Location-based Data Dissemination with Human Mobility Using Online Density Estimation

Viet Duc Le, Hans Scholten and P.J.M Havinga
Pervasive Systems, University of Twente
7522 NB Enschede, The Netherlands
Email: {v.d.le, hans.scholten, p.j.m.havinga}@utwente.nl

Hung Ngo
IDSIA, University of Lugano, SUPSI
6928 Manno-Lugano, Switzerland
Email: hung@idsia.ch

***Abstract*—The emerging wave of technology in human-centric devices such as smart phones, tablets, and other small wearable sensor modules facilitates pervasive systems and applications to be economically deployed on a large scale with human participation. To exploit such environment, data gathering and dissemination based on opportunistic contact times among humans is a fundamental requirement. To tackle the lack of contemporaneous end-to-end connectivity in Delay-tolerant Networks (DTNs), most current algorithms assess the probability of the contact times to gradually convey a message towards its destination. These contact-based approaches do not perform well when historical locations of nodes have mixture distribution. In this paper, we formulate routing problems in spatial and spatiotemporal domains as an online unsupervised learning problem given location data. The key insight is that nodes frequently appearing nearer the message destinations are regarded as possessing higher delivery probability even if they have low contact times. We show how to solve the formulated problems with two basic algorithms, Location-Mean and Location-Cluster, by estimating the means of historical locations to calculate delivery probability of nodes. To our best knowledge, this is the first work to tackle DTN routing problem using online unsupervised learning on geographical locations. In the context of human mobility, simulation results of the Location-Mean algorithm show that the online unsupervised learning approach given node locations achieves better routing performances in term of delivery ratio, latency, transmission cost, and computation efficiency compared to the contact-based approach.**

I. Introduction

The increasing number of handheld devices equipped with on-board sensors has inspired the development of pervasive applications for measuring and gathering data of the surrounding environment. In these applications, data dissemination serves a predominant role since the devices are human-centric. Since human mobility is mostly unpredictable, intermittent connectivity and conventional routing algorithms [1], [2] for mobile ad-hoc networks (MANETs) no longer perform well in DTNs. Therefore, new algorithms are required to overcome the contemporaneous end-to-end connectivity in the opportunistic networks. Recently, researchers have proposed a bundle of opportunistic routing algorithms [3]–[12], which can be categorized into two main streams, stochastic and oracle-based algorithms. Stochastic algorithms execute routing based solely on opportunistic contacts. Oracle-based algorithms require network information such as contact times, message utility, and route schedule.

In this paper, we exploit the impact of human mobility on message delivery in opportunistic mobile phones sensor networks. In particular, we consider a delay-tolerant network of human-centric nodes. Conventional DTN's routing protocols have attempted to find the probability of message delivery based on contact times, which indicate how frequently a pair of devices is in connection. Though having showed good performance on message delivery, contact-based algorithms do not perform well when each device frequently appears at different regions, as most people daily do. The reason is that geographic coordinates of nodes have little or no correlation with their contact times. For example, assume that a staff currently has a bundle of messages to send to colleagues either working in or out of his office. Since devices of staffs working in the same office have high contact times, most contact-based algorithms estimate delivery probability of roommates of the staff much higher than that of colleagues in other offices. Therefore, the staff preferably transfers messages to his roommates first, and to a visiting person later. As a visitor just drops in a while, he or she would not have enough time to wait to receive the messages from the staff. As a result, the staff misses a good opportunity to deliver messages to other offices.

Although several contact-based algorithms also consider transitive probability such as Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) [7] and MaxProp [13], these algorithms will fail in case the mobile phone of aforementioned visitor is frequently out of communication range with other mobile phones in his office. This can happen if the visitor is unwilling to participate in dissemination. However, if the message is carried by the visitor, its information can be delivered to the destination by other communication channels because they are in the same office.

The above scenario points out the limitations of current routing approaches based on contact times, and motivates us to propose a new approach based on historical locations of

mobile nodes, called location-based routing. The key insight is that, nodes frequently appearing nearer the message destinations are regarded as possessing higher delivery probability. In our approach, at regular time intervals, each node records its current location in a first-in-first-out (FIFO) buffer, which has an aging parameter deciding its length. If the buffer overflows, the oldest location will be removed to make room for a new recorded location. An appropriate unsupervised learning or clustering algorithm will be applied to infer the full information of location densities from recorded data. When having obtained the information, we estimate the distribution of nodes in the past to select better candidates to carry messages. To our best knowledge, this is the first work to formulate DTN routing problem as an online unsupervised learning problem on historical locations to predict message delivery probability of mobile phone users.

We validate our method by simulating a DTN network with a real map and realistic human movement models. By comparing the results of most well-known contact-based algorithms with a naive online machine learning given node locations, we show that the location-based approach using online machine learning has a great potential for improving routing performances in terms of delivery ratio, latency, number of message transfers, and computing. Results also show that our approach is suitable for large-scale networks in the long term. In addition, online unsupervised learning algorithms can adapt themselves to changing in human movement patterns. This implies using better machine learning algorithms for estimating location densities in both spatial and spatiotemporal domains is promising, and invokes further research in this direction for opportunistic routing algorithms

The paper is organized as follows. Section II summarizes related work and discusses in more detail the novelty of this paper. In Section III, we formulate the routing problem as estimating a mixture of density functions with unknown parameters, which can be solved by online unsupervised learning and clustering tools. Section III introduces two basic methods to solve the formulated problem. Section IV presents simulation results, and Section V concludes the paper with a brief summary of contributions and discussion on future work.

II. Related Work

Stochastic routing protocols, such as Epidemic [4], First Contact (FC) [5], and Direct Delivery (DD) [6] solely broadcast messages to any encountered node, in order to increase the delivery ratio. Epidemic routing diffuses messages similar to the way viruses or bacteria spread in biology. Whenever encountering another node, a node replicates and transfers messages. A node, which just received the messages, will move to other places, and continuously replicate and deliver the messages to other encountered ones. First Contact, a variant of single-copy scheme, sends messages to the first encountered node without copying the messages. Spray-and-wait comprises the trade-off between epidemic and first-contact by finding an optimal number of copies of messages. Creating more copies of a message increases the message delivery but decreases the network lifetime. These stochastic routing approaches consider the destinations of messages as nodes but locations. Messages may be sent to nodes which never visit the place of delivery, particularly when nodes just ramble within a specific area and the destination of messages is in another area. Under such circumstance, stochastic routing protocols have a poor performance.

Unlike stochastic routing, current contact-based routing uses a selective mechanism to choose the most appropriate nodes conveying messages to the destinations based on historical contact information, for instance, contact times, contact duration, and contact cycle. Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) [7] is a well-known Context-based routing protocol based on the history of encounters. PRoPHET estimates the delivery predictability for each known destination at each node before passing a message. The estimation is based on the history of encounters between nodes. SimBet [8] uses historical contacts to calculate two metrics, *similarity* and *betweenness*. The similarity, which is calculated by how frequently a node and its destination have met, is meant of how socially connected such two nodes are. The betweenness, which is calculated by how many nodes which a node has met, shows how interconnected a node is. However, if the utility metrics are equal, SimBet will prevent its forwarding behavior. To improve this flaw, BUBBLE [9] adds the knowledge of community structure to ensure message diffusion. Since the social knowledge varies over time, information used by BUBBLE may be outdated. In addition, the betweenness may be useless if the message is near its destination. An improved version of SimBet, called SimBetAge [11], was proposed to address these shortcomings.

As mentioned in introduction, we observe that although improving the delivery ratios and deducting delivery cost, most current contact-based algorithms [7]–[9], [11], [14] estimate the delivery probability of a node based on the information from pairwise contact times, which does not truly reflect a delivery probability when nodes frequently appear at several regions and are unwilling to dispatch messages automatically.

With the increasing computing power in smart phones, applying online unsupervised learning on historical locations for routing has great potential. However, it has not drawn much researchers' attention to location-based routing in DTNs. Recent work [15] does routing by simply calculating the delivery probability for a node to be at a location in MobySpace, which is a high dimensional Euclidean space based on the pre-known mobility model. However, the required assumption of that each node has the knowledge about mobility patterns of other nodes in the network makes this work unpractical in realistic scenarios.

To this end, we introduce a new concept of routing in Delay-tolerant Networks and present promising solutions based on Machine Learning. By online learning the distribution of nodes in the past, we estimate the probability of a node can deliver a message to a destination. This approach can also be applied to nodes having mobility patterns that are hard to predict.

III. Location-based DTN Routing

A. Problem Formulations

In this section, we formulate two routing problems in spatial and spatiotemporal domains. The problem in spatial domain is simpler but less reliable than that in spatiotemporal domain. In addition, storing both time and coordinates of mobile phones carried by users requires a more sophisticated security scheme to protect private data. Depending on the requirements of applications, researchers can select the suitable domain to apply.

1) Spatial Domain: We consider a network of n mobile sensor nodes, denoted by the set $\mathcal{S} = \{s_1, \ldots, s_n\}$, which have unpredictable moving patterns. For each node i (i.e., s_i, used interchangeably) let $\mathcal{D}_i = \{\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,k}\}$ denote the set of its k most recent locations recorded with time interval Δ. With a slightly abuse of notation, denote the set of encountered nodes of s_i, including itself, as $\mathcal{E}_i = \{s_i, s_{i+1}, ..., s_{i+e}\}$, where e is the number of nodes currently connected to s_i. The set of location history induced by the set $\mathcal{E}_i$ is defined as $\mathcal{D}_i^e = \{\mathcal{D}_i, \mathcal{D}_{i+1}, \ldots, \mathcal{D}_{i+e}\}$. The set of l messages which is being held by node s_i, and needs to be delivered, is denoted as $\mathcal{M}_i = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_l\}$. Each message might have various attributes such as destination, time-to-live, message size, message priority, etc. The problem is for each node to decide which node in its encounter set is the best next message carrier for each message so as to quickly and reliably deliver the message.

We propose a probabilistic framework to solve the above problem. For each message $\boldsymbol{m}_j \in \mathcal{M}_i$, the destination coordinate is the only relevant attribute concerned in this paper to decide a successful delivery, which in turn depends on the set $\mathcal{D}_i$. Other attributes can be used to sort $\mathcal{M}_i$ in advance by a buffer management [16], [17]. By assuming $\mathcal{D}_i$ is parameterized by a vector of unknown parameters $\boldsymbol{\theta}_i$, the delivery probability can be defined as the probability density function $p_{i,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_i)$ conditioned on the parameter vector $\boldsymbol{\theta}_i$. Here we use a shorthand notation $p_{i,j}(\boldsymbol{m}_j|\cdot)$ to denote the event of a successful delivery of message $\boldsymbol{m}_j$. The set of delivery probabilities of all nodes in $\mathcal{E}_i$ for message $\boldsymbol{m}_j$ is

$$P_{i,j} = \{p_{i+k,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_{i+k})\}_{k=0}^{e}. \tag{1}$$

Let $\boldsymbol{\Theta}_i^e = \{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_{i+e}\}$, the set of unknown parameters for the nodes in $\mathcal{E}_i$. The set of all parameter vectors at all nodes is called *location distribution*, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$. Our first goal is to estimate the set of parameters $\boldsymbol{\Theta}$ using the location history at all nodes, $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$. Once the estimate $\widehat{\boldsymbol{\Theta}}$ (and hence $\widehat{\boldsymbol{\Theta}}_i^e$) is available, each node s_i can calculate the set of delivery probabilities $P_{i,j}$ for each message $\boldsymbol{m}_j$. Subsequently, by making pairwise comparisons, node s_i will find a candidate with higher delivery probability and not holding the message to transfer.

Let us reconsider the problem of learning location distributions from historical data unlabeled to which geographical regions. These data sets may be geometrically viewed as clouds of points in a d-dimensional space ($d = 2, 3$). Finding a location distribution in our approach is a typical unsupervised learning and clustering problem. A location distribution of node i generally falls into two categories, a single distribution or a mixture of z_i distributions. In essence, given historical-location sets $\mathcal{D}$, we have to find the estimate $\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_n\}$ of $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$. In particular, given historical location set $\mathcal{D}_i$ of node i, the problem is to find the estimate $\widehat{\boldsymbol{\theta}}_i = \{\widehat{\boldsymbol{\theta}}_{i,1}, \ldots, \widehat{\boldsymbol{\theta}}_{i,z_i}\}$ of full parameter $\boldsymbol{\theta}_i = \{\boldsymbol{\theta}_{i,1}, \ldots, \boldsymbol{\theta}_{i,z_i}\}$. Then, we can estimate the delivery probability of node i to deliver a message $\boldsymbol{m}_j$ by

$$\widehat{p}_{i,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_i) = \max_{k=1,\ldots,z_i} \{p_i(\boldsymbol{m}_j|\omega_{i,k}, \widehat{\boldsymbol{\theta}}_{i,k}) P_i(\omega_{i,k})\}, \tag{2}$$

where $P_i(\omega_{i,k})$ is the prior probability of each class with state of nature $\omega_{i,k}$. This is a simplified solution to the mixture of distribution, in which we put all weight on the single best distribution. Finally, for each message $\boldsymbol{m}_j$, node i selects the next carrier $s_c(i,j)$ to deliver the message as the one with highest estimate delivery probability,

$$s_c(i,j) = \arg \max_{k=0,\ldots,e} \{\widehat{p}_{i+k,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_{i+k})\}. \tag{3}$$

2) Spatiotemporal Domain: Now we seek to formulate the above problem in spatiotemporal domain. At a time slot t during a cycle of $\mathcal{T}$ time slots, which can be a day, a week, or a month, let $\mathcal{D}_i^t = \{\boldsymbol{x}_{i,1}^t, \ldots, \boldsymbol{x}_{i,k}^t\}$ denote the set of node i's locations recorded at time slot t of k most recent cycles. The set of location history induced by the set $\mathcal{E}_i$ becomes a matrix $\mathcal{D}_i^e = \{\mathcal{D}_i^t, \mathcal{D}_{i+1}^t, \ldots, \mathcal{D}_{i+e}^t\}_{t=1}^{\mathcal{T}}$. Besides the destination coordinate, the expected delivery time attribute needs to be considered to estimate the delivery probability of node i to deliver a message $\boldsymbol{m}_j$. Therefore, the delivery probability of node i to delivery message $\boldsymbol{m}_j$ on expected time slot t is defined as the probability density function $p_{i,j}^t(\boldsymbol{m}_j|\boldsymbol{\theta}_i^t)$ conditioned on the parameter vector $\boldsymbol{\theta}_i^t$, which parameterizes $\mathcal{D}_i^t$. For time slot t in spatiotemporal domain, equations (1), (2) and (3) can be rewritten as

$$P_{i,j}^t = \{p_{i+k,j}^t(\boldsymbol{m}_j|\boldsymbol{\theta}_{i+k}^t)\}_{k=0}^{e}, \tag{4}$$

$$\widehat{p}_{i,j}^t(\boldsymbol{m}_j|\boldsymbol{\theta}_i^t) = \max_{k=1,\ldots,z_i} \{p_i^t(\boldsymbol{m}_j|\omega_{i,k}^t, \widehat{\boldsymbol{\theta}}^t{}_{i,k}) P_i^t(\omega_{i,k}^t)\}, \tag{5}$$

and

$$s_c^t(i,j) = \arg \max_{k=0,\ldots,e} \{\widehat{p}_{i+k,j}^t(\boldsymbol{m}_j|\boldsymbol{\theta}_{i+k}^t)\}. \tag{6}$$

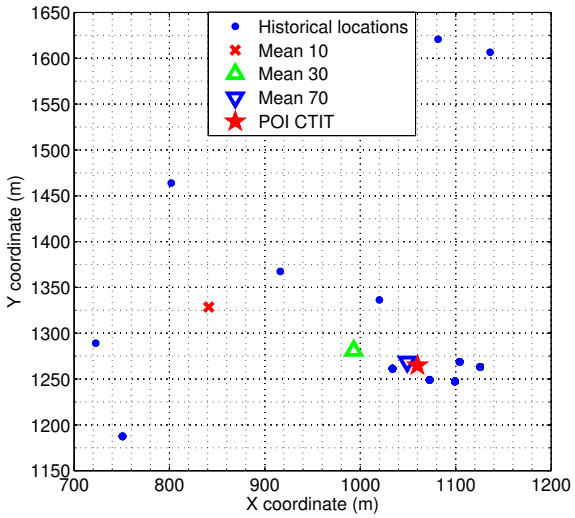

Fig. 1. Illustration of the Location-Mean approach from simulation data.

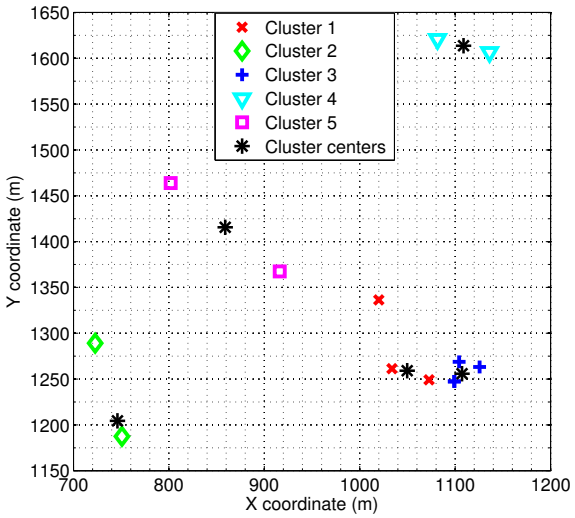

Fig. 2. Illustration of the Location-Cluster approach from simulation data.

Note that if one can solve the problem in spatial domain, a similar solution can be applied to the problem in spatiotemporal domain by drawing historical locations regarding the expected delivery time slot.

The computational complexity depends on online machine learning algorithm. For example, the Location-Mean to be described in Section III-B1, only needs $O(n)$ to update parameters of n nodes. In spatiotemporal domain, the computational complexity becomes $O(n\mathcal{T})$ with the number of expected time slot bounded by $\mathcal{T}$. The memory cost on each node for the problem depends on dimensions of parameters, which are much smaller than original data.

It is infeasible to compare the complexity between location-based and encounter-based approaches because they are based on two different elements, which are not well correlated. Computation load of the location-based scheme depends on the location updating intervals, which are independent from encounters. The shorter intervals, the heavier computation. Meanwhile, computation load of the encounter-based scheme depends on human mobility and density, which decides how frequently people meet each other.

B. Unsupervised Learning Approaches

Due to space limit, we only describe how to estimate the distribution parameters for the spatial domain. The spatiotemporal domain can be solved analogously. Our approach is to use recorded locations to estimate the unknown location distributions of a single best distribution of the mixture densities. For simplicity, assume that the densities follow a Gaussian mixture

$$p_{i,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_i) \sim \aleph(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2), \tag{7}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are vectors of z dimensions, with only z is known as the second case in Table I. The check mark (✓) and question mark (?) indicate known and unknown parameters. Therefore, this problem can be solved by existing classification and clustering tools, such as Gaussian Mixture Model (GMM), Kalman filter, or Support Vector Machine(SVM) [18]. Once the problem is solved, we will obtain estimate prior probability $\widehat{P}_i(\omega_{i,k})$, mean $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\sigma}}_i$ vectors of node i. Afterwards, the delivery probability in Eq. 2 of node i to deliver message $\boldsymbol{m}_j$ is computed by

$$\widehat{p}_{i,j}(\boldsymbol{m}_j|\boldsymbol{\theta}_i) = \max_{k=1,\ldots,z_i} \left\{ \frac{\parallel \boldsymbol{m}_j - \widehat{\boldsymbol{\mu}}_{i,k} \parallel}{\sum_{k=1}^{z_i} \parallel \boldsymbol{m}_j - \widehat{\boldsymbol{\mu}}_{i,k} \parallel} \widehat{P}_i(\omega_{i,k}) \right\}, \tag{8}$$

where $\boldsymbol{m}_j$ and $\boldsymbol{\mu}_{i,k}$ are also coordinates of the message destination and cluster centers, respectively.

Following we present two simplest tools to solve the problem of location densities, Location-Mean (mean) and Location-Cluster (k-means) for $k = 1$ and $k = z_i$, respectively.

TABLE I
THREE CASES OF MIXTURE GAUSSIAN ESTIMATION [18]

Cases	$\boldsymbol{\mu}_i$	$\boldsymbol{\sigma}_i^2$	$P_i(\omega_{i,k})$	z_i
1	?	✓	✓	✓
2	?	?	?	✓
3	?	?	?	?

1) Location-Mean: Suppose we knew the distribution of historical locations of node i came from a single normal distribution with a mean $\boldsymbol{\mu}_i$ and standard deviation $\boldsymbol{\sigma}_i$. Essentially, these two parameters constitute a compact representation of

the movement pattern. If the mobile phone user actually stays most of the time in a specific place, such as his office building, the historical locations has a mean that tends to fall in the region where the user mostly stays. Of course, if the samples from a user is not normally distributed, the Location-Mean approach can give very misleading description of movement pattern, and the estimate delivery probability will be wrong. At each time t, when $\boldsymbol{x}_{i,t}$ is updated, the mean $\widehat{\boldsymbol{\mu}}_i$ can be updated incrementally as:

$$\widehat{\boldsymbol{\mu}}_i \leftarrow \frac{N_i}{N_i+1}\widehat{\boldsymbol{\mu}}_i + \frac{1}{N_i+1}\boldsymbol{x}_{i,t} \tag{9}$$

where N_i is the current number of historical locations of node i.

Figure 1 illustrates the historical coordinates and their means of a node moving according to the movement model to be described in Section IV. We tested a variety of historical lengths. Means 10, 30, 70 are the mean values of the 10, 30, and 70 recorded locations of the node representing a user at the CTIT institute. Because the simulated user spends most of the time in his building, of which the main gate is marked as the star 'POI CTIT', the mean values of 30 and 70 are quite close. This gives a clue that choosing the length of 30 latest historical locations is sufficient. Note that the time interval we chose to record locations in the simulation is randomly drawn between $250-350$ seconds.

2) Location-Cluster: If we consider a longer period of human activity than working hours, the model of a mobile phone user locations probably is a mixture distribution instead of a single normal distribution. We observe that most people still spend most of their time in several places, such as their house, offices, bars, sport centers, etc. Therefore, the normal mixture with unknown number of class z can give a close description of the location densities, as case 3 in Table I. There are several machine learning methods to estimate the number of classes z of each person. The number of classes z can also be obtained by asking the mobile phone user. Note that each node i has its own value of class number, z_i.

We use k-means to solve the defined problem since this technique has efficient online update, thus it can simplify the computation and accelerate convergence. In particular, k-means computes the squared Euclidean distances $\|\boldsymbol{x}_{i,t} - \widehat{\boldsymbol{\mu}}_{i,k}\|^2$ at each time t to find the mean $\widehat{\boldsymbol{\mu}}_{i,k}$ nearest to $\boldsymbol{x}_{i,t}$ with $k = 1, ..., z_i$. K-means does not require to know z_i in advance; instead, z_i can be inferred from the given data. When the distance to $\widehat{\boldsymbol{\mu}}_{i,k}$ is greater than a given threshold δ, we increase z_i by one $z_i = z_i + 1$. The mean of the newly created cluster is $x_{i,t}$. At each time t, when $\boldsymbol{x}_{i,t}$ is assigned to cluster k, the mean $\widehat{\boldsymbol{\mu}}_{i,k}$ can be updated incrementally as:

$$\widehat{\boldsymbol{\mu}}_{i,k} \leftarrow \frac{N_{i,k}}{N_{i,k}+1}\widehat{\boldsymbol{\mu}}_{i,k} + \frac{1}{N_{i,k}+1}\boldsymbol{x}_{i,t} \tag{10}$$

where $N_{i,k}$ is the current number of data points assigned to cluster k. So each node needs to store only tuple of cluster mean values and their number of assignments, instead of storing the whole data it receives for its whole lifetime.

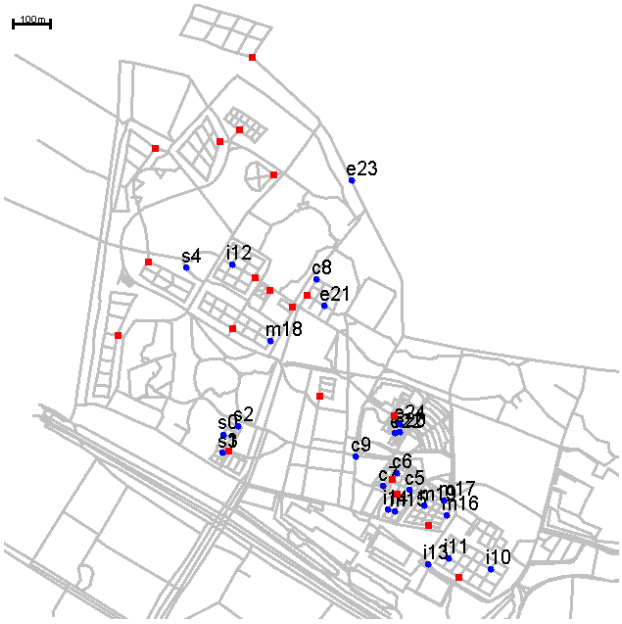

Fig. 3. Screenshot of simulation. WiFi access points marked as square and pedestrians marked as circles.

It is interesting to see how the k-means operates on the example data we used in Figure 1. Figure 2 shows 5 cluster centers and theirs clustered locations. These cluster centers give a compact description of 5 places in campus the node frequently occurs.

We note that there are better Machine Learning tools to solve (2), for instance, Maximum Likelihood, Support Vector Machine, Decision Tree, and Gaussian Mixture Model. Some of them may have very high computation that should be considered since mobile phones have limited computation capability and battery.

IV. Simulation

In this section, we will present our preliminary results, the Location-Mean performances. Note that the recent work studying the nature of human mobility [10], [19]–[21] has proved that suitable movement models can sufficiently present the behavior of human mobility. A realistic model of human mobility does not mean that the movement pattern is predictable; instead, it better characterizes the unpredictable human mobility rather than the simple Random Walk [22].

A. Simulation Settings

The simulation is based on a realistic scenario of the University of Twente campus shown in Figure 3. The routes and Points of Interests (POIs) such as offices, sport centers, stadiums, tennis courts, libraries, restaurants, shops, supermarkets,

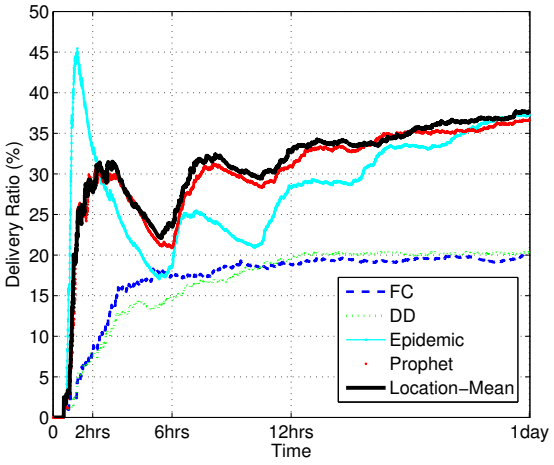

Fig. 4. Convergence of delivery ratios.

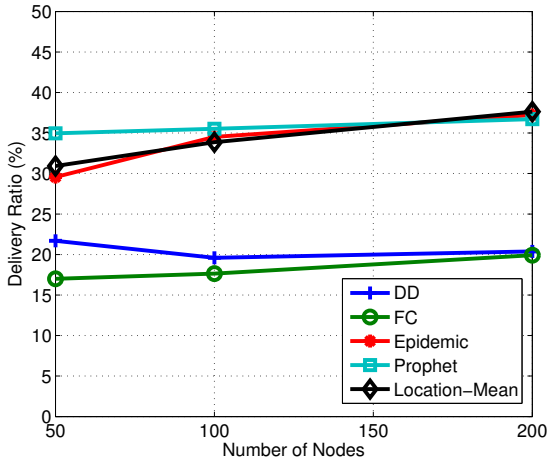

Fig. 5. Delivery ratio vs. Number of mobile nodes.

staff houses, and dormitories are mapped into the simulation. For each place, there is a WiFi access point installed at each main entrance, which is marked as square in Figure 3. These nineteen access points are also the sinks of messages that are randomly generated at one of mobile phones, which are carried by students and staff. We assume that the speed of pedestrians remains almost constant, $0.5-1.5$ m/s. Therefore, the mobility speed has a minor effect on performance results.

The mobile phones and sinks are supposed to possess a WiFi interface at net data rate of 11 Mbit/s with 30 m radio range. Every thirty seconds, a new message with size of 500 Bytes to 1 KBytes is created, and its destination is one of the access points. Buffer sizes of mobile phones and sinks are 25 KBytes and 25 MBytes, respectively. The First In First Out (FIFO) is applied on buffer management.

In addition, students and staff are split into 5 sub-groups, called STAFF, CTIT, IMPACT, MESA++, and ELAN as the name of research institutes in University of Twente. Their movements are modeled with the Shortest Path Map Based Movement, which is presented by [23], and various POIs. At a certain moment, a node will choose one of nineteen POIs with predefined probability. In particular, the probability of a node to visit libraries, shops, or sport centers are 5%, 10%, and 5%, respectively. Since we concern the effect of humans in data dissemination during day time, 60% of the time students and staff stay in their offices, and only 10% of the time they visit their homes for a while. Moreover, we assume that every 30 to 60 minutes there is at least a person entering or leaving a building in the simulation.

With above settings, our proposed algorithm is evaluated and compared with a number of well-known opportunistic routing protocols: Direct Delivery (DD) [6], FirstContact (FC) [5], Epidemic [4], and PRoPHET [7]. Since Location-Mean is a very naive algorithm as an example for location-based, we do not include results of better contact-based algorithms, such as BUBBLE and SimBetAge, to make a fair comparison.

B. Evaluation Metrics

Three metrics are used to evaluate the aforementioned performance requirements of different routing algorithms: delivery ratio, latency, and transmission cost. Note that the hop-count metric is no longer an informative metric to assess the delivery cost in time and distance in DTNs as it is used in connected ad-hoc WSNs so that we do not use it to evaluate our work.

- *Delivery ratio R:* The total number of successfully delivered unique messages, denoted by Q, divided by the total number of created unique messages, denoted by P. Each unique message is created at certain time, and has an unique identification number to be distinguished from others in the network.

$$R = \frac{Q}{P}. \tag{11}$$

- *Latency (L):* The average of delays between the moment that unique message i is originated, denoted by Ts_i, and the time when the first replicate of unique message i arrives at the destination, denoted by Td_i.. The replicate is a copy of an unique message. The number of replicates depends on the methodology of the DTN routing algorithm, single or multiple-copies.

$$L = \frac{1}{Q}\sum_{i=1}^{Q}(Td_i - Ts_i). \tag{12}$$

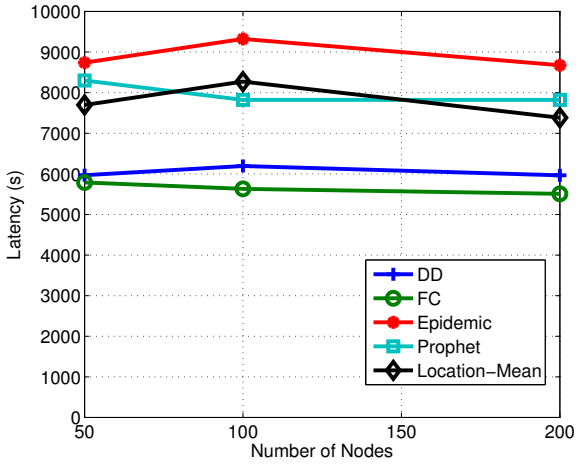

Fig. 6. Latency vs. Number of mobile nodes.

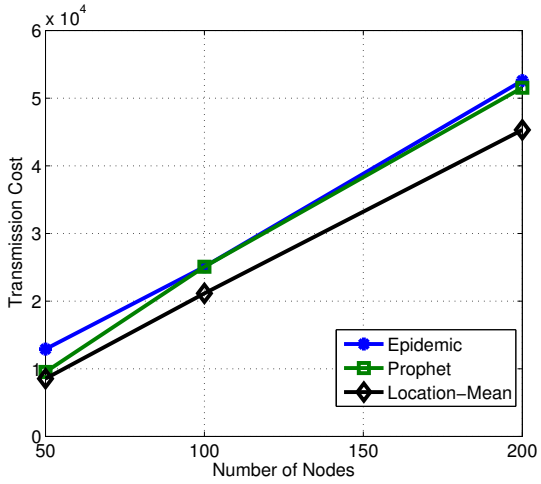

Fig. 7. Transmission cost vs. Number of mobile nodes.

- *Transmission cost (C):* The total number message transmissions, denoted by T, divided by the number of successfully delivered messages.

$$C = \frac{T}{Q}. \tag{13}$$

C. Simulation Results

All results are averaged over 5 runs with difference random seeds to simulate one day in real time. Figure 4 shows the convergence of delivery ratios when time increases. Because of the limited buffer size and contact durations, delivery ratios quickly converge after first two hours. The delivery ratios of DD and FC firmly converge after 12 hrs while those of Epidemic, Prophet and Location-Mean still slightly raise up. We also observe that the delivery ratio given by Location-Mean reaches 41% after two days and still keep slightly increasing. This is explained by the longer period, the closer estimate the means of the locations.

We also evaluate the delivery ratios of the algorithms by varying the number of participants in the above scenario, from 50 to 200. We remark that the number of mobile nodes here represents the people moving in and out of building, not the number of total students and staffs. In such way, 200 can represent 2000 people working in the campus, and thus save a lot of simulation time.

Figure 5 shows the percentages, after one day, slightly rise when the number of nodes increase. This is expected since the more mobile nodes produces more contact opportunities. Remark that this observed increase is not hold with DD because it only sends a message to the destination node. This can be proved using Random Walk theory [22]. Location-Mean scores best among compared algorithms when increasing the number of nodes. This implies that Location-Mean knows better which nodes to transmit the messages than the others, which seem like randomly dissemination.

We also examine how the Location-Mean performs in terms of latency as shown in Figure 6 since the time taken to deliver messages is important. We measure the average delays when changing the number of nodes from 50 to 200 as we evaluate the delivery ratio. Since our algorithms can predict the potential nodes better to avoid the traffic loads, its latency is lower others multi-copies schemes when increasing the number of nodes. We remarked that the average delay obtained by our algorithm, about 120 minutes, is quite long for some applications. However, it makes sense because the delivery totally relies on walking speed in a campus of 4 km^2, and mobile phone users stay in their office or class most of the time. Note that the problem caused by latency can be solved by prioritizing messages based on required delivery time.

We remarked that latency obtained by Epidemic is higher than by Prophet in Figure 6 is reasonable. Only under ideal conditions such as unlimited buffer sizes and all messages can be exchanged during any contact duration, Epidemic will give the lowest latency. However, our simulation is set with limited buffer sizes (message queue), contact durations, and very short communication ranges. This makes Epidemic have longer delay than Prophet, which is consistent with investigation in [7].

Resource consumption is always a key metric in evaluating routing algorithms in mobile phone sensor networks. Figure 7 shows the transmission costs, defined in Section 3, of our proposed algorithm and some existing algorithms. DD has the lowest transmission cost since it only transfers messages to the destinations. FC also has very low transmission cost since it is single-copy routing. Therefore, we subtract them from the plot to have a clearer visualization. Among the multi-

copy schemes, Location-Mean has lowest transmission cost as we expected. Since Location-Mean infers movement patterns of nodes based on locations, it hands the messages to better candidates to avoid roaming messages.

In addition, we observed that simulation with Location-Mean ran faster than that of Epidemic and Prophet for identical settings, which validate computation complexity discussed in Section III. For instance, simulation time for Location-Mean is 3802 seconds with described 200-node scenario, 20% and 12% shorter than Prophet and Epidemic, respectively.

V. Conclusion and Future Work

Addressing routing algorithms for opportunistic mobile phones sensor network with unpredictable mobility of humans, this paper draws up guidelines on approaches by applying online unsupervised learning on the historical locations of nodes. Given recent locations, the delivery probability of a node is estimated through solving a mixture densities problem. Through realistic simulation scenarios and movement models, the results are consistent with the theory of the proposed Location-Mean solution. This paper also gives implications for further development of opportunistic routing algorithms with online unsupervised learning and clustering for location densities in both spatial and spatiotemporal domains. With better density parameter estimation methods compared to the simple ones used in this paper, we expect the performance to be improved with large margin. Following this research, a combination of support vector machine (SVM) and decision tree [24] is planned to be implemented with WiFi 802.11 b/g/n on Nexus 7.

Acknowledgment

This work is supported by the SenSafety project in the Dutch Commit program, www.sensafety.nl.

References

[1] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (aodv) routing," RFC Editor, United States, Tech. Rep., 2003.

[2] D. B. Johnson, D. A. Maltz, and J. Broch, "Ad hoc networking." Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001, ch. DSR: the dynamic source routing protocol for multihop wireless ad hoc networks, pp. 139–172.

[3] V.-D. Le, H. Scholten, and P. Havinga, "Unified routing for data dissemination in smart city networks," in *Proc. of the 3rd International Conference on the Internet of Things (IoT2012)*, 2012.

[4] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Department of Computer Science, Duke Univeristy, Durham, NC, Tech. Rep., 2000.

[5] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proc. of ACM SIGCOMM on Wireless and Delay-Tolerant Networks*, 2004.

[6] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Single-copy routing in intermittently connected mobile networks," in *Proc. of Sensor and Ad Hoc Communications and Networks (SECON)*, 2004, pp. 235–244.

[7] A. Lindgren and A. Droia, "Probabilistic routing protocol for intermittently connected networks," *Internet Draft draft-lindgren-dtnrg-prophet-02, Work in Progress*, 2006.

[8] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '07. New York, NY, USA: ACM, 2007, pp. 32–40.

[9] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '08. New York, NY, USA: ACM, 2008, pp. 241–250.

[10] V.-D. Le, H. Scholten, and P. Havinga, "Towards opportunistic data dissemination in mobile phone sensor networks," in *Proc. of The Eleventh International Conference on Networks (ICN 2012)*, 2012.

[11] J. A. Bitsch Link, N. Viol, A. Goliath, and K. Wehrle, "Simbetage: utilizing temporal changes in social networks for pocket switched networks," in *Proceedings of the 1st ACM workshop on User-provided networking: challenges and opportunities*, ser. U-NET '09. New York, NY, USA: ACM, 2009, pp. 13–18.

[12] V.-D. Le, H. Scholten, and P. Havinga, "Evaluation of opportunistic routing algorithms on opportunistic mobile sensor networks with infrastructure assistance," *International Journal On Advances in Networks and Services*, vol. 5, no. 3 and 4, pp. 279–290, 2012.

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

[14] C. Liu and J. Wu, "Routing in a cyclic mobispace," in *Proc. of in ACM MobiHoc08*, 2008.

[15] J. Leguay, T. Friedman, and V. Conan, "Dtn routing in a mobility pattern space," in *Proc. of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, 2005.

[16] Y. Li, M. Qian, D. Jin, L. Su, and L. Zeng, "Adaptive optimal buffer management policies for realistic dtn," in *Proc. of the 28th IEEE conference on Global telecommunications (GLOBECOM'09)*, 2009, pp. 2683–2687.

[17] G.Fathima and R. Wahidabaru, "Buffer management for preferential delivery in opportunistic delay tolerant networks," *International Journal of Wireless and Mobile Networks (IJWMN)*, vol. 3, pp. 15–28, 2011.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[19] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms," in *Proc. IEEE Infocom*, 2006.

[20] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 19, pp. 189–205, 2011.

[21] A. Lindgren, T. Karkkainen, and J. Ott, "Simulating mobility and dtns with the one," *Journal of Communications*, 2010.

[22] D. Aldous and J. Fill, *Reversible markov chains and random walks on graphs. (monograph in preparation)*. [Online]. Available: http://www.stat.berkeley.edu/ aldous/RWG/book.html

[23] A. Keranen, J. Ott, and T. Karkkainen, "The one simulator for dtn protocol evaluation," in *Proc. of the 2nd International Conference on Simulation Tools and Techniques(SIMUTools)*, 2009.

[24] K. Bennett and J. A. Blue, "A support vector machine approach to decision trees," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, vol. 3, 1998, pp. 2396–2401.