinite node buffer size. We vary $B_{\max}$ from 200kB to 900kB, so that each relay can at least carry one data item, but not all of them. In Figure 8(a), N_I increases when relays have larger buffer to carry data items. When $B_{\max}$ is increased from 200kB to 900kB, N_I of our approach increases by 56%, and its difference from that of flooding correspondingly decreases by 50%. With various buffer constraints, our approach keeps 25%-30% performance advantage over ContentPlace.

The dissemination cost-effectiveness surprisingly decreases when $B_{\max}$ increases, as shown in Figure 8(b). Considering the increase of N_I in Figure 8(a), the main reason is that N_R increases when larger buffer size is available. Nevertheless, when $B_{\max}$ increases from 200kB to 900kB, the cost-effectiveness of our approach only decreases by 25%, which



(a) (b)

Fig. 8. Data dissemination with different buffer constraints: (a) The number of interesters (N_I), (b) The dissemination cost-effectiveness (N_I/N_R)



(a) (b)

Fig. 9. Data dissemination with different rages for maintaining network information: (a) The number of interesters (N_I), (b) The dissemination cost-effectiveness (N_I/N_R) and overhead ($N_{avg} \cdot L_{avg}$)

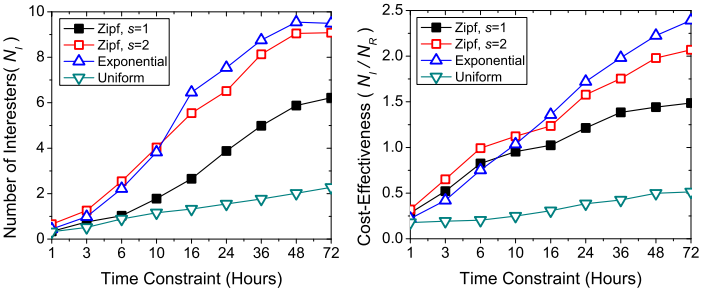
is much smaller than the 65% decrease of ContentPlace. The cost-effectiveness of flooding-based schemes remains stable due to the random strategy for data item selection.

*3) **Scope of Maintaining Network Information**:* The results on various scopes of maintaining network information are shown in Figure 9. In Figure 9(a), N_I increases when network information is maintained in a larger scope, and this increase is larger for smaller r. When r increases from 1 to 3, N_I increases by 64%. When r furthermore increases from 3 to 5, such benefit is reduced to 7.7%.

Figure 9(b) shows the tradeoff between the cost-effectiveness of data dissemination and the maintenance overhead. The overhead is measured by $N_{avg} \cdot L_{avg}$, where N_{avg} is the average number of nodes whose information is maintained, and L_{avg} is the average length of maintained opportunistic paths. It is shown that they both increase at similar rates when r increases, which is consistent with our theoretical analysis in Section VI-B.

C. Distribution of User Interest

In this section, we evaluate the cost-effectiveness of our approach under different distributions of user interest on the *Infocom06* trace. The experiment results are shown in Figure 10. Since our approach prefers to disseminate popular data items, the values of both N_I and N_I/N_R mainly depend on the user interest probability in popular data items. N_I and N_I/N_R have the lowest values when user interest is uniformly distributed. When the exponent s of Zipf distribution increases,



(a) (b)

Fig. 10. Data dissemination with different distributions of user interests: (a) the number of interesters (N_I), (b) the cost-effectiveness (N_I/N_R)

the user interest concentrate more on popular data items, and our approach therefore performs 25% better.

The values of N_I and N_R in cases of exponential distribution and Zipf distribution with $s = 2$ are worth special attention. As shown in Figure 10, our approach performs best in case of exponential distribution when T is longer than 10 hours. This is mainly because user interest concentrates more on popular data items when it is exponentially distributed, as shown Figure 6(b). When T is longer, popular data items have higher chances to be disseminated to more interesters, and hence improves the dissemination cost-effectiveness. In contrast, when T is short, the preference on popular data items reduces the chances of other data items to be disseminated and affects the dissemination cost-effectiveness.

VIII. Conclusions

In this paper, we proposed a novel social-based approach to user-centric data dissemination in DTNs, which considers user interests and improves data dissemination cost-effectiveness. We propose a probabilistic model of user interest, and expand the centrality concept for effective relay selection by considering the social contact patterns and interests of mobile nodes simultaneously. In the future, we will conduct more detailed performance evaluation of our approach, especially in cases where nodes dynamically join and leave the network and network data is randomly being updated. Future research can also benefit from our work by following the concept of user-centric data dissemination to further investigate the roles and impacts of user interests in DTNs.

References

[1] A. Balasubramanian, B. Levine, and A. Venkataramani. DTN Routing as a Resource Allocation Problem. In *Proceedings of SIGCOMM*, pages 373–384, 2007.

[2] C. Boldrini, M. Conti, and A. Passarella. ContentPlace: social-aware data dissemination in opportunistic networks. In *Proceedings of MSWiM*, pages 203–210, 2008.

[3] C. Boldrini, M. Conti, and A. Passarella. Modelling data dissemination in opportunistic networks. In *Proceedings of ACM Workshop on Challenged Networks (CHANTS)*, pages 89–96, 2008.

[4] J. Burgess, B. Gallagher, D. Jensen, and B. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networks. *Proc. INFOCOM*, 2006.

[5] A. Chaintreau, A. Mtibaa, L. Massoulie, and C. Diot. The diameter of opportunistic mobile networks. In *Proceedings of the 2007 ACM CoNEXT conference*. ACM New York, NY, USA, 2007.

[6] V. Conan, J. Leguay, and T. Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. In *Proceedings of the 1st Int'l Conf on Autonomic Computing and Communication Systems*, 2007.
[7] P. Costa, C. Mascolo, M. Musolesi, and G. Picco. Socially Aware Routing for Publish-Subscribe in Delay-Tolerant Mobile Ad Hoc Networks. *IEEE Journal on Selected Areas in Communications*, 26(5):748–760, 2008.
[8] E. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant MANETs. *Proc. MobiHoc*, 2007.
[9] V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot. Delegation Forwarding. *Proc. MobiHoc*, 2008.
[10] K. Fall. A delay-tolerant network architecture for challenged internets. *Proc. SIGCOMM*, pages 27–34, 2003.
[11] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.
[12] W. Gao and G. Cao. Fine-Grained Mobility Characterization: steady and transient state behaviors. In *Proceedings of MobiHoc*, pages 61–70. ACM, 2010.
[13] W. Gao and G. Cao. On Exploiting Transient Contact Patterns for Data Forwarding in Delay Tolerant Networks. In *Proceedings of ICNP*, pages 193–202, 2010.
[14] W. Gao, Q. Li, B. Zhao, and G. Cao. Multicasting in delay tolerant networks: a social network perspective. In *Proceedings of MobiHoc*, pages 299–308, 2009.
[15] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. *Proc. MobiHoc*, pages 241–250, 2008.
[16] P. Hui, E. Yoneki, S. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. *Proc. MobiArch*, 2007.
[17] S. Ioannidis, A. Chaintreau, and L. Massoulie. Optimal and scalable distribution of content updates over a mobile social network. *Proc. INFOCOM*, 2009.
[18] T. Karagiannis, J.-Y. Boudec, and M. Vojnovic. Power law and exponential decay of inter contact times between mobile devices. *Proc. MobiCom*, pages 183–194, 2007.
[19] G. Karlsson, V. Lenders, and M. May. Delay-Tolerant Broadcasting. In *Proceedings of ACM Workshop on Challenged Networks (CHANTS)*, pages 197–204, 2006.
[20] V. Lenders, G. Karlsson, and M. May. Wireless ad hoc podcasting. In *Proceedings of SECON*, pages 273–283, 2007.
[21] F. Li and J. Wu. Mops: Providing content-based service in disruption-tolerant networks. In *Proceedings of Int'l Conf. on Distributed Computing Systems (ICDCS)*, pages 526–533, 2009.
[22] Q. Li, S. Zhu, and G. Cao. Routing in Socially Selfish Delay Tolerant Networks. *Proc. INFOCOM*, 2010.
[23] A. Lindgren, A. Doria, and O. Schelen. Probabilistic routing in intermittently connected networks. *ACM SIGMOBILE CCR*, 7(3):19–20, 2003.
[24] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
[25] S. M. Ross. *Introduction to probability models*. Academic Press, 2006.
[26] A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. *Technical Report CS-200006, Duke University*, 2000.
[27] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft. A socio-aware overlay for publish/subscribe communication in delay tolerant networks. *Proc. MSWiM*, pages 225–234, 2007.
[28] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. M. Ni. Recognizing Exponential Inter-Contact Time in VANETs. In *Proceedings of INFOCOM*, 2010.

APPENDIX

A. Proof of Lemma 2

According to Definition 3, the centrality value $C_i(t)$ of node i decreases when t increases. Therefore, based on the relay selection strategy described in Section V-B, we have

$$\frac{N_I(t)}{N_R(t)} \geq \frac{\sum_{i\in\mathcal{R}(t)} C_i(t)}{|\mathcal{R}(t)|} \geq \frac{\sum_{i\in\mathcal{R}(t),j\in\mathcal{N}_i^1}(1-e^{-\lambda_{ij}(T-t)})\cdot p_{\min}}{|\mathcal{R}(t)|}.$$

Since

$$2-e^{-\lambda_1 t}-e^{-\lambda_2 t}\geq 1-e^{-(\lambda_1+\lambda_2)t},$$

for $\forall\lambda_1,\lambda_2\geq 0$ and $t\geq 0$, according to the definition of s_G in Eq. (12) we have

$$\frac{N_I(t)}{N_R(t)} \geq \frac{\sum_{i\in\mathcal{R}(t)}(1-e^{-s_G(T-t)})\cdot p_{\min}}{|\mathcal{R}(t)|} = (1-e^{-s_G(T-t)})\cdot p_{\min}.$$

B. Proof of Lemma 3

The probability $\mathbb{P}(T_R^{(k+1)}\leq t)$ is equal to $P_1\cdot P_2$, where P_1 is the probability that at least one node in $V\setminus\mathcal{R}(t_0)$ is contacted by nodes in $\mathcal{R}(t_0)$, and P_2 is the probability that at least one node being contacted is selected as the relay. We prove this lemma by analyzing the two probabilities separately.

First, the time until at least one node in $V\setminus\mathcal{R}(t_0)$ is contacted by nodes in $\mathcal{R}(t_0)$ is exponentially distributed with $\lambda=\sum_{i\in\mathcal{R}(t_0),j\in V\setminus\mathcal{R}(t_0)}\lambda_{ij}$. Therefore, according to the definition of c_G in Eq. (14) we have

$$P_1 = 1-e^{-\lambda t}\leq 1-e^{-(N-k)c_G t}, \tag{19}$$

where $1\leq k\leq N/2$.

Second, according to the relay selection strategy in Section V-B, we have

$$P_2\leq\mathbb{P}\bigg(C_i(t)\geq\frac{N_I(t)}{N_R(t)}\bigg).$$

From Lemma 2, along with the memoryless nature of Poisson processes and the assumption that the user centrality values are uniformly distributed, we have

$$P_2\leq(1-\frac{N_I(t)}{N_R(t)})\leq(1-(1-e^{-s_G t})\cdot p_{min}). \tag{20}$$

The lemma hence follows by combining Eqs. (19) and (20).

C. Proof of Lemma 5

The correctness of two cases are proved respectively.

Case 1: According to Eq. (2), for any $j\in\mathcal{N}_i^1$, its contribution to i's centrality value is $p_jC_{ij}(T-t)$. Therefore, if node i is selected as the relay, node j's contribution to $\frac{N_I(t)}{N_R(t)}$ is no larger than $\sum_{k\in\mathcal{N}_j^{R-1}}p_kC_{ik}(T-t)$. Since $\mathcal{N}_i=V\setminus\{i\}\supset\mathcal{N}_j^{R-1}$, generally we have $\sum_{k\in\mathcal{N}_j^{R-1}}p_kC_{ik}(T-t)\leq C_i(t)$, and therefore the first case of the lemma follows.

Case 2: This case can be proved by contradiction. If the selection of any $j\in\mathcal{N}_i^1$ as the relay decreases $\frac{N_I(t)}{N_R(t)}$, this comes to be the same with Case 1, and can be concluded as $C_i(t)<\frac{N_I(t)}{N_R(t)}$, which contradicts the prerequisite of Case 2.

D. Proof of Lemma 6

The difficulty of calculating $p_{AB}(T)$ in an iterative manner mainly comes from the properties of the coefficients $C_k^{(r)}$ in Eq. (5). When a new edge (N_{r-1},B) with weight λ_r is added into a path AN_{r-1}, such coefficients are modified as

$$C_k^{(r)}=\begin{cases}C_k^{(r-1)}\cdot\frac{\lambda_k}{\lambda_r-\lambda_k}, & k\neq r\\ \prod_{s=1}^{r-1}\frac{\lambda_s}{\lambda_s-\lambda_r}, & k=r\end{cases} \tag{21}$$

Our observations from Eq. (21) are two-fold. First, each coefficient $C_k^{(r)}$ $(k\neq r)$ is updated by multiplying a distinct value $\frac{\lambda_k}{\lambda_{r+1}-\lambda_k}$. Second, the calculation of $C_r^{(r)}$ involves all the edge weights $\lambda_1,...,\lambda_{r-1}$. Both of the two observations makes it impossible to calculate $p_{AB}(T)$ solely from $p_{AN_{r-1}}(T)$ and λ_r.