On Exploiting Hotspot and Entropy for Data Forwarding in Delay Tolerant Networks

Peiyuan Yuan†, Huadong Ma‡, Shaojie Tang†
†Department of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
‡Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA
peiyuan@htu.cn, mhd@bupt.edu.cn, tangshaojie@gmail.com

Abstract—Performance of data forwarding in Delay Tolerant Networks (DTNs) benefits considerably if one can make use of human mobility in terms of social structures. However, it is difficult and time-consuming to calculate the centrality and similarity of nodes by using solutions for traditional social networks, this is mainly because of the transient node contact and the intermittently connected environment. In this work, we are interested in the following question: Can we explore some other stable social attributes to quantify the centrality and similarity of nodes? Taking GPS traces of human walks from the real world, we find that there exist two known phenomena. One is public hotspot, the other is personal hotspot. Motivated by this observation, we present Hoten (hotspot and entropy), a novel routing metric to improve routing performance in DTNs. First, we use the relative entropy between the public hotspots and the personal hotspots to compute the centrality of nodes. Then we utilize the inverse symmetrized entropy of the personal hotspots between two nodes to compute the similarity between them. Third, we exploit the entropy of personal hotspots of a node to estimate its personality. Besides, we propose a method to ascertain the optimized size of hotspot. Finally, we compare our routing strategy with other state-of-the-art routing schemes through extensive trace-driven simulations, the results show that Hoten largely outperforms other solutions, especially in terms of combined overhead/packet delivery ratio and the average number of hops per message.

I. INTRODUCTION

Delay tolerant networks [1] have been applied into many applications, such as the interplanetary internet [2], vehicle ad-hoc networks [3] and content delivery system [4] [5] etc. In these scenarios, routing is one of the most challenging problems, due to the lack of an end-to-end path between source and destination. Obviously, this new feature leads to a considerable performance degradation for conventional wireless routing protocols such as AODV or DSR, as they are originally designed for stable network topology. Hence, new data forwarding algorithms should be designed for DTNs.

In the past few years, several DTNs routing schemes (e.g., epidemic [6] and data MULEs [7]) have been proposed to deal with this problem. Among them, epidemic scheme seems to be a feasible solution to forward messages from a sender to a potential receiver when nothing is known about the mobility of nodes (in the rest of this paper, without loss of generality, we use the terms “people” and “node” interchangeably), since it tries to send each message over all possible paths in the network. Apparently, epidemic scheme has the merits of low mean delivery delay (MDD) and high packet delivery ratio (PDR), at the same time, it also incurs a high price of system resources because of the large amount of redundant copies.

This deficiency has motivated researchers to develop other novel data forwarding algorithms, which make a better tradeoff between packet delivery ratio and the consumption of system resources by taking advantage of different contexts (e.g., [8] [9] [10] [11] [12]). For these schemes, the routing performance depends heavily on the contexts they used to estimate the better relay nodes to the destination. Furthermore, most existing schemes do not take the social structures into account. Whereas, human walks gradually play a critical role in the network performance [13] with the recent popularization of personal hand-held mobile devices, since devices may lose connection when people move around. Hence, the social structures of humans walks acquired by mobility characterization techniques are of great importance on designing data forwarding metrics. Therefore, we focus on how to integrate social structures into the data forwarding algorithms in DTNs. It is a critical while challenging task especially in an intermittently connected environment.

Recently, there are a few works that explicitly consider some social structures in DTNs routing (e.g., [14] [15]). However, none of them fully exploit social structures extracted from real human traces. For instance, some existing schemes only exploit virtual community structure to identify the friendship among nodes and use centrality or similarity of nodes to estimate the utilities of such nodes as potential relays. The reason behind these schemes is that the underlying social structure is more stable compared with the network topology, and hence can be used for better relay selections. By analyzing GPS traces of human walks from the real world, we confirm that there also exist two known phenomena as the indications in [16] [17]. One is that people always move around a set of well popular locations which are called public hotspots, instead of purely random motions. The other is that each people shows preference for some particular locations which are called personal hotspots in this paper. We believe that both kinds of hotspots are more stable than underlying social structure mentioned above as public hotspots are formed by superimposing personal hotspots together and personal hotspots/habits are stable over time and across situations [33]. Moreover, the evaluation for centrality and similarity of nodes in existing schemes takes traditional methods of social networks or ego networks. We argue that these approaches are difficult and
time-consuming, due to the transient node contact and the intermittently connected environment.

Taking all above issues into account, in this paper, we exploit hotspots to design a new routing metric. In specific, we investigate the following two kinds of hotspots. (i) The public hotspots: this implies that there exists a bigger chance to meet the destination in these locations than other places. Hence, we also need to address how to identify those nodes which have a higher centrality than others. (ii) The personal hotspots: this implies that if we can deliver a message to one of the most $k$ popular personal hotspots of the destination, the message will be quickly received by the destination. As such, we have to answer the problem of how to estimate the similarity between a potential relay and the destination. Besides, since each person has his/her own personality we still need to incorporate this factor into the data forwarding process.

In this paper, we present a novel metric, called Hoten, to address these challenges. We first use the relative entropy between the public hotspots and the personal hotspots to evaluate the centrality of nodes. Then we utilize the inverse symmetrized entropy of the personal hotspots of two nodes to weigh the similarity between them. Third, different from the related works, we integrate a new factor, into our Hoten metric and exploit the entropy of personal hotspots to estimate node personality. Besides, we propose a method to ascertain the optimized size of hotspot. Our main contributions can be summarized as follows:

- We introduce the entropy theory into opportunistic forwarding. Rather than exchange neighbor’s adjacency matrix [14] or count the number of times a node acts as a relay for other nodes on all the shortest delay paths [15], we exploit hotspot and entropy to quantify the centrality and similarity of nodes, which guarantees Hoten is concise and low time complexity.
- We take personality of nodes into account, which makes Hoten prediction more accurate than the existing works since each people has his/her own personal habit.
- We exploit the values of Hurst parameter to explore the optimized size of hotspot and try to reduce the influence of the number of hotspots on the bursty dispersion of traces.
- We conduct extensive experiments to compare Hoten and several state-of-the-art works based on five real DTNs traces, experiment results show that Hoten largely outperforms other solutions, especially in terms of combined overhead/packet delivery ratio and the average number of hops per message.

We organize the remainder of this paper as follows. Section II reviews the related work. Section III presents the process for identifying hotspots. Section IV describes our approaches to evaluate centrality, similarity and personality metrics. In Section V, we make a performance evaluation. Finally, we conclude our paper and discuss some future research issues in Section VI.

II. RELATED WORK

It is challenging to deliver messages through disconnected parts of the network. In the past, several schemes have been proposed to solve this issue. On the basis of contexts they used, we classify them into the following two categories: (i) data forwarding without social structures, (ii) data forwarding with social structures.

- **A. Data Forwarding without Social Structures**

**Periodic information based:** Several schemes utilize the periodic information inherent to some mobility patterns to route message in DTNs. S. Merugu et al. [18] assumed that the global knowledge of the mobility of nodes could be predicted over a finite or indefinite time scale, due to the periodicity in node movement. They delivered messages over a space-time routing table with knowledge of when the relay would be encountered. Likewise, S. Jain et al. [19] took a modified Dijkstra algorithm to compute the shortest path between the source and the destination by assuming that the time when a message will arrive at a particular node must be predicted. They presented several schemes and evaluated their performance based on different knowledge oracles which they acquired from the network. On exploiting past traces of buses to predict future behavior, the authors of [20] presented MaxProp, which shows better performance than protocols that depend on proactive knowledge. Besides, the authors of [21] proposed a source routing in DTNs, they took the expected minimum delay as forwarding metric based on that the motion of real objects was repetitive but non-deterministic.

**Opportunity based:** The deficiency of epidemic scheme has motivated researchers to develop opportunity based data forwarding algorithms (e.g. [8] [9] [10] [11] [12]). Most of them make a better tradeoff between packet delivery ratio and the consumption of system resources by taking advantage of different contexts. For these schemes, the routing performance depends heavily on the contexts they used to estimate the better relay nodes to the destination. For instance, A. Lindgren et al. [8] presented PROPHET, a probabilistic routing protocol for DTNs. They exploited past histories of encounters to predict the probability of future encounters. Similar to [8], CAR (context aware routing) was proposed in [9], which exploited Kalman filters and the context information such as the changing rate of neighbors of a node and its current energy level to predict the delivery probability. In addition, J. Leguay et al. [10] presented MobySpace, a high-dimensional Euclidean space constructed by the past motion patterns of nodes.

- **B. Data Forwarding with Social Structures**

Note that most aforementioned schemes do not take the social structures into account. However, with the recent popularization of personal hand-held mobile devices, human walks gradually play a critical role in the network performance, since devices may fail to connect with each other when people move around. Recently, a few works attempt to uncover the underlying stable network structure in real traces by using
social networks analysis technology [22]. For example, SimBet [14] exploited betweenness centrality and social similarity of ego networks [23] to differentiate nodes. Messages will be forwarded to such nodes which have relatively big SimBet values to increase the probability of finding better relays to the final destination. P. Hui et al. [15] proposed BUBBLE, which combined node centrality and community structure to make forwarding decisions. They assumed that each node had a global rank across the whole system and a local rank within its local community. When a message is out of the community of the destination, it is forwarded to the node with a high global rank, when the message enters into the range of the destination community, it is delivered to the node with a high local rank in that community.

III. IDENTIFYING HOTSPOT

We present the experimental datasets used in the paper in Section III.A. In Section III.B, we give a detailed presentation about the hotspots division and weight computation. In Section III.C, we discuss the bursty dispersion of hotspots.

A. Experimental Data-sets

We use the following five real DTNs data-sets gathered by [17] [24] over almost two years (from 2006-08-26 to 2008-04-18), referred to as KAIST, NCSU, New York City, Orlando and North Carolina State Fair. The characteristics of these datasets such as intra-inter-contact distribution have been explored in several studies (e.g., [17] [24]) and applied into different scenarios (e.g., message deletion mechanism in [25]). Interestingly, by analyzing these traces, we find that they cover a rich diversity of environments ranging from well connected area (State fair) to quite sparse situation (New York City). We summarize the main features of the five data-sets in Table I.

![Table](image)

| Site          | No. of traces | volunteers | start date       | end date       |
|---------------|---------------|------------|------------------|----------------|
| KAIST         | 92            | 34         | 2006-09-26       | 2007-10-03     |
| NCSU          | 55            | 20         | 2006-08-26       | 2006-11-16     |
| New York      | 39            | 10         | 2006-10-23       | 2008-04-18     |
| Orlando       | 41            | 18         | 2006-11-19       | 2008-01-09     |
| State fair    | 19            | 18         | 2006-10-24       | 2007-10-21     |

B. Hotspots Division and Weight Computation

In this subsection, we first clarify some terms used in this paper such as GPS log, GPS trace, stay point, and then present our solutions to hotspots division and weight computation.

GPS log and GPS trace: The data collected by the GPS devices carried by participants are form of GPS log, which is a sequence of three-tuples (Timestamp, X-coordinate, Y-coordinate). As depicted in Fig 1 on a two dimensional plane, we can connect these three-tuples into a GPS trace according to their time sequences.

Stay point: A stay point $P$ denotes a physical location where a participant stays more than a threshold. There are two categories of stay points. The first means that a participant remains stationary for a while exceeding the threshold, and the second denotes that a person wanders around within a certain small spatial region for a time period (in [17], the default values of the threshold and the radius of small region are 30 seconds and 5 meters, respectively). The authors of [26] proposed an algorithm for stay point detection.

Hotspot division: Different from the virtual community structure used in [15], a hotspot is defined to be a physical region with an area of $d$ by $d$. Since different values of $d$ result in different number of hotspots, which in turn influences the self-similarity of traces [16], we need to find the optimized value of $d$. In this paper, we exploit the values of Hurst parameter to explore the influence of the number of hotspots on the bursty dispersion of traces. We take the maximum of Hurst parameter to ascertain the optimized size of hotspots. Mathematically, let $D$ denote the set of $d$, let $H$ denote the Hurst parameter of traces and function $f$ denote the mapping from $D$ to $H$, we have

$$f : D \rightarrow H$$

and there exists $d_{\text{optimized}} \in D$, such that $h_{\text{max}} = \max(f)$, where $h_{\text{max}} \in H$. We use iterative process to observe the effect of selecting different values of $d$ on the values of $h$, which are estimated by using the aggregated variance method. Fig.2 shows the results. Among them, we set $d_{\text{optimized}} = f^{-1}(h_{\text{max}})$, where $f^{-1}$ is the inverse function of $f$.

Weight computation: As stated above, we divide the five scenarios by non-overlapping $d$ by $d$ squares, each square indicates one hotspot. We use the weight of each hotspot to denote its popularity. The larger the weight value is, the more popular the hotspot is. There are several methods to estimate the weight of hotspots [27], we here take a simple but efficient solution, called count process. We count the number of stay points within each hotspot and then compute the weight of each hotspot by normalizing the sampled count.

Let $K$ denote the number of hotspots in the network, let $n_i$ denote the number of stay points in hotspot $i$ and $w_i$ denote the weight of that hotspot, we have:

$$w_i = \frac{n_i}{\sum_{i=1}^{K} n_i} \quad (1)$$

Similarly, let $n^{j}_{\text{personal}}$ denote the number of the $i$th person’s stay points in $j$th hotspot and $w^{j}_{\text{personal}}$ denote the weight of that hotspot influenced by the $i$th person, we have:

$$w^{j}_{\text{personal}} = \frac{n^{j}_{\text{personal}}}{\sum_{j=1}^{K} n^{j}_{\text{personal}}} \quad (2)$$



![Fig. 1. GPS trace and stay point](image)
An online approach to identify public hotspots: Clearly, it is not possible for a DTNs user to acquire a global knowledge (e.g., the public hotspots). We here exploit the aggregated personal hotspots to identify the public hotspots of system. That is, each node carries a hotspot matrix $H_{N \times K}$ with initial elements $h_{i,j} = w_{personal, i}$ and 0 otherwise (where we take node $i$ as an example and $N$ is the number of nodes). When two nodes meet up, they exchange their own $H$ and update the values of $H$ using information from their neighbor. After that, they estimate the weight $w_j$ of the $j$th public hotspot by summing the elements in $V_j$, where $V_j$ is the $j$th column of $H$. Finally, they normalize each $w_j$ and use them to compute the betweenness centrality (please refer to Section IV.D).

C. Bursty Dispersion of Hotspots

The phenomenon of bursty dispersion (i.e., self-similarity) of hotspot implies that people always tend to swarm near to a few very popular locations, which means we can only use few particular locations to identify the individual trace. Hence, the size of control packet will be reduced considerably.

Bursty dispersion of public hotspots: The bursty of public hotspots means that popular locations become more popular as individual bursty traces are superimposed together. Fig. 2 portrays the distribution of public hotspots of the five sites, which shows a clearly bursty pattern and coincides with the theory of preferential attachment proposed in [28]. Bursty dispersion of personal hotspots: The bursty of personal hotspots implies that individual user spends most time in some special locations consciously or unconsciously. On average, only about of 1-14.7% hotspots are visited by each participant as shown in Table II (we here only consider the top $k$ hotspots whose sum of weights is bigger than or equal to 0.9, which has at least 90% confidence guarantee). Fig. 4 depicts the distribution of personal hotspots of State fair (in each scenario, we randomly choose two people as samples. The rest scenarios show the similar features, we omit them here due to the space limitation), which also shows a bursty pattern as that of the public hotspots. Notice that there exist two phenomena in these figures (Fig. 2 and Fig. 4). One is that different people may have different preferred locations, i.e., different personal habits, the second is that the bursty degree of personal hotspots is fiercer than that of the public hotspots. Both the two phenomena inspire us to estimate the centrality, personality and similarity of people.

|       | KAIST | NCSU | New York | Orlando | State fair |
|-------|-------|------|----------|---------|------------|
| Ratio | 0.01  | 0.018| 0.07     | 0.01    | 0.147      |

IV. IMPLEMENTING HOTSPOTS INTO HOTEN

We present our solution in this section. In Section IV.A, we explore the centrality of a node. We analyze the similarity between nodes in Section IV.B. In Section IV.C, we present personality. We finally exploit entropy and hotspot to design Hoten metric in Section IV.D.

A. Centrality

Node centrality reflects the relative importance of nodes in the network (i.e., how popular a person is within a social
network). The more important the person is, the bigger the chance to meet other people is. Freeman [29] [30] proposed three most widely used methods to estimate centrality, called degree, closeness and betweenness measures.

**Degree centrality:** Degree centrality is measured as the number of one-hop neighbors of a given node \( i \), which reflects the direct relationship between the node \( i \) and its neighbors. A node with higher degree centrality means it can directly contact with more other nodes. Degree centrality of node \( i \) is counted as:

\[
C_D^i = \sum_{j=1, j\neq i}^{N} p_{ij}
\]

where \( N \) is the number of nodes in the network and \( p_{ij} = 1 \) if node \( j \) is one of neighbors of node \( i \).

It is not easy to compute degree centrality in DTNs as the number of direct contacts that involve a node is varying from time to time. One optional method is that we can set a time window and count the number of neighbors of nodes within it. However, we can not ascertain how the optimal size of the time window is.

**Closeness centrality:** Closeness centrality shows the “closeness” of a node to all other reachable nodes. Freeman took the reciprocal of the average geodesic length \( d(i, j) \) (i.e., the shortest path from node \( i \) to all other reachable nodes) to measure it [30]. Closeness centrality of a node also reflects the node’s freedom from the network, which is calculated as:

\[
C_C^i = \frac{N - 1}{\sum_{j=1, j\neq i}^{N} d(i, j)}
\]

In DTNs, it is hard to work out the geodesic length \( d(i, j) \), due to the unguaranteed end-to-end path between node \( i \) and node \( j \).

**Betweenness centrality:** Betweenness centrality reflects the controlling capability of a node to other nodes, which measures the extent to which a node falls on the shortest path between two other nodes. The higher the betweenness centrality of a node is, the bigger the ability it has to facilitate communication to other nodes within the network is. Betweenness centrality of a node \( i \) is computed as:

\[
C_B^i = \sum_{j=1}^{N} \sum_{k=1}^{N} g_{jk}(i) / g_{jk}
\]

where \( g_{jk} \) is the total number of shortest path between node \( j \) and node \( k \), and \( g_{jk}(i) \) is the number of those paths that include node \( i \).

Obviously, the betweenness centrality is difficult to be evaluated with the increasing number of nodes, due to the high time complexity. Besides, similar to that of closeness centrality, it is more difficult to work it out in DTNs. For example, the authors of [14] used an adjacency matrix \( A \) to represent node contacts, which has elements \( A_{ij} = 1 \) if there has been at least one contact between node \( i \) and \( j \) at any past time and \( A_{ij} = 0 \) otherwise. The betweenness centrality thus can be estimated as:

\[
A^2[1 - A]_{i,j}
\]

Apparently, the matrixes will get more and more identical with the contacts aggregation as shown in Fig.3 (when two nodes meet up each other, they swap their own neighbor list to update the matrix). As a consequence, heterogeneity of the nodes can not be well reflected, which in turn will impair the network performance (please refer to the Section V). On the other hand, if we use the sliding time window as the authors of [15] did, we have to ascertain the optimal size of time window, whereas, answering this question is non-trivial as well. We discuss how to exploit hotspot to solve this problem in the Section IV.D.

**B. Similarity**

Similarity reflects the associations between nodes in the network. Sociologists have observed the phenomenon long before, which is called “clustering” in physics, that if two people have one or more common friends, they can also be friends with high probability.

The number of common neighbors between nodes has an important influence on the dissemination speed of messages in DTNs. When the neighbors of nodes contact each other frequently, the message diffusion process can be expected to take faster than when the association between nodes is weaker. That is, nodes having a stronger association with a given node are good relay candidates for message diffusion to that node. The generalized method exploits some contexts to estimate the degree of association. For example, the authors of [51] took advantage of the mail list to match the relationship between people in real world. The authors of [32] reflected the associations between bloggers by analyzing the linking objects existing in the large number of blogs.

However, it is difficult to count the number of common neighbors (or others such as the common mail list items [51] or common linking objects [32]), due to the same reasons mentioned in the subsection IV.A.

![Fig. 3. Adjacency matrixes for node 0 and 1 at different time instants.](attachment:image.png)
C. Personality

Personality reflects the unique characteristic (or behavior) of a given person. The famous psychologist Allport, G. W [33] defined the personality as “a general neuropsychic structure unique to the individual with the capacity to render many stimuli functionally equivalent and to instigate and guide consistent (equivalent) forms of adaptive and stylistic behavior.” Allport, G. W suggested that personality characteristics are relatively stable over time and are stable across situations.

The personality characteristics mainly include tendentiousness, complexity, uniqueness, positiveness and stability etc. We believe that the personal hotspots at least can reflect the tendentiousness, uniqueness and stability of personality as shown in Fig.2 and Fig.4, since public hotspots are superimposed by personal hotspots, and moreover, each person has his/her own personal habit and the personal habit is stable once it is forming. Hence, it is necessary and significative to exploit personal hotspots to make comparisons across people. In the next subsection, we introduce how to integrate it into the Hoten metric.

D. Hoten

In this subsection, we use the entropy theory to compute betweenness centrality, similarity and personality of nodes as it denotes the degree of disorder or randomness in a system, that is, the bigger the entropy value is, the more disordered the system would be. More specifically, we utilize the relative entropy between the public hotspots and the personal hotspots to evaluate betweenness centrality of a node, we then exploit the inverse symmetrized entropy of the personal hotspots between two nodes to compute the similarity between them, we finally use the entropy of personal hotspots of a node to estimate its personality.

Let random variable $X_i$ denote the distribution of personal hotspots of node $i$, let random variable $Y$ denote the distribution of public hotspots, let $p(x_i^j)$ and $p(y_j)$ denote the weights of $j$th personal hotspot and public hotspot, respectively, we have:

$$p(x_i^j) = w_{\text{personal},i}^j$$

$$p(y_j) = w_j$$  \hspace{1cm} (7)\hspace{1cm} (8)

**Betweenness centrality computation:** Relative entropy (also called Kullback-Leibler divergence) can be used to differentiate the divergence between two random distributions. If the relative entropy value equals to zero, we call that the two random variables have the same distribution (i.e., if $X_i$ has the same distribution as $Y$, we call that node $i$ has the highest betweenness centrality in the network). Let $C_b^i$ denote the betweenness centrality of node $i$, we have

$$C_b^i = \left( \sum_{j=1}^{K} p(x_i^j) \log(p(x_i^j)/p(y_j)) \right)^{-1}$$  \hspace{1cm} (9)

Replace equations (7) and (8) into equation (9), we have

$$C_b^i = \left( \sum_{j=1}^{K} w_{\text{personal},i}^j \log(w_{\text{personal},i}^j/w_j) \right)^{-1}$$  \hspace{1cm} (10)

Compared with equations (5) and (6), it is clear to see that our solution is more concise and has a low time complexity $\Theta(K)$, which is only related to the number of hotspots and independent of the number of nodes in the network. **Similarity computation:** The relative entropy does not keep symmetry, i.e., the relative entropy of $X_i$ over $X_j$ does not equal to that of $X_j$ over $X_i$. We here use inverse symmetrized entropy to estimate the similarity $Sim(i,j)$ between node $i$ and node $j$, we have

$$Sim(i,j) = \left( Sim(i/j) + Sim(j/i) \right)^{-1}$$  \hspace{1cm} (11)

where, $Sim(i/j) = \sum_{i=1}^{K} w_{\text{personal},i}^j \log(w_{\text{personal},i}^j/w_{\text{personal},i}^j)$ and $Sim(j/i) = \sum_{j=1}^{K} w_{\text{personal},j}^i \log(w_{\text{personal},j}^i/w_{\text{personal},j}^i)$.

**Personality computation:** Let $Per_i$ denote the personality of node $i$, according to the definition of entropy, we have

$$Per_i = -\frac{K}{\sum_{l=1}^{K} w_{\text{personal},l}^i \log(w_{\text{personal},l}^i)}$$  \hspace{1cm} (12)

To make the above equations hardness, we set $w_j = \delta$ and $w_{\text{personal},i}^j = \delta$ if they equal to zero, where $\delta$ is a constant. **Hoten metric:** The Hoten metric is a value between 0 and 1 and is calculated by integrating the above three components. Hence, the question of selecting the best relay for the message becomes a multi-objective optimization problem. This is achieved by linear weighting method. Let $BetUtil_i$, $SimUtil_i(n_d)$ and $PerUtil_i$ denote the betweenness utility, similarity utility and personality utility of node $i$ for delivering a message to destination node $n_d$ when meeting up node $j$, respectively. Exploiting the normalized relative weights of these attributes, we have

$$BetUtil_i = \frac{C_b^i}{C_b^i + C_b^i}$$  \hspace{1cm} (13)

$$SimUtil_i(n_d) = \frac{Sim(i,n_d)}{Sim(i,n_d) + Sim(j,n_d)}$$  \hspace{1cm} (14)

$$PerUtil_i = \frac{Per_i}{Per_i + Per_j}$$  \hspace{1cm} (15)

According to the linear weighting method, we have

$$Hoten_i(n_d) = \alpha BetUtil_i + \beta SimUtil_i(n_d) + \gamma PerUtil_i$$  \hspace{1cm} (16)

where $\alpha$, $\beta$, and $\gamma$ are system parameters and $\alpha + \beta + \gamma = 1$.

**Hoten routing:** We outline the Hoten routing in Algorithm 1, which presents the communication process between node $i$ and node $j$. Take node $i$ as an example. When meeting up node $j$, for any message $m$ that $i$ carries, if its destination
\( m_d \) is node \( j \), node \( i \) delivers it to node \( j \) and removes it from \( i \)'s message queue. Otherwise, if node \( j \) does not hold this message, the two nodes swap their own Hoten utility. If \( \text{Hoten}_i(m_d) \) is smaller than \( \text{Hoten}_j(m_d) \), node \( i \) delivers the message to node \( j \) and removes \( m \) from its buffer space, i.e., Hoten takes a single copy scheme.

**Algorithm 1** Hoten Algorithm, pseudo-code of node \( i \)

1. upon meeting up node \( j \) do
2. for any message \( m \) in \( i \)'s queue do
3. if \( m_d == j \) then
4. deliverMsg\((m)\)
5. remove\((m)\)
6. else if \( m \notin j \) then
7. \( i \leftarrow \text{Hoten}_j(m_d) \)
8. isForwarding\((m)\) \{make forwarding decision\}
9. end if
10. end for
11. isForwarding\((m)\)
12. if \( \text{Hoten}_i(m_d) < \text{Hoten}_j(m_d) \) then
13. forwardingMsg\((m)\)
14. remove\((m)\)
15. end if

V. PERFORMANCE EVALUATION

We take Epidemic routing as a baseline to compare Hoten performance to SimBet metric.

A. Simulation Setup

We exploit the aforementioned five real DTNs traces to test the premise of routing based on social structures. Since each trace has different run times, for the four DTNs traces (KAIST, NCSU, New York City and Orlando), we use the minimum runtime (15000s) in KAIST as the baseline, for State fair traces, the runtime is set to 6000s. Thus, we get 92,32,26,39 and 19 traces respectively, which are slightly smaller than the original numbers (please refer to Table I). The value of \( \delta \) is set to 0.000001. The parameters for the Hoten metric in equation (16) are all set to 1/3, which assigns an equal importance to them. According to Table II the ratio of \( k/K \) is set to 15%. The nodal transmission range is set to 250m, a typical value of WiFi. In addition, all nodes are both sources and destinations, i.e., each node sends a single message for all other nodes.

B. Performance Criteria

We evaluate the performances of the three routing protocols taking the following criteria into account.

**Cumulative packet delivery ratio (CPDR):** This criterion represents the delivery performance in the network in terms of the number of successfully received messages over that the sent messages. We evaluate the delivery performance of the three metrics under different message TTLs.

**Mean delivery delay:** Although delay is tolerant in DTNs, a low end-to-end delay is still desirable as long delay means more system resources are occupied for longer.

**Average ratio of infected nodes:** We use this criterion to quantify the overhead in the network. Since Hoten and SimBet only take a single copy scheme, both are expected to perform similarly in this respect.

**Average number of hops per message:** The least hop does not mean the shortest delay in DTNs, since it is measured as the successful forwarding times of a message until the destination receives it. Whereas, we still try to minimize this criterion due to the two aspects of considerations, the channel interference and battery power. Minimizing the number of hops also reduces the probability of channel interference and the consumption of battery power.

C. Cumulative Packet Delivery Ratio

Fig.6 illustrates the performance of packet delivery ratio under different message TTLs. Epidemic has the highest CPDR than the other two as expected. Compared to SimBet, it is clear to see that Hoten improves the packet delivery ratio. The reason behind this is that Hoten exploits hotspot and entropy to estimate the centrality and similarity of nodes and takes nodal personality into account, which make Hoten prediction more accurate than that of SimBet. An exception happens at State Fair, where the CPDR of SimBet is better than that of Hoten, this is mainly because that the adjacent matrixes among nodes will quickly become identical in well connected scenarios, hence, the heterogeneity of the nodes can not be well reflected, which in turn makes SimBet tend to flood the messages.

D. Mean Delivery Delay

Looking at the mean delivery delay (Fig.7). Epidemic has a better MDD performance than Hoten and SimBet, also as expected. Compared to SimBet, Hoten indeed prolongs the MDD. Whereas, we notice that Hoten improves the CPDR metric in most scenarios (Fig.6), hence, we conjecture that the extra delay may be caused by those messages which could be dropped under SimBet, but now are able to be delivered to their destinations under Hoten.

E. Average Ratio of Infected Nodes

Fig.8 clarifies that both Hoten and SimBet achieve the better performance in terms of average ratio of infected nodes as expected. It is obvious to see that Epidemic almost infects every nodes in the network.

F. Average Number of Hops per Message

Fig.9 illustrates the average number of hops per message. Hoten metric obviously outperforms the other two schemes. For example, at KAIST, the average number of hops per message achieved by Hoten is near to 7, whereas Epidemic and SimBet lead to longer routing paths almost resulting in an average hop value of 27 and 41 respectively. Interestingly, we find that the average number of hops per messages resulted from SimBet is even bigger than that of Epidemic. This outlier is, we conjecture, due to the repeated infection caused by SimBet, i.e., when node \( i \) has delivered message \( m \) to node \( j \), they are more likely to exchange messages with each other.
VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel routing metric, called Hoten, to route messages in DTNs. We exploit hotspot and entropy to design utility function. We first use the relative entropy between the public hotspots and the personal hotspots to evaluate the centrality of nodes. Then we utilize the inverse symmetrized entropy of the personal hotspots of two nodes to compute the similarity between them. Third, we exploit the entropy of the personal hotspots to estimate node personality. Besides, we propose a method to explore the optimized size of hotspot. Trace-driven simulation results show that Hoten largely outperforms other solutions, especially in terms of combined overhead/packet delivery ratio and the average number of hops per message.

One significant topic for future work is to study the influence of temporal correlation of stay points on the Hoten performance.

REFERENCES

[1] Kevin Fall, Stephen Farrell. DTN: An Architectural Retrospective. IEEE Journal on Selected Areas in Communications, 2008, 26(5):828-836.
[2] W. Iancic, W. M. Eddy, D. Stewart, et al. Experience with delay-tolerant networking from orbit. In Proceedings of ASMS 2008, IEEE, pp.173-178.
[3] Rua Qin, Zi Li, Yanfei Wang, et al. An Integrated Network of Roadside Sensors and Vehicles for Driving Safety: Concept, Design and Experiments. In Proceedings of Percom 2010, IEEE, pp.79-87.
[4] Yang Zhang, Wei Gao, Guohong Cao, et al. Social-Aware Data Diffusion in Delay Tolerant MANETs. Work. Springer Publisher. Retrieved from http://mcn.cse.psu.edu/paper/yangzhang/book-chapter10.pdf 2011.
[5] Wei Gao, Guohong Cao, Arun Iyengar, et al. Supporting Cooperative Caching in Disruption Tolerant Networks. In Proceedings of ICDCS 2011, IEEE.
[6] Vahdat A, Becker D. Epidemic routing for partially connected ad hoc networks. [c22200006] Durham, North Carolina : Duke University, 2000.
[7] R. C. Shah, S. Roy, Sushant Jain, et al. Data MULEs: modeling and analysis of a three-tier architecture for sparse sensor networks. Ad Hoc Networks, 2003,1(2−3):215-233.
[8] A. Lindgren, A. Doria, O. Schelen. Probabilistic Routing in Intermittently Connected Networks. Lecture Notes in Computer Science, 2004,3126:239-254.
[9] M. Musolesi, S. Hailes, C. Mascolo. Adaptive Routing for Intermittently Connected Mobile Ad Hoc Networks. In Proceedings of WoWMoM 2005, IEEE, pp.183-189.
[10] J. Leguay, T. Friedman, V. Conan. Evaluating Mobility Pattern Space Routing for DTNs. In Proceedings of Infocom 2006, IEEE, pp.1-10.
[11] T. Spyropoulos, K. Psounis, C.S. Raghavendra. Efficient Routing in Intermittently Connected Mobile Networks: The Multiple-Copy Case. IEEE/ACM Transactions on Networking, 2008, 16(1):77-90.
[12] V. Erramilli, M. Crovella, A. Chaintreau, et al. Delegation Forwarding. In Proceedings of MobiHoc 2008, ACM, pp.251-259.
[13] Wei Gao, Guohong Cao. Fine-Grained Mobility Characterization: Steady and Transient State Behaviors. In Proceedings of MobiHoc 2010, ACM, pp.61-70.
[14] E. Daly, M. Haahr. Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs. In Proceedings of MobiHoc 2007, ACM, pp.32-40.
[15] Pan Hui, J. Crowcroft, E. Yoncki. BUBBLE Rap: Social-based Forwarding in Delay Tolerant Networks. In Proceedings of MobiHoc 2008, ACM, pp.241-250.
[16] J. Yoon, B.D. Noble, Mingyan Liu, et al. Building Realistic Mobility Models from CoarseGrained Traces. In Proceedings of MobiSys 2006, ACM, pp.177-190.

j, it deletes m and may receive m again when meeting up another node v, which also carries m and has a lower SimBet metric than node i.

K. Lee, S. Hong, S. J. Kim. Demystifying Levy Walk Patterns in Human Walks. Technical Report, CSC, NCSU, 2008.
S. Merugu, M. Anmar, E. Zegura. Routing in Space and Time in Networks with Predictable Mobility. Technical Report GIT-CC-04-7, Georgia Institute of Technology.
S. Jain, K. Fall, S. Patra. Routing in a delay tolerant network. In Proceedings of SIGCOMM 2004, ACM, pp.145-158.
J. Burgess, B. Gallagher, D. Jensen et al. MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networks. In Proceedings of Infocom 2006, IEEE, pp. 1-11.
Cong Liu, Jie Wu. Routing in a Cyclic MobiSpace. In Proceedings of MobiHoc 2008, ACM, pp. 351-360.
P. V. Marsden. Egocentric and sociocentric measures of network centrality. Social Networks, 2002, 24:407-422.
M. EVERETT, S. P. BORGATTI. Ego network betweenness. Social networks, 2005, 27(1):31-38.
I. Rhee, M. Shin, S. Hong et al. On the Levy-walk Nature of Human Mobility. In Proceedings of Infocom 2008, IEEE, pp. 924-932.
S. Kaveevivitchai,H. Esaki. Independent DTNs Message Deletion Mechanism for Multi-copy Routing Scheme. In Proceedings of AINTEC 2010, ACM, pp. 48-55.
Quannan Li, Yu Zheng, Xing Xie et al. Mining User Similarity Based on Location History. In Proceedings of GIS 2008, ACM.
M. Kim, D. Kotz, S. Kim. Extracting a mobility model from real user traces. In Proceedings of Infocom 2006, IEEE, pp. 1-13.
A.-L. Barabasi, R. Albert. Emergence of scaling in random networks. Nature, 1999, 286:509-512.
FREEMAN, L. C. A set of measures of centrality based on betweenness. Sociometry, 1977, 35-41.
FREEMAN, L. C. Centrality in social networks conceptual clarification. Social networks, 1979, 215-239.
L. A. Adamic, E. Adar. How to search a social network. Social Networks, 2005, 27(3):187-203.
T. Fukuhara, T. Murayama, T. Nishida. Analyzing concerns of people from Weblog articles. AI Soc., 2007, 22(2):253-263.
Allport, G. W. Personality traits: Their classification and measurement. Journal of Abnormal and Social Psychology, 1921, 16:6-40.