Zakhary, Sameh and Radenkovic, Milena and Benslimane, Abderrahim (2014) Efficient location privacy-aware forwarding in opportunistic mobile networks. IEEE Transactions on Vehicular Technology, 63 (2). pp. 893-906. ISSN 0018-9545

Efficient Location Privacy-Aware Forwarding in Opportunistic Mobile Networks

Sameh Zakhary, Milena Radenkovic and Abderrahim Benslimane, *Senior Member, IEEE*

Abstract

This paper proposes a novel fully distributed and collaborative k-anonymity protocol (*LPAF*) to protect users' location information and ensure better privacy while forwarding queries/replies to/from un-trusted Location-based Service (LBS) over opportunistic mobile networks (*OppMNet*). We utilize a lightweight multi-hop Markov-based stochastic model for location prediction to guide queries towards the LBS's location as well as to reduce required resources in terms of retransmission overheads. We develop a formal analytical model and present theoretical analysis and simulation of the proposed protocol performance. We further validate our results by performing extensive simulation experiments over pseudo realistic city-map using map-based mobility models and using real-world data trace to compare *LPAF* to existing location privacy and benchmark protocols. We show that *LPAF* manages to keep higher privacy levels in terms of k-anonymity, and quality of service in terms of success ratio and delay, compared to other protocols while maintaining lower overheads. Simulation results show that *LPAF* achieves up to 11% improvement in success ratio for pseudo realistic scenarios, while real-world data trace experiments show up to 24% improvement with a slight increase in the average delay.

Index Terms

Mobile ad-hoc network, Distributed computing, Anonymity, Location privacy.

I. Introduction

Location information has become a modern commodity where it is being used by many businesses to provide user tailored services known as location-based services (*LBSs*). The wide penetration of mobile devices capable of detecting, storing and sharing users' location information raises many privacy issues. Privacy-conscious users are more aware of the risks and potential threat of such widely accessible information and tracking capability [1]. *OppMNet* is a special class of Delay-Tolerant Networks (DTNs), where opportunistic communications occur over multi-hop store-carry-forward between mobile devices when they are in the communication area of each other

S. Zakhary is with the School of Computer Science, University of Nottingham, UK (e-mail: szz@cs.nott.ac.uk).

M. Radenkovik is with the School of Computer Science, University of Nottingham, UK (e-mail: mvr@cs.nott.ac.uk).

A. Benslimane is with the Computer Science Laboratory of Avignon (LIA), University of Avignon, France (e-mail: abderrahim.benslimane@univ-avignon.fr).

(called *encounter*), with no prior knowledge of when these encounters will occur. Opportunistic communication relies on users' cooperation to forward queries over a multi-hop path. Privacy-concerned users typically disable their devices opportunistic capability in order to preserve their own location privacy, thus causing a breakdown of communication [2].

Incorporating location information in mobile networks sparked many innovative applications, such as building robust recommendation system [3], [4], publishing and sharing cycling routes [5], or inferring users future travel destinations [6]. Online social networks, such as Facebook and Google+, capture the real-world social relationship between users. However, ensuring users' location-privacy while accessing location-based services in *OppMNet* is a challenging problem due to sharing of users' location information in order to receive tailored services. Moreover, users cannot rely on a trusted third party (TTP) to anonymize or obfuscating their location information due to the disconnections and lack of infrastructure.

This paper is concerned with the source k-anonymity location-privacy when contacting an LBS in *OppMNet* through obfuscation. *Obfuscation* refers to the collaborative activities by nodes to deliberately degrade the quality of information collected by the LBS about the source of the LBS query. We focus on a class of LBSs that does not require user identity in order to provide the service. Other anonymity and security aspects in *OppMNet* are outside the scope of this paper. We utilize a stochastic model for location predication, and propose a lightweight Markov model to drive the privacy preserving protocol. We recognize two possibilities based on whether the nodes are capable of knowing their exact location (such as GPS coordinates). For the first case, where the nodes are able to detect their exact location coordinates, inspired by [7], we propose a path prediction maintained locally at the individual nodes using Markov model proximity. As for the second case, where the nodes are unable to determine their exact location coordinates, we use the recorded Ids for the sighting of fixed infrastructure points (either access-points or GSM cell IDs) as the location identifier. Unlike [7], our proposed model utilizes multi-hop prediction. Nodes exchange their local predictions and that of their own friends when they meet each other, during opportunistic encounters, to help obtain more accurate/updated prediction estimate through shared knowledge exchanged in a distributed way.

Due to *OppMNet* challenges, users need to utilize each encounter as well as leverage the social relationships efficiently with other nodes. Intelligent location predication is needed in order to better enable queries to reach their final destination (LBS) during the obfuscation phase and hence increase both efficiency (*i.e.*, increase overall network success ratio and decrease the number of retransmissions) and delay (as query reaches its destination). In this paper, we present a forwarding protocol that aims to balance between the following: increasing users' location-privacy (k-anonymity), maintaining high quality of service (success ratio and delay) and using limited resources (retransmission overhead). We assume that users trust their friends in their social group with providing location obfuscation when an encounter happens. During obfuscation, a node that carries a number of queries to obfuscate searches for the best neighbors to forward these queries to in order to increase the source privacy by keeping the query within the social group.

As in [8], [9] and [10], a social network represents a pre-existing social ties which can be determined by using either: 1) similarity between users' profiles [10] (*e.g.*, interest, social affiliation), as shown in section III-B1; 2) nodes' encounter patterns [11], as shown in section V-B. Two users are called *friends* if they are directly connected over the social network. We use the term *social group* to refer to the subset of the *social network* that contains users that are friends of every other user in that group. A user can belong to more than one *social group*. The *social group* represents a set of all *friends*, similar to figures (10a, and 10b). A user can build a representation of her own social graph representing a subset of the social network, where vertexes represent users, and edge represent a friendship relationship. A user is either: connected directly to her *friends* (*i.e.*, friends are one edge away), or indirectly through intermediary *friends* to her *indirect friends* (*e.g.*, friends-of-friends) over two or more edges.

The major contributions of this paper can be summarized as follows:

1) We present *LPAF*, a k-anonymity based protocol, offering protection to the users locations information and better privacy while forwarding queries/replies to/from un-trusted LBS in *OppMNet*.
2) We develop a lightweight multi-hop Markov-based location prediction model to adaptively inform the obfuscation protocol to achieve higher privacy through mobile node limited/temporal neighborhood knowledge.
3) We present a formal mathematical model to analytically evaluate and simulate the proposed location-privacy protocol in opportunistic multi-hop communication networks.

The rest of the paper is organized as follows: section II discusses the problem of location-privacy and identify unique challenges in *OppMNet*. Section III presents protocol design and underlaying metrics. Section IV presents analytical analysis including the model and its evaluation. In section V, we show performance evaluation comparing proposed protocol to other benchmark and contemporary location-privacy protocols in pseudo realistic simulation and using real-world data traces. Section VI presents related work, while section VII concludes this paper.

II. Location-Privacy Problem analysis in *OppMNet*

A. Problem Definition

This work focuses on the location privacy from the source (query originator) point of view and not the LBS (Provider). The motivation for our work is location-privacy leakage and users' increased concerns while accessing LBS services [9], [12]. The fact that most LBS services used in mobile opportunistic networks are centralized and are offered by large private businesses (such as google, facebook), triggered many research efforts to understand the privacy implications [2], [13], and propose solutions as we will show in section VI. Users are concerned about the aggregation of their private location-data at the provider side, as well as any unauthorized access/disclosure of such data. In addition to that, most LBSs servers are hosted at centralized data centers and the results are delivered back through Internet then delivered via wireless medium through access points, base-stations or ad-hoc communication. The jurisdiction over and privacy/security offered by these data-centers varies, and hence privacy policies applicable to users at one country might not meet requirements of another.

More generally, we can formulate our research question as "*how to enable the users in* OppMNet *to communicate with an LBS while maintaining their location privacy*". More specifically, we investigate the possibility of allowing nodes to communicate without leaving a digital footprint that can be represented as <Node_ID_a, time_b, Location_c> at the LBSs, or without revealing information that allow LBSs to know that two nodes have been at the same location, which can be represented relatively as <Node_ID_a, Node_ID_b, Time_c>. This is a difficult problem because as nodes try to communicate in *OppMNet*, they need to make use of each encounter efficiently. A number of trade-off decisions need to be considered as the users set their privacy-level requirement. Firstly, Users setting higher privacy constraints will suffer from lower data utility; this is defined in terms of query delivery to LBSs and number of answered queries. Secondly, the level of privacy achievable is dependent on many factors and users do not know in advance if they are setting their privacy constraint at appropriate level for their own circumstances. For example, topology sparseness and users own social group size affect greatly the level of achievable privacy as will be discussed in more details at later sections.

In our research question, we refer to the users need for location privacy. There have been many various definitions of what "Location Privacy" is. Beresford and Stajano [14] defines it as "the ability to prevent other parties from learning one's current or past location", Duckham and Kulik [15] define location privacy as: "a special type of information privacy which concerns the claim of individuals to determine for themselves when, how, and to what extent their location-information is communicated to others". The later definition is more aligned with this paper.

B. Why encryption is not the solution to location privacy

This paper is not concerned with security issues in *OppMNet*, and for that reason we aim to explain briefly the challenges and limitation of using *Encryption* to address location-privacy. *Encryption* is widely used to ensure confidentiality, integrity and authenticity of the information transferred between two endpoints. Over traditional networks, data encryption is used during transmission over un-trusted communication medium or when data is saved to storage devices to prevent unauthorized access or modification [17], [18]. Security of any such encryption schema must not rely on the obfuscation of their algorithm, but only on the secrecy of the decryption key. Hence, distribution and management of keys among all nodes in *OppMNet* is critical to any such schema. Due to the lack of a central authority that acts as TTP, using *Encryption* is not suitable to *OppMNet*. Considering the two main types of encryption schemes, symmetric and asymmetric; where symmetric means that both encryption and decryption are performed with the same key, while asymmetric requires two different keys (*e.g.*, Public-key Infrastructure–*PKI*). Using asymmetric keys require both sender and receiver to exchange public keys that will be used before performing any secure communication. Therefore, two nodes who never met will be unable to directly use such scheme, *e.g.*, over multi-hop opportunistic communication. This also implies that each node has to obtain the different keys to communicate securely with all other nodes in the network. Nevertheless, symmetric schemes present the advantages of being computationally cheaper and easier to set up than asymmetric, but do not offer the same protection.

In addition to the *OppMNet* challenges discussed in section II-A, the problem of user location tracking is aggravated through the ability of all participating nodes to eave-drop on the forwarded query/response between source and LBS which then compromises users' location-privacy. And the assumption of an always online TTP or the handset ability to perform strong encryption is generally not applicable in *OppMNet* due to the limited energy on the mobile devices and the long periods of disconnections. In the case of untrusted LBS, as it is in this paper, encrypting the query so that only the destination can receive/decrypt it

III. Achieving k-Anonymity with Location Privacy-aware Forwarding in *OppMNet*s

Many locaction-based services are being delivered over *OppMNet* because: 1) as infrastructure coverage is not always available in urban congested areas, and mostly in rural developing areas; location-based service discovery and delivery over opportunistic networks has gained considerable focus in recent years [19]–[21]. 2) users interested in location-specific information (*e.g.*, query social events [22]–[24], location-based news [25], or upload sensing information [12]) normally be within a specific geographical location area for the LBS reply to reach them, and in other cases the reply delay does not affect the validity for query result (*e.g.*, querying for location-based advertisements [26] provides personalized advertisements during a shopping trip that are relevant to the user who enters specific shops). 3) using *OppMNet* provides a free-of-charge way for accessing services which helps us bridge digital divide as disconnections still exist in the rural [27] and urban scenarios [28]. This has gained considerable attention lately to support digital-inclusion [29] of members of the society helps bring equality between poor/deprived people and those from richer backgrounds. Although some LBS queries can have high sensitivity to delays, such as real-time step-by-step navigation directions for a fast moving vehicle (which is not the scope of this work). In this work we focuse on other type of queries which are less delay-sensitive (*e.g.*, a tourist who sends a query from a historical castle "which kings lived here?"). For LBSs, delay-sensivity will largely depend on the type of the query asked, the geographical scale for which the query is relevant (*e.g.*, city wide, natural reserves, or tourists attractions) and the urgency of that request.

In this section we present *LPAF* algorithm and introduce the different collaborative measures that are collected, calculated and shared locally between nodes in order to provide obfuscation service to their social group.

A. LPAF *Algorithm*

TABLE I: Notation Reference Table.

Parameter	**Meanings**
k	k-anonymity *privacy-level* requirement defined as the number of hops to obfuscate each query.
MP	Node's own predicted "Markov Proximity" as calculated in equation (3).
MMP	Node's "Multi-hop Markov Proximity" as calculated in equation (4).
P_c	the social-profile attributes matching-criteria for obfuscation (*i.e.*, female and/or walking users or *Facebook*[2] friend).
$sd_{ij}(P_c)$	*social-distance* between the profiles of nodes i and j, under criteria P_c as calculated in equation (1).
P_{low}/P_{high}	Low/High matching threshold for the social distance.
P	the profile attributes of the node.
N_r	original node sending the LBS query.
N_{curr}	current node holding a message to be obfuscated. $h_r(m) = k(m) \iff N_{curr} = N_r$
$Nghs_{curr}$	set containing all the immediate (first-hop) neighbors for the current node.
OUT_n	set of possible obfuscation events organized as tuples of <*neighbor, message, score*> to send to possible next-hop from current neighbors.
M_{obf}	set of all messages at N_{curr} that requires obfuscation.
M	set of buffered and incoming new messages at N_{curr} before checking obfuscation requirements.
$k(m)$	function to extract the privacy-level requirements for message m, *i.e.*, the requested minimum number of obfuscation hops.
$h_r(m)$	function to calculate the remaining number of obfuscation hops for message m.

A user, who wishes to send a query to an LBS, searches for nearby friends (P_{high}) and forwards a copy of her query to one of the available friends. This friend then forwards the query intra-social group using social forwarding for a defined number of hops (k) within the user's social group. After the first k hops, the query can be forwarded using any DTN forwarding protocol (*SnW* in this case). This allows the source not to reveal his location by directly contacting the LBS, but rely on the social forwarding protocol to form an obfuscation path to the LBS. If any intermediary node was unable to find a set of socially matching neighbors, then it attempts to use other more socially distant neighbors if they meet a lower privacy threshold (P_{low}), and they are more likely to deliver the query to its final destination as predicted by their reported MMP.

We propose two *Privacy Qualities*: "*High Privacy*" and "*Best-effort Privacy*". The protocol adaptively tries to achieve one of these depending on the network topology (*i.e.*, how often users meet their friends) and whether intermediaries have enough socially-related neighbors to perform location obfuscation. "*High Privacy*" ($P_{high} = 1$) refers to the source's k-anonymity when her query has been forwarded over at least k obfuscation hops, and all intermediate users belong to the source's *social group*. "*Best-effort Privacy*" ($P_{low} < 1$), on the other hand, refers to the source's k-anonymity when any of the intermediate nodes does not belong to the source's *social group* (*e.g.*, obfuscation carried by friends-of-friends). An intermediate node aims to maintain the source's "High Privacy" quality while forwarding a query. It searches through the current neighbors for friends of the source of that query, and forward the obfuscated query to one of them. If none of the current neighbors is a friend of the source, it tries to offer "Best-effort Privacy" quality by forwarding the query to one of its own friends to obfuscate, and so on. In which case, the intermediate node matches the neighbors against a lower privacy_threshold and uses nodes that pass this test. The detailed algorithm to be followed by source or intermediary nodes participating in the obfuscation is shown in Algorithm 1, and can be explained as follows:

Line $1 \cdots$ Line 11: This part of the code is where the user privacy requirements are specified (line 1), and initial preparation is carried. A loop (line 3) goes over all incoming messages to extract queries that require obfuscation into M_{obf}. Each query is checked for any remaining obfuscation hops (line 4). If more obfuscation hops are needed, then it is added to M_{obf} (line 5). Else, (line 4) the message has been fully obfuscated and the source $real_{id}$ and the full path information is removed in preparation for forwarding outside the *social group*. If no queries require obfuscation, the algorithm exists at line 11.

Line $12 \cdots$ Line 24: This part of the code is where nodes filter their neighbors and decide on the best one to perform obfuscation. Each node search for all existing neighbors and match them with available queries to obfuscate M_{obf}. If a neighbor has already obfuscated a query before, then it is not considered further for this query (line 14). Else, we calculate privacy related measures as in (line 15). (line 16) If the current neighbor social distance meets P_{high} (*i.e.*, "High Privacy"), then the query is scheduled for forwarding. If not, (line 19) the social distance is checked against P_{low}, and the query is scheduled for forwarding if passed (*i.e.*, "Best-effort Privacy").

Line $25 \cdots$ Line 30: This part of the code is where obfuscation is carried. The outgoing messages are sorted by priority and privacy measures. For each outgoing messages, the message obfuscation-path is updated (line 27), and the query is sent to that next-hop. In order to forward the reply back to the requester, we propose that each node in the *social group* which has been on the obfuscation path maintains a *mapping table*, that maintain information about the obfuscated query and its source, which is used to speed-up the LBS-reply addressing (line 29). Each of these entries is associated with a timeout value to keep the table size manageable and minimize lookup overhead.

Line $31 \cdots$ Line 41: This code is for maintenance. All obfuscation messages that has expired are reported to the LBS application (App_r) and dropped from the nodes' buffers (line 32). All remaining outgoing queries are sent using DTN forwarding protocol (line 37). Intermediary nodes follow the same process in algorithm 1 for $k-1$ times.

Because the LBS only knows the $pseudo_{id}$ of the incoming query, it reply to the $pseudo_{id}$. Only nodes on the request obfuscation path know the $real_{id}$ of the requester. The forwarding of reply will use the mapping between $pseudo_{id}$ to $real_{id}$to forward the reply back to the requester. Due to the asynchronous operation of forwarding in *OppMNet*, it is possible that the LBS-reply never reaches any of the nodes $k-1$ on the original obfuscation

Algorithm 1: SENDING PRIVATE LBS QUERY USING *LPAF*.

```
1: User of node ($N_r$) sets the anonymity requirement of the LBS application ($App_r$) running on the node $N_r$ in
   terms of ($P_{low}$/$P_{high}$, $k$, $P_c$) that communicate with $LBS_{srvr}$ at different locations $l_r$.
2: $M_{obf} \leftarrow \phi$, Generate node $pseudo_{id}$
3: for all $m$ such that ($m \in M$) do
4:    if $h_r(m) > 0$ then
5:       $M_{obf} \leftarrow M_{obf} \cup \{m\}$
6:    else
7:       remove $real_{id}$, path from $m$
8:       $M \leftarrow M \cup \{m\}$
9:    end if
10: end for
11: if $M_{obf} = \phi$ then goto line 37 end if
12: for all $N_i$ such that ($N_i \in Nghs_{curr}$) do
13:    for all $m$ such that ($m \in M_{obf}$) do
14:       if $(path(m) \cap \{N_i\}) \neq \phi$ then goto line 12 end if
15:       Calculate social distance $sd_{ci}$, multi-hop Markov proximity $MMP_{id}$
16:       if $sd_{ci}(P_c) > P_{high}$ then
17:          $OUT_n \leftarrow OUT_n \cup \{N_i, m, sd_{ci}, MMP_{id}\}$
18:          $M_{obf} \leftarrow M_{obf} \setminus \{m\}$
19:       else if $sd_{ci}(P_c) > P_{low}$ and $MMP_{id} > MMP_{cd}$ then
20:          $OUT_n \leftarrow OUT_n \cup \{N_i, m, sd_{ci}, MMP_{id}\}$
21:          $M_{obf} \leftarrow M_{obf} \setminus \{m\}$
22:       end if
23:    end for
24: end for
25: Sort $OUT_n$ by message priority, $sd_{ci}$, $MMP_{id}$
26: for all $T_i$ such that ($T_i \in OUT_n$) do
27:    $path(m_i) \leftarrow path(m_i) \cup \{N_i\}$
28:    Forward $m_i$ to $N_i$ using DTN protocol
29:    Update $reverse\_addr_{table} \leftarrow reverse\_addr_{table} \cup \{< m_i, pseudo_{id}, real_{id}, time >\}$
30: end for
31: for all $m$ such that ($m \in M_{obf}$) do
32:    if $delay(m_i) > delay\_threshold$ then
33:       print $App_r$ can not send the query and maintain the user pre-set location privacy requirements.
34:       Drop message $(m_i)$
35:    end if
36: end for
37: for all $m$ such that ($m \in M$) do
38:    Forward $m$ using DTN forwarding
39: end for
40: $LBS_{srvr}$ receives the LBS-query which has a location query in ($pseudo_{id}$, $l_r$)
41: $LBS_{srvr}$ generates and forwards the LBS-reply through social forwarding and send reply ($pseudo_{id}$, reply)
```

path in its way back from the LBS. In this case, we use similar scheme as in [30], [31] which use group label in an attempt to reach the source of the query. The outbound node of the set of intermediate nodes, *i.e.*, last node number k on the obfuscation path, removes the identity of the query source (N_r) and replaces it with a social group label, before forwarding the query freely to any neighboring node (*i.e.*, without following the social criteria).

B. Distributed and Collaborative Measurements

Using k-anonymity attempts to reduce the quality of the location information either in space (spatial) or in time (temporal), hence prevent meaningful use by various LBSs, especially with low user density scenarios [7]. Our idea of utilizing the social graph as an overlay to provide the basis for a trust model between *OppMNet* users, where friends in social network trust each other to obfuscate their location. Users build a collaborative obfuscation path over asynchronous social-based k hop intermediate nodes, and hence form a temporal location-privacy obfuscation path between the source and the LBS where intermediaries are the users' own social acquaintances. As shown in algorithm 1, nodes require social awareness to select next-hop (section III-B1), and location awareness (section III-B2) to forward queries towards the destination LBS while building the obfuscation path.

1) Social Awareness

We utilize social-links (*i.e.*, friendship relations between users) in *OppMNet* to maintain the users' location-privacy. In this section, we consider the general case where social-relations are defined in terms of users' *social-profile* similarity [10]. We structure the user's *social-profile* as an n-tuple *profile attributes*, and each attribute is assigned a different weight that reflects the relevance of this attribute to the whole profile. For example, affiliation attribute is given 40% importance, where the gender is given 5% importance in matching the two nodes profiles. The source of the query, or an intermediary node obfuscating it, calculates the *social-distance* between the profiles of itself as i and the possible next-hop as j, under criteria P_c (sd_{ij}), which is defined as the weighted number of matching profile attributes between the profiles of the two nodes (p_i and p_j), and it is calculated as:

$$sd_{ij}(P_c) = \frac{M\left(p_i, p_j, P_c\right)}{C\left(p_i\right)} \tag{1}$$

Where the function $M()$ searches for the matching degree between the two given social profiles using criteria P_c (as defined in table I). $C()$ returns the number of attributes in the given profile. p_n represents the social profile of a node n. Matching degree reflects both the number of matched profile-attributes and the weight each of these attributes in the social profile.

2) Location Awareness — Markov Proximity

We represent the location-visits using first-order Markov chains as a directed weighted graph between map regions. Each map is denoted as a region R^+ which is split into smaller regions with a total of $n \times m$ squares. Each time a node moves from one region a to another region b, the weight of the edge connecting the two vertexes $(a, b \in R^+)$ on the graph is denoted by E_{ab}. The weight of each directed edge reflects the probability that the node is going to move from the source-vertex representing location $l_x(\forall l_x, x \in a)$ to the destination-vertex representing location $l_y(\forall l_y, y \in b)$. Probability generation function is using a first-order Markov model, in which the probability $P(b \mid a)$ that the node will move to region b given that it is currently at region a is:

$$P_j\left(b|a\right) = \frac{E_{ab}}{\sum_{l\in R^+} E_{al}} \tag{2}$$

Then a node j is able to calculate its "Markov Proximity" (MP) as the probability of moving from the current location l_c to the next location l_n as the value $P_j\left(n \mid c\right)$.

Assuming that two socially related nodes i and j meet at location l_j, and i has a message that it wishes to obfuscated to reach node d (the untrusted LBS). Then node i will send a query to j asking for the estimate MP_{jd}. Note that at the time of encounter, the two nodes are within Bluetooth communication area of about 10 meters in radius, so we assume that the current location is the same. MP_{jd} query can be piggy-packed on the normal transmitted data messages or inside the hello message as its size is negligible compared to the payload size. MP_{jd} is calculated using equation (3) and is tracked locally by each node and only shared with each node social group. Nodes only respond and calculate MP_{jd} for their current location in order to prevent external exposure of location information.

$$MP_{jd} = \begin{cases} 0 & \text{for} \qquad v_{jd} \in \emptyset \\ P_j(l_d|l_j) & \qquad otherwise \end{cases} \tag{3}$$

Where v_{jd} represents the directed edge connecting the two vertexes corresponding to locations l_j and l_d, and the direction of movement is from l_j to l_d.

3) Distributed multi-hop location-obfuscation prediction

We now consider the multi-level prediction problem, where two meeting nodes wish to calculate the prediction, not only considering the probability of the other node visiting the LBSs location directly, using the combined probability of all their respective contacts. For example, node A meeting two nodes B and C, where A is interested in location obfuscation service towards LBS. A asks both nodes B and C to provide their probability for visiting the LBS considering all their contacts. If node B rarely visits LBS location directly, but its contacts do with higher probability than that of node C, then node A will chose node B as the next-hop and not C. More formally, let the set of nodes that B encounters over time called N_b, then we define the "Multi-hop Markov Proximity" (MMP) as:

$$MMP_{jd}\left(A,B\right) = \begin{cases} 0 & , \qquad N_b \in \emptyset \\ \alpha \times MP_{jd} + \beta \times \frac{\sum_{i \in N_b} MP_{l_i d} \times MP_{j l_i}}{N_b} & , \quad otherwise \end{cases} \tag{4}$$

Where α and β are weighting factors. The values of both will be adaptively determined depending on the stage of the obfuscation phase. Let k be the privacy requirements indication the obfuscation-path length, and h_r the remaining number of obfuscation hops to achieve this privacy requirement. This allows intermediate nodes to use multi-hop prediction using localized view of their next-hops about their various contacts, instead of relying on greedy and naive single-node prediction. During the obfuscation phase α and β are defined as follows:

$$\alpha = \begin{cases} \frac{k-h_r}{k} & , \quad h_r < k, k \neq 0 \\ 1 & , \qquad h_r = 0 \end{cases} \tag{5}$$

$$\beta = \begin{cases} 1-\alpha & , \quad h_r < k \\ 0 & , \quad h_r = 0 \end{cases} \tag{6}$$

From the above equations (5) and (6), we can see that the values of both α and β are dependent on k and h_r. The value of α is increasing with increasing k, and it decreases with fewer remaining obfuscation hops (*i.e.*, h_r). The value of β is the complement of α to one, which allow nodes to give more weight to the first term in equation (4) as queries are about to achieve their anonymity requirement of k hops. While, at the start of the obfuscation process, the protocol gives more weight to using intermediate nodes to reach the LBS which subsequently offer better privacy. When a query has passed k obfuscation hops (*i.e.*, $h_r = 0$), intermediate nodes give all the weight to their next-hops direct MP prediction (or $\alpha = 1$ and $\beta = 0$) in equation (4).

TABLE II: Analytical Model: List of parameters and their meaning.

Parameter	Meanings
R^+	The mobility area $L \times L$ representing the spatial dimension where nodes are moving.
N	The set of all the nodes in the network.
S	The social group which is a subset of network nodes that are socially related (*i.e.*, nodes with common social links) $S \subseteq N$.
S_a	The set of social friend of node a.
λ	The spatial Poisson process rate of event.
$R\left(v\right)$	Communication area of node v with a transmission radius $r : R = \pi r^2$.
$\#R\left(v\right)$	The number of nodes inside the communication area R of node v.
$E\left(v\right)$	The set of intermediate nodes (hops) that form the multi-hop path for node v to reach the destination using store-carry-forward paradigm.
$\llbracket l_1, l_2 \rrbracket$	The Euclidean distance between two location points $l_1, l_2 \in R^+$.
$P^1\left(\#(R\left(v\right))\right)$	The probability that a node v has at least one neighbor.
v_a	Average node speed in the network.
v_{rel}	The relative speed between two nodes.
k	Social obfuscation-path length, or anonymity level requirement.
$\varphi_e\left(t\right)$	The ratio of nodes receiving a packet after time t.

IV. *LPAF* ANALYTICAL ANALYSIS

A. Analytical Model

In this section we model our protocol using probabilistic analytical model. A number of models have been proposed for modeling opportunistic communication in ad-hoc networks [32]–[34]. For the purpose of our study, we focus on driving a model that is not only temporal with respect to timing of event, but also spatial model (incorporating location knowledge). We use a spatial Poisson process [35], also known as spatial point process, with an associated rate of events λ. Spatial Poisson process is useful for modeling ad-hoc wireless networks [36] in general, and especially location-dependent protocol. This is because the process is based on the Poisson distribution and has the additional spatial dimension that provides a mathematical means for modeling the distribution of nodes over the network area. Using the spatial dimension, we can calculate the probability of a node being in a certain location once the system has reached a steady-state, and the probability of nodes being neighbors. Let N be the set of nodes moving within a region on the map denoted as R^+, R^+ is an $L \times L$ square area. Table II shows a list of model parameters along with their interpretation.

In our study, we model communication as a Poisson process. Communication can occur only between two nodes that are within the communication area of each other. Assuming all nodes have the same communication area R and that communication can take place without interference from either physical obstacles or on the wireless channel. Let two nodes $a, b \in N$, and let the location of the two nodes at time t be $l_{a,t}, l_{b,t} \in R^+$. The two nodes can communicate directly if and only if $\llbracket l_1, l_2 \rrbracket \leq r$. We define the opportunistic communication probability according to the spatial Poisson theory. Let $P\left(\#(R\left(v\right))\right)$ be the probability that i nodes exist within the communication area $R\left(v\right) \subset R^+$. $P\left(\#(R\left(v\right)) = i\right)$ is formally defined as follows:

$$P\left(\#(R\left(v\right)) = i\right) = \frac{e^{-\lambda \pi r^2}\left(\lambda \pi r^2\right)^i}{i!}, \lambda = \frac{|N|}{L^2} \tag{7}$$

Subsequently, we define the probability that a node v has at most i nodes within the communication area $R\left(v\right) \subset R^+$ as follows:

$$P\left(\#(R\left(v\right)) \leq i\right) = \sum_{x=0}^{i} P\left(\#(R\left(v\right)) = x\right) \tag{8}$$

Let $P^1\left(\#(R\left(v\right))\right)$ be the probability that a node v has at least one neighbor, *i.e.*, the two nodes are able to communicate with each other. According to the Poisson theory, the inter-arrival distance between events is i.i.d (independent and identically distributed) random variable. This is because the Poisson model has a memorylessness property causing the number of arrival events at any location in R^+ occurring after time t to be independent from the number of events occurring before time t.

$$P^1\left(\#\left(R\left(v\right)\right)\right) = P\left(\#\left(R\left(v\right)\right) \geq 1\right)$$

From the probability theory, we know that $\sum_{i\geq 0} P\left(\#\left(R\left(v\right)\right) = i\right) = 1$, we can rewrite it as $P\left(\#\left(R\left(v\right)\right) \geq 1\right) + P\left(\#\left(R\left(v\right)\right) = 0\right) = 1$. We can then substitute $P\left(\#\left(R\left(v\right)\right) \geq 1\right) = 1 - P\left(\#\left(R\left(v\right)\right) = 0\right)$ as follows:

$$\begin{aligned} P^1\left(\#\left(R\left(v\right)\right)\right) &= 1 - P\left(\#\left(R\left(v\right)\right) = 0\right) \\ &= 1 - e^{-\lambda\pi r^2} \end{aligned} \tag{9}$$

For the purpose of developing the *LPAF* model, we need to focus on the obfuscation path rather than the immediate neighboring nodes. We are interested in modeling the multi-hop communication path used for obfuscation. Let $P\left(\left|E\left(v\right)\right| \geq i\right)$ be the probability of node v forming i multi-hops path of successive encounters (*i.e.*, using any available nodes) using Epidemic protocol. This is equal to the probability that each node from the set of the i intermediate nodes has at least one successful encounter at its first hop (because the inter-arrival of events in the Poisson process is i.i.d). $P\left(\left|E\left(v\right)\right| \geq i\right)$ can be formally defined as follows:

$$\begin{aligned} P\left(\left|E\left(v\right)\right| \geq i\right) &= \prod_i P\left(\#\left(R\left(v\right)\right) \geq i\right) \\ &= \left(P^1\left(\#\left(R\left(v\right)\right)\right)\right)^i \qquad \text{from (9)} \\ &= \left(1 - e^{-\lambda\pi r^2}\right)^i \end{aligned} \tag{10}$$

When the node's communication area is very small compared to the dimensions of the network, *i.e.*, $R \ll L$, [33] shows that we can obtain an approximate rate of encounters (or inter-meeting time) for any node in the network (p). This rate of encounter can be obtained using the following formula:

$$p = c\frac{v_{rel}r}{L^2} \quad , \quad c : \text{mobility constant} \tag{11}$$

$$v_{rel} = \frac{c}{\pi v_a^2}\int_0^{2v_a} \frac{x^2}{\sqrt{1-\left(1-\frac{x^2}{2v_a^2}\right)^2}}dx \tag{12}$$

where c is a mobility model related constant. For the purpose of our analytical evaluation, we consider only the random direction model with $c = 1$ and epidemic packet forwarding. For more complex real-life validation, we evaluate *LPAF* through intensive simulation in different scenarios. And v_{rel} is the relative speed between nodes and can be calculated assuming an average stable node speed in the network (v_a). According to [32], we can obtain the number of nodes $I\left(t\right)$ that has received a packet after time t assuming a single initial source as follows:

$$I\left(t\right) = \frac{|N|}{1 + e^{-p|N|t}\left(|N| - 1\right)} \tag{13}$$

Where p is the rate of encounters as calculated in (11). So, we can then calculate the ratio of nodes that has received a packet after time t as $\varphi_e\left(t\right)$ for the Epidemic protocol:

$$\varphi_e\left(t\right) = \frac{I\left(t\right)}{|N|} = \frac{1}{1 + e^{-p|N|t}\left(|N| - 1\right)} \tag{14}$$

In the case of the proposed protocol *LPAF*, the probability that a node communicates is restricted during the obfuscation phase to only within its social community. So a node has less probability of forwarding packets due to its privacy constraints. We develop a new probabilistic model to reflect this behavior. Consider two nodes a and b, where node a wants to obfuscate a query by sending it through node b. Let event A be that node a has at least one neighbor b (*i.e.*, at least a and b are within the communication area of each other), and Let B be the event that one of these neighbors is from the social group of a (*i.e.*, b is a social friend of a). More formally, let A: $[\![l_{a,t}, l_{b,t}]\!] \leq r$ and B: $b \in S_a$. Let $P\left(A \mid B\right)$ be the conditional probability that event A will occur given the knowledge that event

B occurred. Note that since we are using random direction model, the probability of encounter between socially connected nodes is not affected, *i.e.*, it is shown in the literature that human encounter are not random, and that friends meet each other more often than they would meet someone who is a complete strangers [37], [38]. Taking that into consideration, we emphasis that the analytical study presents the worst-case view of *LPAF* performance due to the lack of this connection between the social layer and nodes' mobility. This can be expressed as follows:

$$P(A \mid B) = P(A) \tag{15}$$

Let $P_f\left(i\right)$ denote the probability that a node v meets i successive disjoint friend nodes (*i.e.*, different nodes belonging to node v social group S_v). The probability $P_f\left(i\right)$ can be expressed as:

$$\begin{aligned} P_f\left(\left|E\left(v\right)\right| = i\right) &= \prod_{x=0}^{i-1} \frac{\left(\left|S\right| - 1 - x\right)}{\left|N\right|} && ,0 < i < \left|S\right| \\ &= \frac{\left(\left|S\right| - 1\right)!}{\left|N\right|^i \times \left(\left|S\right| - i - 1\right)!} && ,0 < i < \left|S\right| \end{aligned} \tag{16}$$

Now, we would like to calculate the probability that a node v will successfully communicate under a privacy requirement of one hop obfuscation. We use the set of intermediate disjoint nodes from the multi-hop path $E\left(v\right)$ from node v to the destination, and not the set of nodes within the transmission range of v, because we are calculating the probability of forming a multi-hop path, rather than the nodes immediate neighborhood (*i.e.*, one hop). To this end, we define two events A and B as follows:

A: $\{$*meeting at least one node* $\}$

B: $\{$*a node is a friend (*i.e.*, belongs to the source's social group)*$\}$

Using probability theory, we can find the probability $P_s\left(\left|E\left(v\right)\right| = 1\right)$ that a node v will successfully communicate under a privacy requirement of one hop obfuscation can be expressed as follows:

$$\begin{aligned} P_s\left(\left|E\left(v\right)\right| = 1\right) &= P\left(A \cap B\right) \\ &= P\left(A \mid B\right) \times P\left(B\right) && \text{from (15)} \\ &= P\left(A\right) \times P\left(B\right) \end{aligned} \tag{17}$$

The first operand in (17) is the probability that a node v will encounter at least one node. The second operand can be calculated using equation (16), as P_f when $i = 1$.

$$\begin{aligned} P_s\left(\left|E\left(v\right)\right| = 1\right) &= P\left(\left|E\left(v\right)\right| \geq 1\right) \times P_f\left(\left|E\left(v\right)\right| = 1\right) \\ &= P\left(\#\!\left(R\left(v\right)\right) \geq 1\right) \times P_f\left(\left|E\left(v\right)\right| = 1\right) \\ &= P^1\left(\#\!\left(R\left(v\right)\right)\right) \times P_f\left(\left|E\left(v\right)\right| = 1\right) \end{aligned} \tag{18}$$

Using definitions from equations (9) and (16):

$$\begin{aligned} P_s\left(\left|E\left(v\right)\right| = 1\right) &= \left(1 - e^{-\lambda\pi r^2}\right) * \frac{\left(\left|S\right| - 1\right)!}{\left|N\right|^1 * \left(\left|S\right| - 2\right)!} \\ &= \left(1 - e^{-\lambda\pi r^2}\right) * \frac{\left(\left|S\right| - 1\right)}{\left|N\right|} \end{aligned} \tag{19}$$

From the definition of *LPAF*, a node v needs to form an i hop obfuscation path using i disjoint nodes for the communication with an LBS to be successful (*i.e.*, in *LPAF* case, the nodes are more constrained by the privacy requirements and will not be able to forward packets as freely as when using Epidemic). Let events A, B denote the following:

A: $\{$*meeting at least one node for* i *successive hops*$\}$, B: $\{i$ *nodes belong to the source's social group*$\}$

Further, let $P_s\left(\left|E\left(v\right)\right| \geq i\right)$ be the probability that a node v will successfully communicate over i hops obfuscation-path. $P_s\left(\left|E\left(v\right)\right| \geq i\right)$ is defined formally as follows:

$$P_s\left(\left|E\left(v\right)\right| \geq i\right) = P\left(A \cap B\right) = P\left(A\right) * P\left(B\right) \qquad ,0 < i < \left|S\right|$$

Let us focus only on the event A, the probability of forming multi-hop path using i nodes can be obtained using (10). The second operand which represent the probability that event B occurs can be obtained using (16).

$$\begin{aligned} P_s\left(\left|E\left(v\right)\right| \geq i\right) &= \left(P\left(\left|E\left(v\right)\right| \geq i\right)\right) * P_f\left(\left|E\left(v\right)\right| = i\right) \\ &= \left(1 - e^{-\lambda \pi r^2}\right)^i \times \frac{\left(\left|S\right| - 1\right)!}{\left|N\right|^i * \left(\left|S\right| - i - 1\right)!} \end{aligned} \tag{20}$$

From the definition of *LPAF*, it is important to note that we are interested in the events where i nodes will encounter each other following the DTN store-carry-forward paradigm. Hence, event A relates to the nodes' ability to form i multi-hop path to reach the destination. This event is quite different from the event of having a neighborhood that has i neighbors. Substituting into equation (20), the probability that a node v will successfully communicate over i hops obfuscation-path can be expressed as follows:

$$P_s\left(\left|E\left(v\right)\right| \geq i\right) = \begin{cases} \left(1 - e^{-\lambda \pi r^2}\right)^i \times \prod_{x=0}^{i-1} \frac{\left(\left|S\right| - x - 1\right)}{\left|N\right|} & , \quad 0 < i < \left|S\right| \\ 0 & , \quad otherwise \end{cases} \tag{21}$$

Let γ be the ratio between the probability that a node v successfully communicate using *LPAF* over i hops obfuscation-path, and the probability that a node can form i hops path using any nodes, *i.e.*, without obfuscation restrictions using Epidemic routing. We define γ as follows:

$$\gamma = \frac{P_s\left(\left|E\left(v\right)\right| \geq i\right)}{P\left(\left|E\left(v\right)\right| \geq 1\right)^i} \tag{22}$$

Using equations (9) and (21), γ can be expressed as follows:

$$\begin{aligned} \gamma &= \frac{\left(P^1\left(\#\left(R\left(v\right)\right)\right)\right)^i * P_f\left(\left|E\left(v\right)\right| = i\right)}{\left(P^1\left(\#\left(R\left(v\right)\right)\right)\right)^i} \\ &= P_f\left(\left|E\left(v\right)\right| = i\right) \\ &= \frac{\left(\left|S\right| - 1\right)!}{\left|N\right|^i * \left(\left|S\right| - i - 1\right)!} \end{aligned} \tag{23}$$

For the proposed protocol *LPAF*, we can then calculate the number of nodes that received a query after time t assuming a single initial source $I_s\left(t\right)$. We substitute γ into equation (13) taking into consideration the value calculated in (23):

$$I_s\left(t\right) = \frac{\left|N\right|}{1 + e^{-p\gamma\left|N\right|t}\left(\left|N\right| - 1\right)} \tag{24}$$

Let φ_s be the ratio of nodes that has received a query after time t using *LPAF*. φ_s is defined as follows:

$$\varphi_s\left(t\right) = \frac{I_s\left(t\right)}{\left|N\right|} = \frac{1}{1 + e^{-p\gamma\left|N\right|t}\left(\left|N\right| - 1\right)} \tag{25}$$

B. Analytical Model Evaluation

In this section, we propose an analytical model that enables the study of *LPAF*, in terms of logical and quantitative relationships, in order to see how the model reacts to different parameters such as privacy-levels (k) and social group sizes ($\left|S\right|$) and what impact each has on quality of service such as delivery ratio and delay.

Hence, we perform extensive simulations, using the proposed analytical model for *LPAF* to study the feasibility/limits of conceiving obfuscation path for different privacy-levels (k) and social group sizes ($\left|S\right|$). We show that using the proposed *LPAF*, an obfuscation path can still be constructed even at high privacy-levels (k), at the expense of lower quality of service represented by lower delivery ratio and higher delays. It is worth mentioning that although we show that the protocol can work in *OppMNet* at extreme situations (characterized by low social groups sizes $\left|S\right| \leq 30\%$, and high privacy-level of $k = 8$), these situations are not likely to exist in real live and present a worst-case scenario for *LPAF*. This is because the analytical model does not consider the network topology and how it is dynamically influenced by the social ties.

TABLE III: List of parameters and their values.

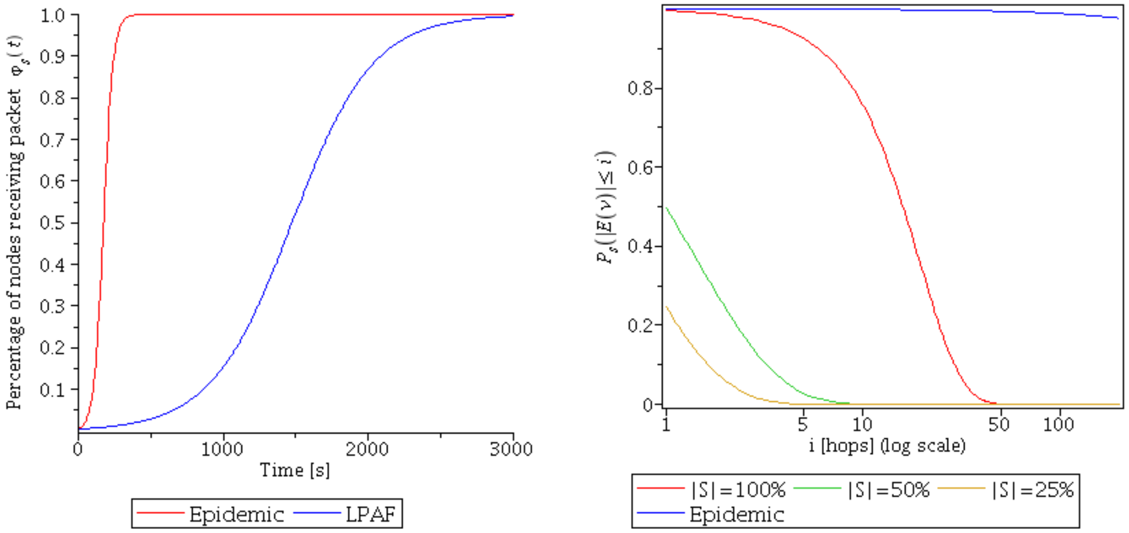
Parameter	**Meanings**
L^2	$1000m \times 1000m$
$\|N\|$	200 nodes
$\|S\|$	25–200
r	120m
v_1	1 m/s
k	2–8 hops

To validate the proposed analytical model for *LPAF*, we start by comparing its performance to Epidemic forwarding as in figure 2a. *LPAF* has 3 hops obfuscation path at social group size of 50% of the total nodes in the network. We choose a moderate social group size using information obtained from experiments shown in figure 2b, and the number of obfuscation hops ($k = 3$) was chosen at the number of hops where the protocol achieves about 25% probability to reaching all nodes, or at 50% of the protocol maximum achievable probability (P_s). The results shown assume a single initial source sending one query. We can see that Epidemic protocol is capable of disseminating the packet quicker, and that *LPAF* manages to follow with a slight delay due to the higher location anonymity provided.

In order to understand the impact that the social group size $|S|$ has on the probability of forming obfuscation-path, we conduct the experiments shown in figure 2b. The figure shows the probability of forming obfuscation-path of different length i for an increasing trend of the social group size $|S|$, and we compare them to the Epidemic probability of forming the same number of hops i using any node even several times. The size of social group ranged from 50 nodes to 200 nodes, or 25% to 100%. We can see a decreasing probability of forming a successful delivery with longer obfuscation path (as i increases), which is due to the extra restriction on finding socially related nodes (of course, not yet used). Finding such friend-nodes is affected primarily by the social group size $|S|$, among other factors such as nodes average speed and communication area R, and the total area of R^+. We can see that increasing $|S|$ allows for an increased probability of reaching higher privacy. As $|S|$ approaches $|N|$, *i.e.*, theoretically the whole network is one social group, the probability falls as i increases, and this is because intermediate nodes are selected only once. This validates the privacy-restrictions developed through the analytical model for obfuscation based communication using *LPAF*. Interestingly, this shows that the probablity of forming obfuscation path drops logarithemically with respect to the number of obfuscation hops i.

We can see the probability to form the same i multi-hop path using Epidemic protocol in figure 2b as derived using equation (10). It is considerably much higher compared to *LPAF*, and it is approaching 0.99 for $i < 15$ hops. This is because Epidemic protocol can form any multi-hop path and re-use nodes previously been on that path, *i.e.*, nodes that has participated in the query forwarding before. While for *LPAF*, a node cannot re-forward a query to a node that has obfuscated it before, because this does not increase the source's privacy (*i.e.*, the anonymity set does not increase).

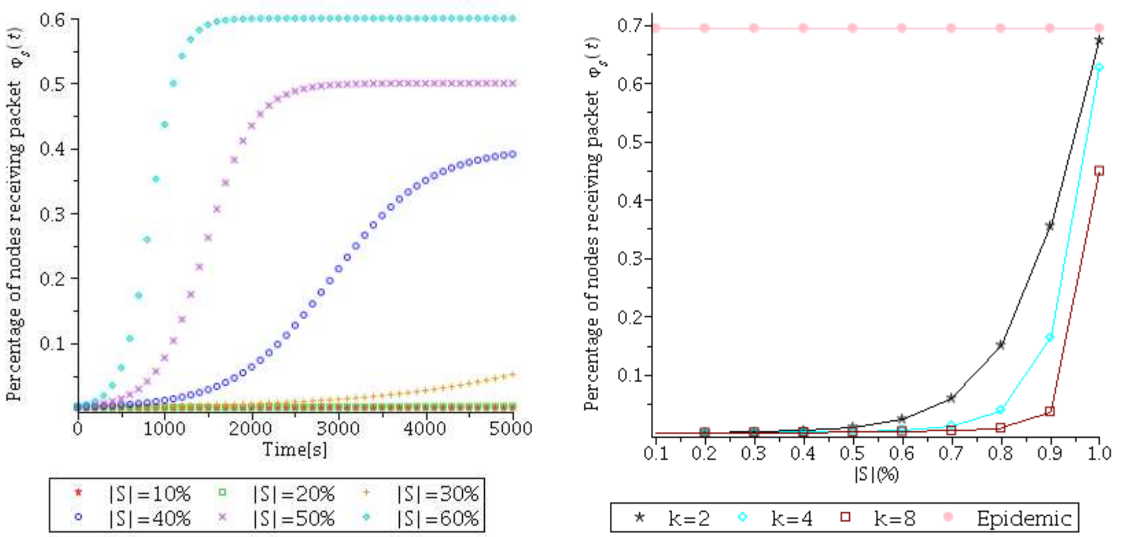
We evaluate the *LPAF* performance convergence over time with respect to social group size $|S|$ (in figure 3a) and location-privacy requirements k (in figure 3b). Figure 3a shows *LPAF* performance under different conditions of social group sizes ranging between 10% and 60% of the total number of nodes $|N|$ under moderate location privacy (k=3). The graph shows that *LPAF* manages to disseminate the packets to all nodes in the social group. For example, about 40% of the social group nodes receive packets in less than 1600 seconds, for average social group size ($|S| \geq 50\%$). We can see that as the social group size grow, the percentage of nodes receiving the



(a) Percentage of nodes (y-axis) that receive a packet after time t (x-axis). $\varphi_e(t)$ for Epidemic and $\varphi_s(t)$ for *LPAF* using ($k = 3$ and $|S| = 50\%$).

(b) Probability (y-axis) to form various obfuscation path length i (x-axis): under different number of social group size.$P_s(|E(v)| \geq i)$ for *LPAF* and $P(|E(v)| \geq i)$ for Epidemic.

Fig. 2



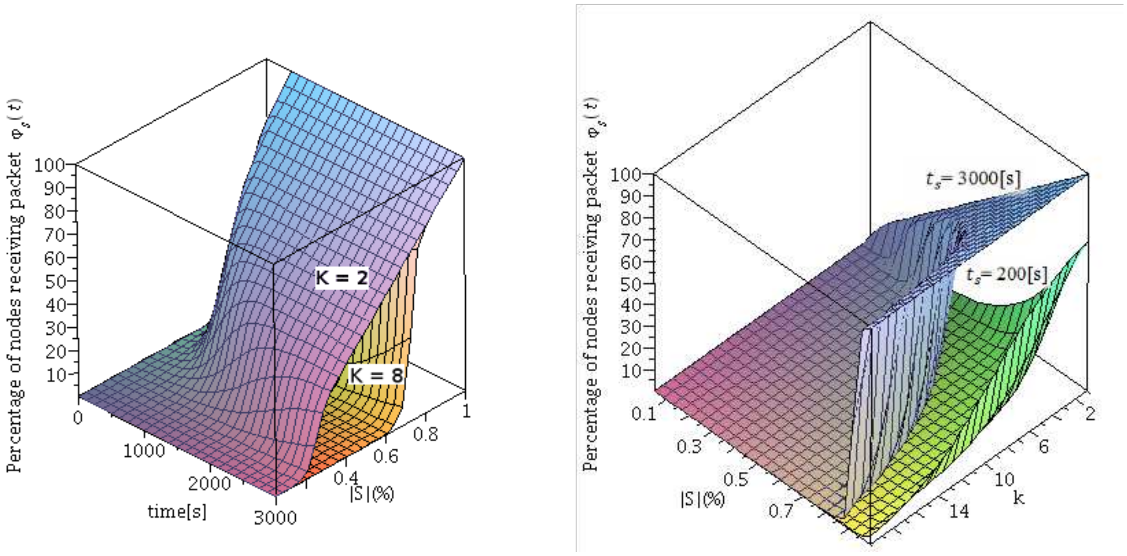
(a) Percentage of nodes (y-axis) that receive a packet after time t (x-axis) with different values of $|S|$.
$\varphi_e(t)$ for Epidemic and $\varphi_s(t)$ for *LPAF*.

(b) Percentage of nodes (y-axis) that receive a packet for various social group size (x-axis).
$\varphi_e(t)$ for Epidemic and $\varphi_s(t)$ for *LPAF*, k range between 2 and 8 hops.

Fig. 3

packet increases sharply, *i.e.*, the time required to reach the group gets shorter as more friends participate in the forwarding. This is because the probability of meeting socially related nodes becomes higher, which allows the packet to be forwarded faster and to more nodes than in scenarios with small social group size.

In order to choose the best suitable protocol configuration parameters in a given envionrment, we need to combine the main parameters to show their impact on the protocol performance (*e.g.*, percentage of nodes receiving a packet $\varphi_s(t)$). Taking both time and $|S|$ as the protocol two degrees of freedoms is shown in figure 3b, while $|S|$ and k is shown in figure 4a. Figure 3b shows *LPAF* theoretical performance under various location-privacy



(a) $|S|$ on x-axis, time on y-axis and $\varphi_s(t)$ on the z-axis at k=2 and 8.

(b) $|S|$ on x-axis, k on y-axis and $\varphi_s(t)$ on the z-axis at t_s=200 and 3000[s].

Fig. 4: *LPAF* performance comparing the percentage of nodes receiving a query $\varphi_s(t)$.

requirements. The function $\varphi_s(t)$ is sampled at $t = t_s$ for comparing performance, where $(t_s = 200[s])$ is chosen because it is the time where the protocols are in-transition with respect to $|S|$ as shown in figure 3a. We can see that $\varphi_s(t)$ generally increases as $|S|$ increases, and that validates the same impact seen in figure 3a. The figure also shows that as k increases, the sloop of the curve becomes steeper (*i.e.*, the social group size has less impact on the protocol performance as k increases). Moreover, for the same value of $|S|$, we see that as k increases, the performance drops (*i.e.*, the curve gets lower). For example, at k=2 and k=4, we see that when the social group size is about 85% of the total number of nodes in the network ($|S|$=85%), the percentage of reached social nodes is in the range of 30% and 10% at $t = t_s$ respectively.

Figure 4a shows a three dimensional performance comparison of two different location privacy levels from low to high (k=2 and 8). The figure clearly shows the impact of the privacy level and social-group size on the delay and percentage of nodes receiving the query. On the x-axis, we show social group size ($|S|$), the time (t) on the y-axis and the result achieved ($\varphi_s(t)$). We can see that the achieved level of packet penetration to other nodes at some social group size can be easily obtained.

Using 4a, we can draw plane perpendicular to the x-axis at a point equal to a given social group size ($|S|$) in the network, and we obtain a 2-D plane showing the various expected performance over time. From the figure, we observe a general increase in the percentage of reached nodes as we move to a bigger group size, which is consistent with both figures 2b and 3a. We can see a faster convergence (less time) to reach more nodes as k decrease (less privacy). This figure can be used to guide the user selection of the level of location privacy (k) under specific network conditions, by determining the expected $\varphi_s(t)$ as the average required user data utility and user tolerated level of delay.

For example, for a network with a known social group size $|S|$ equal to 50% and a user data utility requirement to reach 40% of the nodes within 500 seconds, the user will be able to achieve this combination under privacy-level k=2, but not at a higher privacy-level of k=8. On the other hand, if the number of social group size can be controlled, then the above graph can be used to show the required percentage of social nodes to achieve a certain privacy level. For example, if a user wishes to maintain privacy level of k=2 and a data utility requirement to reach 50% of the nodes within 1000 seconds, then the required percentage of social group nodes must be more than 60%.

Figure 4b shows the maximum achievable percentage of reached nodes after two different sample times ($t_s = 200[s]$ and $3000[s]$). The figure shows that as the number of obfuscation hops increase, there is a decline in the percentage of nodes receiving the packet at the same values of $|S|$. And as the social group size $|S|$ increase, we see an increase in the percentage of nodes receiving the packet. If the user needs to maintain the same percentage of nodes receiving the packet while having higher privacy, then the social group size $|S|$ needs to be increased to provide the additional privacy without degrading the overall performance.

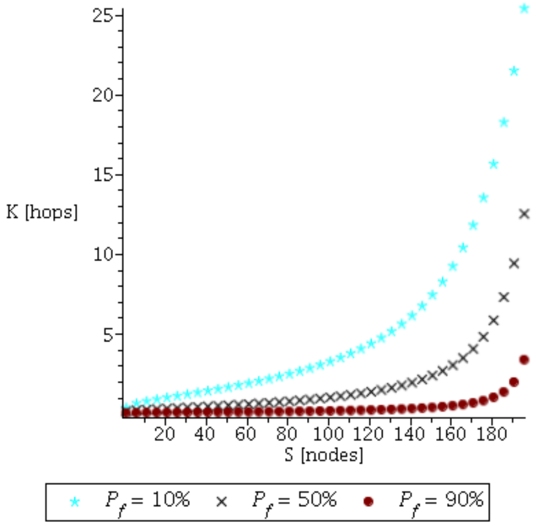


Fig. 5: The maximum number of obfuscation hops (k) (y-axis) for various social group size $|S|$ (x-axis), for k between 1 and 25 hops.

Figure 5 shows *LPAF* achievable privacy-level at various social group sizes in the network for three different probabilities for successfully encountering k friends (P_f). P_f, as defined in equation (16), directly affects the probability that a user will have successful communication with the LBS. The figure presents the trade-off between privacy-level (k) for different social group size $|S|$ in the networks, *i.e.*, the user have higher probability of successful obfuscation for higher k as $|S|$ increases. For example, a user who belongs to a social group of size 120 has 50% probability of successful communication for k range below 2 hops, while only 10% probability in case of higher k (from 4 to 5 obfuscation hops).

V. Performance Evaluation

We compare *LPAF* performance against *No_Privacy*, Oracle-based k-Anonymity [39] (*OBK*), Social-Only [39] (*SLPD*) and *ALAR* [40] protocols. We implement all privacy-preserving protocols as extensions to ONE simulator [41]. We track the performance across three dimensions: 1) Quality of service expressed as success ratio (*SR*) and end-to-end delay, 2) Quality of anonymization reflected in the achieved number of obfuscation hops (k), and 3) Energy efficiency tracked through the retransmission overhead.

In addition to using an urban scenario, we perform evaluation using real-world data trace of Bluetooth and Wi-Fi (*UIUC*) collected at the University of Illinois. For *ALAR*, we set the number of the alternative disjoint path (k=2), and the number of splits (S=2); similar to the settings used in [40]. Other simulation settings are summarized in table IV. Using these evaluation results, we confirm the analytical-model results in section IV-B for an increasing privacy-level (k) and different social group size ($|S|$).

A. Pseudo Simulation in Urban Scenario in Helsinki City-Center

For the purpose of initial evaluation, we experimented with a well known benchmark urban scenario for Helsinki city-center. We combine the Opportunistic Network Environment (ONE) simulator [41] probabilistic *Map-Based Movement Models* with Helsinki map to generate a semi-realistic urban scenario in Helsinki city-center. The objective of these set of experiments is to study the proposed privacy-preserving protocol in pseudo realistic city-center scenario and measure its effectiveness. In order to achieve a realistic experiment, a number of elements needed to be configured and used in ONE. Firstly, we use a real city-map (*Helsinki City Map*) that include details about roads, pedestrian-only streets and tram lines. Secondly, we use a more realistic map-based movement mode, compared to Random Way Point, where different nodes have different constraints moving around the city as it will be detailed in section V-A1.

Without loss of generality, we develop *LPAF* over a quota-based DTN forwarding protocol, namely *Spray And Wait (SnW)* [42]. We evaluate the social-based protocols (*SLPD*, *LPAF*) across varying social matching threshold, *i.e.*, users can select different values for P_{low}/P_{high} (as defined in table I). Users who desire more privacy set both P_{high} and P_{low} to a value close to 1, and vice versa. P_{low} is configurable and was set to 0.5 (50%) for all experiments. Location queries are sent every 20 seconds from a number of nodes belonging to a pool of 40 tourist scattered walking and querying locations around the city. Each member of this group has a defined social profile, and it is used to measure the social distance to other nodes (as explained in section III-B1). We assume that all scenarios have one LBS, and it is moving around the city. Each node coming in contact with the LBS records the current location where the encounter took place and shares it with other socially-related nodes. The nodes use the most-recent location reported by their neighbors as the current LBS location to update and share their *MMP* (as explained in section III-B2).

1) Network Model

We consider a network that has a set of *N* nodes that are actively relaying messages in the *OppMNet*s when requested. The selfishness or other malicious behaviors that would cause nodes to not participate in the forwarding process are out of the scope of this paper. Nodes have limited resources (buffer, energy) and will continue to forward messages as long as these resources are available. We consider heterogeneous node types with different resource levels (such as trams and pedestrians). We assume that trams are equipped with two types of wireless networking interfaces: high-power long range interface and another short range one.

There are many existing mobility models in the literature [43]. We have chosen a map-based mobility model, because it uses the underlaying map data in order to constrain node movement. Hence, nodes of different types follow distinctive rules to move from one location to another, *e.g.*, cars drive on roads or highway, and pedestrian use streets. This provides a number of advantages: 1) It avoids the drawbacks of some mobility models where nodes can move through buildings, or nodes meet at unrealistic points such as cars on pedestrians-only street; 2) It presents information about the underlying map and nodes' location in the network, and hence allows the evaluation of the proposed proximity-prediction efficiency; 3) LBSs typically utilize user locations and map information, and most available real-world traces do not contain spatial information. And while some of the exiting real-world traces can be used (such as UIUC trace used in section V-B), the evaluation is typically limited to a given set of locations in the trace (discrete and limited subset of all possible locations), and we cannot use/re-run the same experiment again with different nodes' locations to ensure statistical significance of the results. Running experiments over Helsinki scenario allows us to evaluate different random topology conditions and observes variations in results. By repeating these experiments many times, we study the statistical significance of any performance results. For the aforementioned reasons, we are complementing the evaluation using pseudo-simulation with UIUC real-world data trace scenario.

This work is not concerned with vehicular technology, and it focuses only on social opportunistic network. We utilize a set of heterogeneous mobile nodes (Pedestrians, Cars, and Trams) positioned on top of the Helsinki-city map, where cars and trams are used to ferry data between pedestrians when the possibility occurs. Communication between pedestrians is performed over *OppMNet* using Bluetooth, while communication between Trams, Cars is over Wi-Fi. The different nodes exhibit different characteristics, People carried mobile devices have smaller buffers compared to other mobile nodes. The *city-center scenario* consists of 80 pedestrians, 40 cars and 6 trams and 10 access points. Pedestrians communicate between them through Bluetooth, and receive information from cars, trams and access-points if within communication range using Wi-Fi.

All mobile elements of our scenario follow a movement model that attempts to closely match real-world, where cars drive on roads or highways, trams have predefined and fixed route that goes across the city, and pedestrians walk on streets. We define a set of Points of Interest, these are various locations around the map where mobile nodes visit more often, *i.e.*, with higher probability than other locations around the city. For example, people go to the city center more often than they visit urban areas of the city. In our scenarios, the points of interest represent areas such as workplace, business centers, market places, and similar locations. In our experiments, we do not model delay due to message collision caused by multiple nodes simultaneous transmission over the wireless medium, as such delays are considerably small compared to average delays due to nodes disconnections. Wireless channel transmission errors and congestion are also not considered in this study.

TABLE IV: Simulation parameters/values.

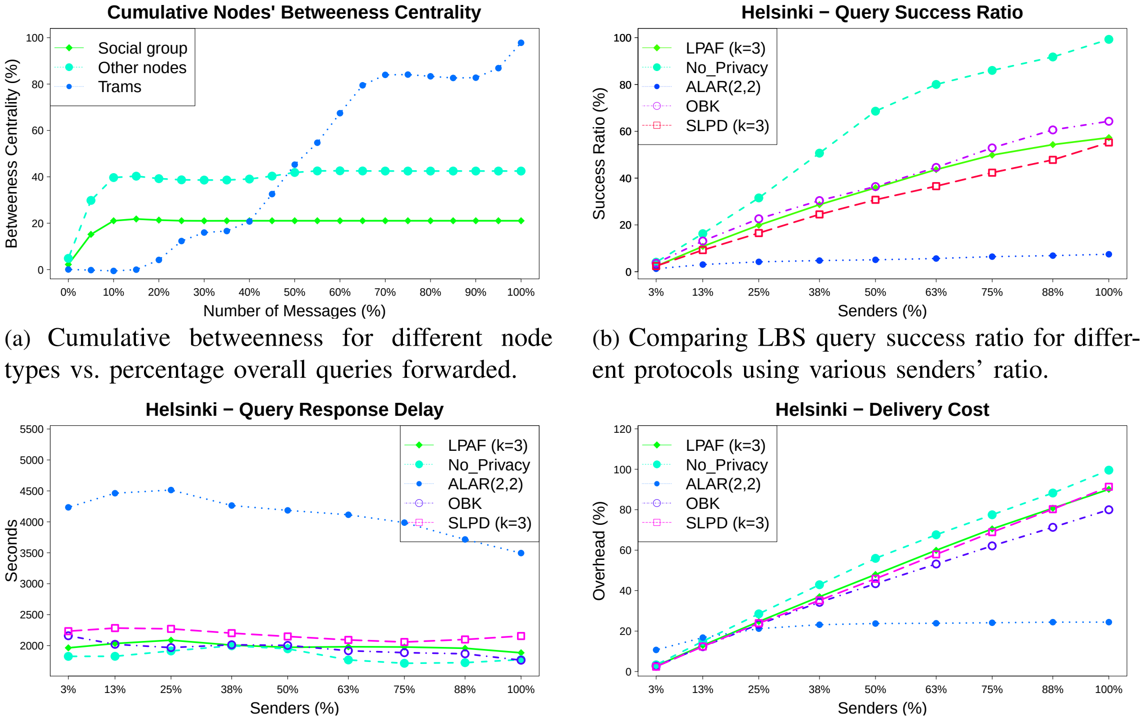
Key	Value	Key	Value
Simulation Time	43200 seconds	Query/Reply Size	512 bytes
Map Size (W x H)	4500 m × 3400 m	Publishing Rate	3 queries/minute
Map Squares (n by m)	10 × 10	Initial Query Replicas	10
Total no. of Nodes ($\|N\|$)	126	Buffer Size	Unlimited
Pedestrian Nodes	80	Pedestrian (Walking Speed)	1.8–5.4 Km/h
Car Nodes	40	Cars' Speed	10–50 Km/h
Trams	6	Opportunistic Interface	Bluetooth
Infrastructure Devices	12	Bluetooth Range	10 m
Users ($\|S\|$)	40	Movement Model (Pedestrian, Cars)	Shortest Path Map-Based
Number of LBSs	1	Movement Model (Trams)	Map-Based Route
Transmit Speed	2 Mbps		

Trams in the network are acting as data mules, *i.e.*, they are not considered as sources or destinations for any traffic. We assume, as in a typical city scenario, that these trams have contact with access-point when in communication range, or otherwise be disconnected. Trams are expected to execute *LPAF* protocol, and that users trust the trams' operator to handle their queries and perform obfuscation operations. If a node encounters a tram, it includes a profile matching-criteria and the required privacy-level for the query to obfuscate, and forward this information over to the tram. Tram extracts the privacy-related information and searches for other nodes using the profile matching-criteria to obfuscate this query.

2) Simulation Results in City-Center

We examine all protocols in scenarios where privacy-conscious users are assumed to not participate in the opportunistic network as shown in [2]. *LPAF* takes into account social relationship with neighbors performing obfuscation, as well as next-hop proximity-prediction (MMP) to the final destination. We evaluate *LPAF* under varying privacy-level requirements (k), varying number of source nodes accessing the LBS (senders), and different social group size ($|S|$). All experiments were conducted at high social matching threshold (*P*=1) for both *SLPD* and *LPAF*. Each point on the graph is the average of 50 runs, and in each run nodes were initially positioned randomly over the map. The senders are selected from a pool of highly social nodes (central nodes or hubs), starting by 1 sender or 3% of the pool size, then 5 (or 13% of the pool size), 10, 15 till 40 senders (or 100%). Increasing the number of senders increases the offered load to the network, hence show protocols' stability. Experiments examining the effect of the social group size were conducted using only one single sender selected at random at each run.

Betweenness centrality [44] measures the likelihood that a node lies on the paths linking different nodes in the network. This centrality measures the nodes' ability to control information flow between different parts of the network. In order to perform experiments with varying number of social nodes, we conducted graph analysis to find different centrality measures for different node types at each scenario. Figure 6ashows the cumulative "*Betweenness*" centrality for different node types in our scenarios, vs. the number of queries each group was responsible for forwarding measured without any privacy constraints. The group of nodes labeled "*Other nodes*" include vehicles and other pedestrian users not belonging to the social group. The figure shows that trams have the highest centrality and are involved in forwarding more queries compared to other groups. This is because trams are more powerful both in terms of their communication area and speed, hence, cover extended areas of the map. All results are normalized, *i.e.*, represented as a relative to the maximum achievable during the different simulation runs.



(a) Cumulative betweenness for different node types vs. percentage overall queries forwarded.

(b) Comparing LBS query success ratio for different protocols using various senders' ratio.

(c) Comparing LBS query average delay for different protocols using various senders' ratio.

(d) Retransmissions overhead for different protocols using various senders' ratio

Fig. 6

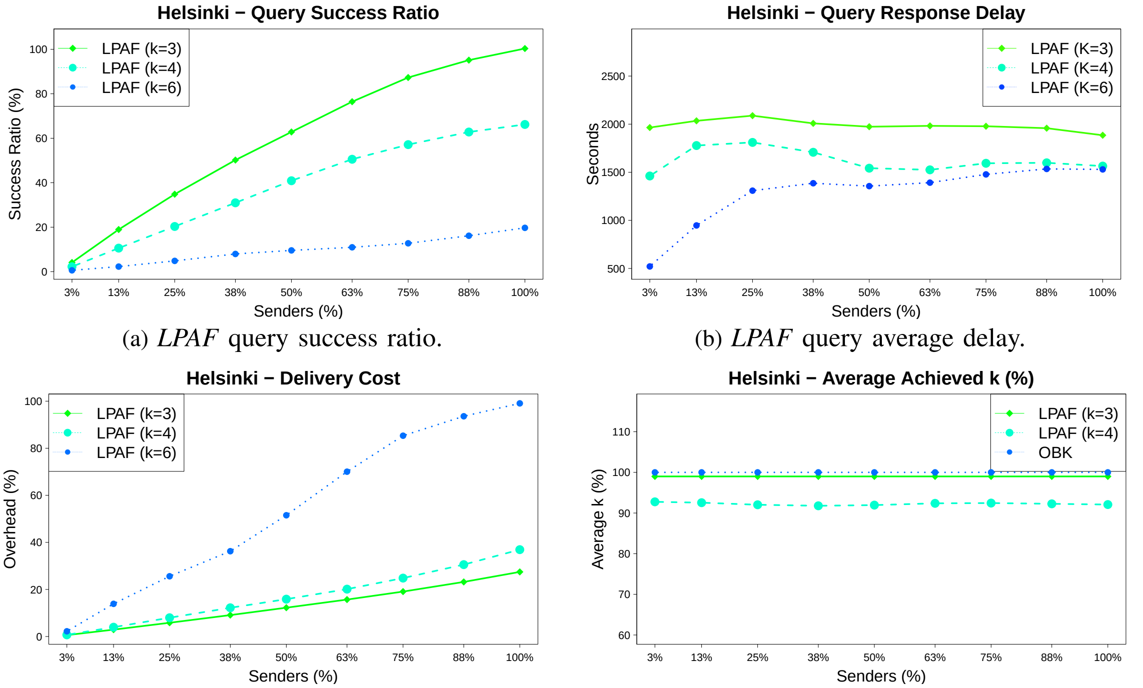
Figure 6b shows the normalized success ratio for *LPAF* compared to *No_Privacy* and the other privacy protocols. For both *SLPD* at $k > 3$ and *LPAF* at $k > 6$ results are 0%, so they are not shown on any of the graphs. The graph shows that *LPAF* is able to deliver queries in a privacy preserving way for $k = 3$ higher than the other location-privacy protocols (namely *SLPD* and *ALAR*). The figure shows that *ALAR* has a steady but low success ratio, and it is considerably lower than other protocols ($\approx 5\%$). *LPAF* has a rising success ratio as number of senders increases, while *SLPD* is constantly below by an average of 11%. *LPAF* is able to achieve a maximum of $\approx 50\%$, which is higher than both *SLPD* and *ALAR*, because nodes collaborate and share prediction to achieve the privacy level and move towards the LBS. *LPAF* is approaching the theoretical maximum success ratio maintained by *OBK* that uses an Oracle.

Figure 6c shows the average delay in seconds for each of the protocols. *No_Privacy* and *OBK* opportunistic forwarding have the lowest delay as they deliver queries without any obfuscation or using an Oracle. We include both protocols as benchmark for comparison, where *OBK* is not suitable for real-world *OppMNet*s scenario because it uses centralized server and assume future knowledge of network events.

In both figures 6b and 6c, we see that *ALAR* has low success ratio and high delay, this is because nodes wait for k distinctive neighbors for each of the S splits belonging to the same query. Which means that each query requires ($S \times k$) distinctive neighbors for all splits to be forwarded out of a source, or a total of 4 unique neighbors in our case, but in two sets one for each split ($S = 2$). Moreover, *ALAR* excludes neighbors from being possible next-hop for a particular query –rapidly– as soon as neighbors have received one split, and that causes additional fragmentation to an already fragmented network topology in *OppMNet*s.

Figures 6d shows the normalized average retransmission overhead, *i.e.*, the average of the total number of wireless transmissions for a particular protocol, divided by the maximum number of relayed queries in each scenario. We observe that *LPAF* retransmission overhead is linearly increasing with the number of senders and below *No_Privacy* forwarding. The level of retransmission overhead is less than the overhead for the *No_Privacy* protocol, and generally greater than *OBK*. For *ALAR*, the figure shows that it reaches a steady overhead level when senders are greater than 25% to reach around 22% which is due to the low success ratio ($\approx 7\%$). *No_Privacy* has higher overhead because it utilizes a more greedy approach to maximize delivery [42]. This is because location-privacy requirements act as a leash that controls and selectively forward queries, hence the less re-transmissions overhead. *LPAF* efficiency is because other protocols send many replicas of a query (or its splits) that never reach the destination and get dropped, but *LPAF* aims to use each replica with prediction to better direct them towards the destination LBS.

Figures 7a and 7b show *LPAF* success ratio and delay at different privacy levels respectively. We can generally see a decrease in the success ratio and an increase in delay as k increases, which validates earlier results from the analytical model evaluation figures (4a, 4b). *LPAF*(k=3) has lower delay –better– than *LPAF*(k=6) for number of senders higher than 20% while having better success ratio. This is because higher privacy causes *LPAF* to attempts to search and construct longer obfuscation path from source towards destination to reach k obfuscation hops, which leads to higher delays. This impact of the higher privacy-level requirement validates the impact seen in figure 3b, as increased k causes lower performance. We can also notice a slight increase in the delay initially, followed by a steady delay level. This is because as the number of senders increases, the interaction between different socially related nodes cause the protocol to build obfuscation path faster using the exchanged prediction information leading to better success ratio. This effect has no impact after the number of senders reaches a certain point as no additional information can be collected.



(a) *LPAF* query success ratio.

(b) *LPAF* query average delay.

(c) Retransmissions overhead relative to success ratio.

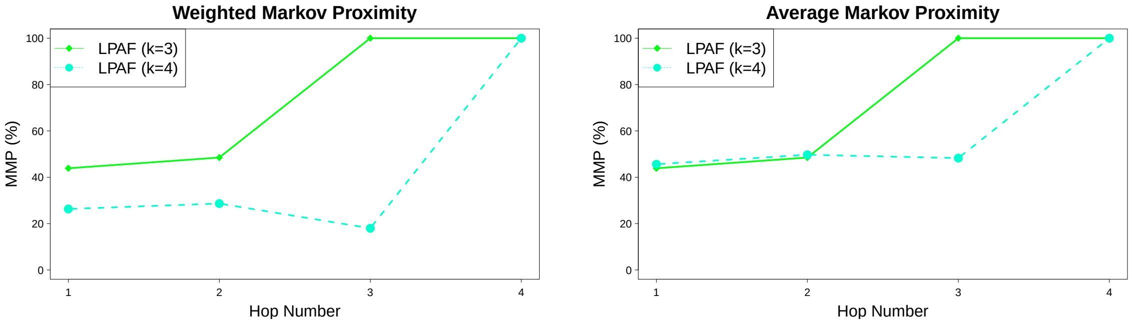
(d) Average achieved k (anonymity hops) accessing LBS service. (*LPAF*(k=3) and *OBK* are at 100%).

Fig. 7: Comparison of *LPAF* performance for different privacy-level requirements (k) and using various ratio of senders.

Figure 7c shows *LPAF* overhead relative to success ratio, *i.e.*, overhead per successfully delivered query. We can see that *LPAF* overhead is increasing linearly with the offered load, which supports the protocol design as consumed resources are manageable under increasing offered load. And as k increases, the overhead increases. This is due to the additional required re-transmissions of a query between k hops to perform the obfuscation.

Figure 7d shows the average achieved number of hops as a percentage of the required k set by the user. Each point is the average of all the percentage achieved k for all delivered queries. The figure shows that both *LPAF* and *SLPD* were able to achieve 100% (or 3 obfuscation hops within the social community). As k increases ($k = 4$), *SLPD* fails to deliver any queries hence achieved k is 0 and it is not shown on the graph, while *LPAF* compromises in order to continue delivering queries with k of $\approx$ 90% (or average $k \ = 3.9$) while maintaining high level of privacy.

Figure 8a shows the average MMP for all the nodes selected as message-relays at each hop for delivered messages. The figure shows that *LPAF* is discovering and selecting next-hop offering the best available proximity at different privacy-level requirement (*i.e.*, different k). MMP is displayed as 100% for hops greater than the requested privacy-level k, as query are freely forwarded beyond this hop. We can see from the graph that MMP of the hops is between 50–80%, and that nodes independently select the best next-hop depending on the encounters the node will come in contact with. Nodes search asynchronously for the best neighbor that offers the highest MMP, which can be greater or less than the previous message hops.



(a)

(b)

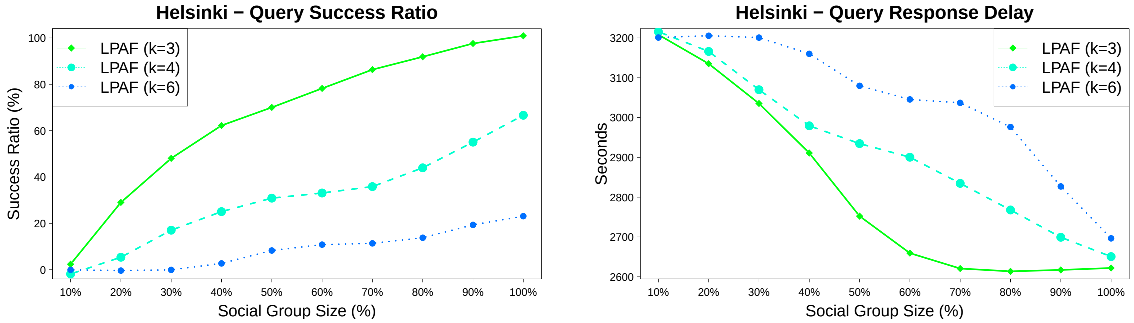
Fig. 8: (a) Average Markov proximity (MMP) at each hop for different k. (b) Weighted MMP relative to number of messages at each hop.

Figure 8b shows a weighted measure of MMP shown in figure 8a. Weighted MMP is denoted as $\overline{MMP}$, and is calculated as follows:

$$\overline{MMP}(h,k) = \frac{MMP \times QRY\,(h,k)}{\max\limits_{i \in k}\,(QRY\,(h,i))} \quad \text{where} \quad h \in [1,4], k \in (3,4)$$

Where the function $QRY()$ returns the number of queries successfully reaching hop h for privacy level k. Using $\overline{MMP}$ has the effect of normalizing the MMP measurement across the different values of k, while maintaining a per-hop distinctive results. The figure reflects the impact of the number of queries reaching each hop on the MMP. It shows that messages have better MMP as are forwarded over each successive hop towards the LBS. The two figures show that, even thou the average MMP is not varying considerably as k increases in both scenarios ($k = 3$ and $k = 4$), the weighted MMP drops as k increases. This is because less queries successfully pass through the whole obfuscation path, which leads to a drop in the weighted Markov proximity obtained through multiple hop prediction, so *LPAF*($k = 4$) has lower weighted proximity compared *LPAF*($k = 3$).

As expected from results we have seen in the analytical model evaluation (section IV-B); We notice in figure 9a that *LPAF* success ratio increases as we increase the social group size ($|S|$), which validates figures 2b and 3a in the evaluation of the analytical model; and decrease with increased value of k validating results in figure 3b. $|S|$ is increased by selecting more nodes from the pool of 40 nodes (around 30% of the total network nodes). Using the pool of nodes allows us to make a reference with other figures above. We can see that as the social group size increases, the success ratio increases because *LPAF* is able to find more socially related nodes. This is shown as a higher probability of forming k obfuscation path seen in the analytical model figure 2b. *LPAF* ability to utilize these extra social-related nodes enhances the overall success ratio as more encounters are being utilized for obfuscation and to achieve the obfuscation privacy-level requirement k. Figure 9b shows the associated overall average message delay, we notice that as k increases we experience greater delay as message gets buffered longer awaiting suitable encounters. On the other hand, we can see as social group size increases, the delay drops to minutes. This drop in delay with higher $|S|$ is consistent with the time delay presented on the x axises on figure 4a with different values of $|S|$ in the analytical model evaluation. As we can see $\varphi_s\,(t)$ rising faster in time (less delay) $|S|$ increases.



(a) Success ratio.

(b) Delay.

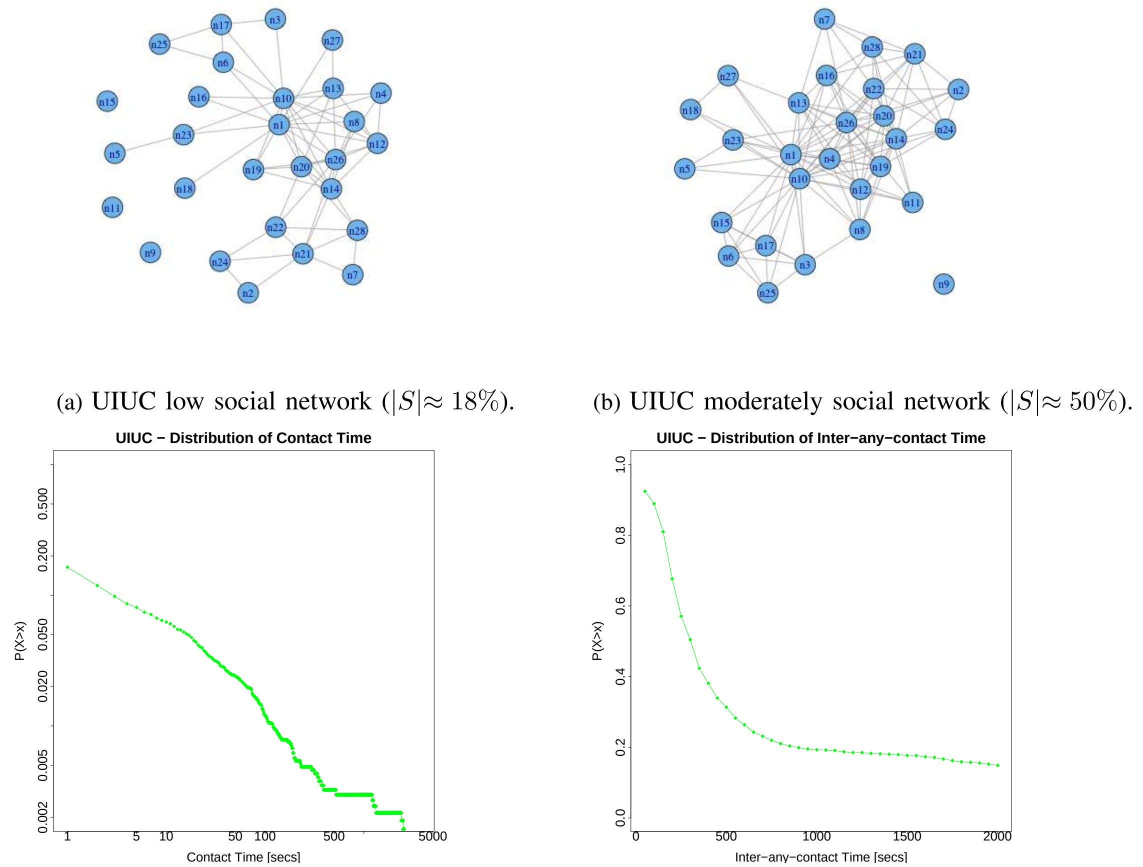
Fig. 9: Comparison of *LPAF* performance at different ratios of social group sizes $|S|$ and for various privacy levels (k).

B. Real-World Data Trace Simulation Using UIUC Dataset

1) Network Model

The UIUC [45] dataset is a real-world data trace collected at the University of Illinois. It contains the Media Access Control (MAC) addresses of Bluetooth devices and Wi-Fi access points collected by the University of Illinois Movement (UIM) framework using Google Android phones. Each phone regularly scans its surrounding, and an *encounter* occur when the scanning device detects another device in its communication area. The data collected is for 28 devices, and it spans a 3 weeks period during March 2010. Despite the lack of nodes' exact location information (*e.g.*, GPS) at the time of recording events, this dataset is particularly useful for evaluating location-privacy as it contains the MACs of Wi-Fi access-point encountered by the mobile devices. As these access-points are stationary, a device that records one of these MACs shows that it was present at the location of this access point. Subsequently, We have used these Wi-Fi MACs to represent the nodes' location at the time when the phone recorded an access-point MAC.

In order to build the social relationship between users, we preform analysis on the regularity of users existing at the same location similar to [9], [11]. We statistically study the frequency of devices encountering each other and encountering similar Wi-Fi access-points, which indicates that users carrying these devices exist at the same location regularly. We then generate a user-to-user (pair-wise) total encounter frequency. As participants in this dataset are faculty, staff, and students; more socially connected nodes meet inside as well as outside the university campus. We run two distinct social network scenarios (*low* and *moderate* social networks $|S|$), to compare performance obtained to that from the analytical model in figures 3a and 4a. In *low* $|S|$, two users are considered friends if their encounter frequency falls within the upper quartile (the top 25%) of the distribution of all node-pairs encounters frequency. The resulting social network has average $|S|$ equal to 5, and a standard deviation of 4; and it is shown in figure 10a (*i.e.*, $|S| \approx 18\%$). While in the *moderate* $|S|$ shown in figure 10b scenario, two users are socially connected if their encounter frequency is greater than the average encounter frequency of all users' pairs. The resulting social network has average $|S|$ equal to 10 friends and standard deviation of 5 (*i.e.*, $|S| \approx 50\%$).



(a) UIUC low social network ($|S|\approx 18\%$).

(b) UIUC moderately social network ($|S|\approx 50\%$).

(c) UIUC distribution of contact time.

(d) UIUC distribution of Inter-any-contact time.

Fig. 10: UIUC real-world data trace analysis.

We adopt the definition of contact time and inter-any-contact time from [46], where contact time refers to a pair-wise measure of the period when two users are within the communication area of each other, and can exchange data if they wish to; and inter-any-contact-time is a user metric that shows the average duration between any contact with other nodes (*i.e.*, isolation periods). Figure 10c shows the contact time distribution between users' pairs. The figure shows that the contact time follows the power-law distribution as observed in [46]. The figure shows that only 5% of the users have a contact time greater than 5 seconds, and $\approx 1\%$ of the nodes have contact time above 100 seconds. These are typical connectivity characteristics of opportunistic communication in DTN. Figure 10d showing the inter-any-contact time which indicates periods of complete user isolation. We observe that 50% of the users have isolation periods of about 300 seconds, and $\approx 20\%$ of all users have isolation periods above 1000 seconds. The inter-any-contact time is useful to analysis, as it shows that users have long isolation periods at which no data exchange can be performed which is not a privacy-related constraint but due to network topology formed through users' movement (or lack of).

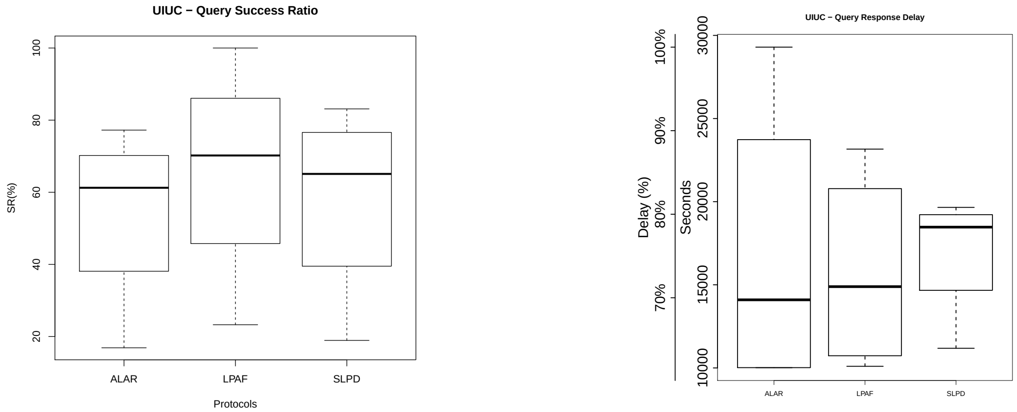
The dataset duration is split into three equal simulation duration of one week (7 days), which represents an adequate window for enough events occurring between nodes to allow data dissemination in the *OppMNet*. All

MACs for external devices were discarded, *i.e.*, only the 28 participants were sending/receiving/relaying message during the simulations. We repeat the simulation three times using the different simulation periods for all protocols, and results were averaged. A socially well-connected node is chosen to be the untrusted LBS (*n1* in figure 10a), and the remaining 27 nodes act as query sources (senders). LBS queries were sent at random intervals with a constant rate of one query every 6 hours.

2) Simulation Results in UIUC

This section shows the success ratio and delay comparing protocols that offer location-privacy in the *OppMNet*. We examine the performance of *LPAF* under two social scenarios (*i.e.*, low and moderate social network). This is in order to validate the impact of the social group size, with results obtained from the analytical model evaluation in section IV-B.

Figure 11a shows a set of experiments comparing location-privacy protocols performance using UIUC real-world data trace. The social network between nodes is the one presented in figure 10b with moderate social group size ($|S| \approx 50\%$). We can see that *LPAF* has better performance; with a higher –better– ($\approx 70\%$) median success ratio, and higher –better– min/max success ratios; compared to *SLPD* and *ALAR*. We can also see that the range between the minimum and maximum success ratios is big, represented by the lower/upper whiskers respectively; due to the difference in the network topology across the three weeks, and the change in participants movements and quality of collected contacts during the trace collection. Figure 11b presents the delay which is calculated as a percentage of the maximum achievable by the three protocols. Even thou *LPAF* median delay is close to *ALAR*, it shows lower –better– min/max delay (≈ 50 and 88%) respectively. And lower –better– quartiles (the boxplot for *LPAF* is lower than that of *ALAR*). This is because *LPAF* is employing MMP to guide the query to reach the LBS during obfuscation.

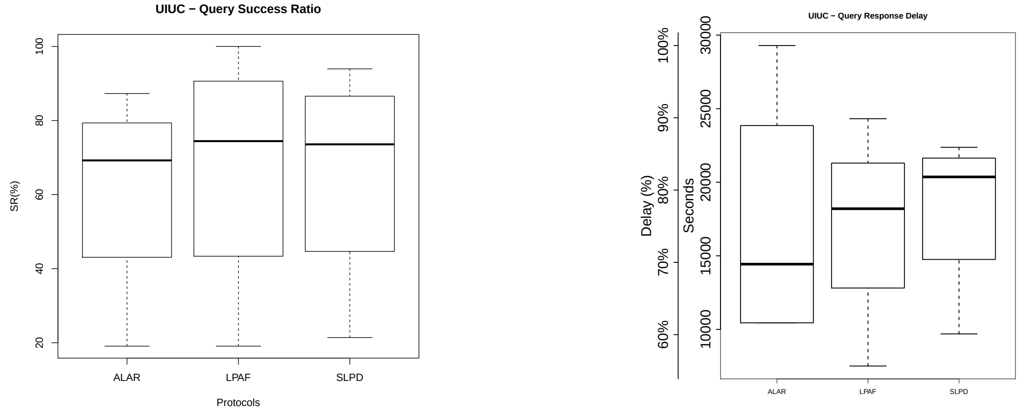


(a) Relative success ratio.

(b) Relative delay.

Fig. 11: *Moderate Social*: Using real-world data trace to compare *LPAF* performance against other location-privacy preserving protocols in *OppMNet* in figure 10b.

Figure 12 shows the same evaluation as in figure 11, but in a low social network scenario. The social network between nodes is the one presented in figure 10a with low social group size ($|S| \approx 18\%$). Figure 12a shows that *LPAF* manages to be slightly higher –better– mean success ratio than the other two protocols, and that *LPAF* performance drops (lower boxplot) to become very close to *SLPD* and *ALAR*. This is due to the restriction imposed by the limited number of friends to obfuscate queries, and the drop in probability of forming k obfuscation path seen in the analytical model figure 2b. *LPAF* drop in success ratio validates the analytical model performance seen in figure 3a, where we see lower percentage of nodes receiving packets as $|S|$ decreases.



(a) Relative success ratio.

(b) Relative delay.

Fig. 12: *Low Social*: Using real-world data trace to compare *LPAF* performance against other location-privacy preserving protocols in *OppMNet* in figure 10a.

Figure 12b shows the delays incurred by the different protocols. Firstly, we can see that *LPAF* and *SLPD* have both suffered additional delays higher than *ALAR*. This is due to the two protocols reliance on the social network size ($|S|$), and hence more time is spent searching for friends in a low social network. Secondly, by comparing this figure to figure 11b, we can see that *LPAF* delay has increased as $|S|$ decreased. This relation validates the earlier finding in the analytical model figure 3a which shows a slower rise –higher delay– as $|S|$ decreases.

C. Security Discussion

Our proposed protocol provides k-anonymity through obfuscation, where the *Location Service Provider* controlling the LBS is unable to distinguish the originating location or identify the source of a query out of –at least– k users (*anonymity set*) who have participated in obfuscating that query. We assume a trusted-community security model, and consider only an external attacker capable of eavesdropping on limited traffic in the network (as it will be discussed next). By trusted-community, we refer to the mutual trust between users who belong to the same social group (as defined in section I), which includes only known friends that the user entrust to perform location obfuscation for different LBS queries. We further assume that members of the social group do not launch attacks, nor collude with an attacker to determine the source's location.

Traffic analysis attacks similar to [47] can be harmful to k-anonymity protocols including our proposed protocol *LPAF*. Using traffic analysis, the attacker collects information about who is communicating with whom and when. We only consider an attacker with limited presence, *i.e.*, able to eavesdrop over all the traffic along the obfuscation path, but only a subset of forwarded queries over the obfuscation path. In which case, the attacker can only monitor forwarding activities less than k over the obfuscation path, and will have incomplete knowledge about the remaining obfuscation hops on that path. For a successful attack, the attacker needs to be able to isolate the path followed by a certain query from the rest of the traffic in the network and identify the source node over that path which is not possible if the attacker do not monitor that path. Our protocol makes such an attack non-trivial because: 1) socially-related users normally are co-located and meet at regular times where they exchange other types of messages which makes it difficult to perform traffic analysis to isolate obfuscated queries, 2) the attacker needs to eavesdrop over all network traffic as the obfuscation path the query takes is unknown. With repeated traffic analysis over time, an attacker can learn about social relation between various nodes which provide includes a group of size $\leq k$, hence does not compromise the k-anonymity.

In the case where the attacker is omni-present (*i.e.*, the attacker can monitor all traffic in the network), a privacy attacks can be easily launched to identify a certain query source. Firstly, the attacker can then link these events to the obfuscated LBS query and trace-back the LBS query to deduce the complete obfuscation path back to the original source with high accuracy. Secondly, this situation in *OppMNet* is equivalent to the attacker being able to physically monitor the wireless medium over the whole network or physically track the victim node, and hence he does not need to intercept the LBS query to know the victim location, as the existence of the wireless signal from the victim node is sufficient to reveal its existence at this area. In this case, obfuscation is not useful to provide higher privacy.

The proposed privacy protocol relies on the user social-network, and subsequently the movement pattern of users in this group can greatly affect the user privacy. For example, considering a simplified scenario of a social group of two friends (A and B). If both friends visit the same set of locations over the day at the same times, then neither of them can offer obfuscation service to the other (*i.e.*, the two users become identical in terms of their movement profile; and if A exists at one location, then it trivial to infer that B exists there too). Considering real-life human-movements, similarity between the users' visited locations is being observed [11], but it is shared among the whole social group, which subsequently provide a larger anonymity set and plausible deniability [48].

VI. Related Work

There has been little attention given to location privacy threats in Opportunistic Mobile Networks, with insufficient user centric analysis and solutions. Most existing research rely on an online TTP, such as mobile phone operators [49] or online centralized server [50], which is generally not applicable to *OppMNet*.

One way to provide higher location privacy is to limit access to location information through access control mechanisms [22], [50]. Wernke et al. [50] present PShare, a secure location sharing protocol. PShare distributes location data across multiple location servers to increase robustness against attack and ensure imprecise user position in case a server is compromised. Puttaswamy and Zhao [22] present a solution where LS servers are treated as un-trusted data-stores and LBSs are offered by friends in the network. All location information sent to the LS is encrypted and only authorized users or applications are granted access. Both schemes assume an online LS, and access control is applied using public-private keys or multi-secret sharing.

Another approach for providing higher location privacy is to obfuscate the location information by generating false location update-events that causes diluted users movement traces at LBSs side. Meyerowitz and Choudhury [7] propose CacheCloak, an anonymization system for communication in Vehicular ad-hoc networks (VANETs). Rongxing et al. propose SPRING [51], where road-side units (RSUs) are positioned at highly-social intersects are used to store temporarily the forwarded packets, and hence prevent an adversary from tracking its source. The two approaches rely on a pre-trusted and controlled nodes (*e.g.*, anonymizing server, or RSUs) to offer obfuscation.

TABLE V: Comparison of Location-Privacy offered by various protocols.

Privacy / Anonymity	GSM	Mobile-IP	SPRING [51]	Mano et al. [52]	SUNC [53]	SpotME [54]	ALAR [40]	*LPAF*
Privacy-aware	No	No	Yes	Yes	No	Yes	Yes	Yes
Delay tolerant	No	No	Yes	No	No	Yes	Yes	Yes
No pre-trust server	No	No	No	No	Yes	Yes	Yes	Yes
k-Anonymity	No	No	No	Yes	No	No	No	Yes
Decentralized	No	No	Yes	No	Yes	Yes	No	Yes
Sparse topology	No	Yes	No	No	No	Yes	No	Yes
Encryption	No	Yes	Yes	No	Yes	No	Yes	Partially
Energy efficient	No	No	No	No	No	No	No	Partially
Location-aware	Yes	Yes	No	Yes	No	Yes	No	Yes
Profile-aware	No	No	No	Yes	No	No	No	Yes
Adaptive	No	No	No	No	No	No	No	Yes
Context aware	No	No	No	Yes	No	No	No	Yes
Location prediction	Yes	Yes	No	No	No	No	No	Yes

Our proposed protocol is based on k-anonymity [55] which is a widely used approach for anonymity in many fields such as communication networks and databases. In traditional networks, k-anonymity is used in privacy solutions such as Tor [56] for Internet, P2P VoIP [57], where the anonymized user is indistinguishable from at least $k-1$ other users called the anonymity set. One popular way of offering location k-anonymity is through cloaking. Cloaking is the technique by which the geo-spatial dimension is divided into areas called cloaks, and the user is k-anonymized if the formed cloaking region is occupied by at least $k-1$ other users, so the user's location is hidden among k anonymity set. Spatial and temporal cloaking introduced by Gruteser and Grunwald [58] are using a centralized location broker service to form the cloaking region. Cloaking techniques are not suitable to *OppMNet* because of the small communication area and nodes mobility.

On the other hand, Mixing offers location privacy by ensuring that user's identity is kept private, and it is based on untraceable email communication proposed by Chaum [59]. A number of Mixing techniques have been proposed [14]. During the mixing process, a trusted entity performs shuffling of users' identities inside a specially constructed mix-zone, so that old and new user identities are not linkable. This process requires that nodes physically move into and stay inside the mix-zone throughout the mixing process in order to prevent eavesdropping by an attacker. Mixing is not suitable for *OppMNet* as it requires that a specified number of nodes co-exist inside the mix zone for the whole duration this operation.

Mano and Ishikawa [52] propose an anonymization for users of location-based services in mobile environments using a trusted third-party entity, which hides the location and profile attributes of the query from un-trusted service. Yanfei, et al. propose SUNC [53], a privacy-preserving scheme to maintain source unobservability in multi-hop wireless networks. SUNC utilizes network coding and specifically designed dummy messages to ensure packet unlinkability and source anonymity. SUNC assumes a multi-hop wireless network, and does not consider the long delays, mobility and disruption impact typically expected in DTNs.

Zakhary and Radenkovic propose "Social-based Location Privacy in DTNs" (*SLPD*) [39]. *SLPD* is a location-privacy protocol that obfuscate queries using the social relationship between nodes and aims to achieve a predefined privacy level. *SLPD* is not adaptive to the topology or individual nodes' connectivity patterns. The authors compare their work to "Oracle-Based k-Anonymity Protocol" (*OBK*), a benchmark k-anonymity location-privacy protocol, which requires an Oracle with complete future knowledge of the network. *OBK* is used to evaluate the effectiveness of the offered k-anonymity privacy by relying on centralized and trusted matchmaker server, and is not practically applicable to *OppMNet*s scenario.

Lu, X., et al. propose "Anti-localization anonymous routing for DTNs" (*ALAR*) [40] to allow a source to send messages without revealing its physical location while communicating opportunistically over DTNs. *ALAR* focuses on ensuring nodes' location-privacy while transmitting sensitive messages by preventing an adversary from obtaining the complete message. The source splits each message into multiple *S* segments (each is called split), and each split is transmitted through an alternative disjoint path (k). *ALAR* uses Epidemic routing to send as many copies, but only when the node (source or intermediary) has k neighbors in its communication area that have not been observed for any of the previously sent splits for the same message.

Quercia et al. propose SpotME [54], a technique for preserving users' location-privacy through reporting of fake locations. The approach uses randomized response algorithm, and is addressing specific targeted applications. Table V shows a summary of comparison between various proposed protocols in different environments, according to a list of identified key criterion for providing location privacy in *OppMNet*s. Song et al. [60] study, comparing four different general techniques for location prediction, shows that using more sophisticated predictors is not necessary and does not increase accuracy significantly compared to Markov predictors. Authors show that using Markov predictors beyond the second order (*i.e.*, making predictions based on history of size $n >= 3$ previous locations) decrease the prediction accuracy.

Geo-routing [61], [62] is a research area where location information is being used to assist in routing data in various ad-hoc networks. The work presented in this paper differs from geo-routing in many aspects. Firstly, geo-routing searches for the shortest path to a given location, but *LPAF* does not necessarily search-for the shortest path. *LPAF* aims to find a path that preserves the user's location privacy through subsequent obfuscations. Secondly, nodes participating in geo-routing need to exchange information about their movement patterns in order to enable centralized or distributed predictions of nodes' locations at different times. On the other hand, *LPAF* prevents nodes from sharing detailed location information, as this negatively impacts the user's location-privacy, and only allows aggregated Markov predictions inside the social group to be shared. Finally, the proposed protocol allows queries to be forwarded over a path that goes in a roundabout way to avoid revealing sources' location information to the LBS, hence the nodes on the obfuscation path do not deliver the query to the LBS even if communication opportunity arose as this leaks the source location-information.

VII. Conclusion and future work

We proposed new privacy-preserving protocol *LPAF* that offers adaptive, fully-distributed, socially-driven location-privacy forwarding. *LPAF* employs a distributed light-weight Markov-based location-prediction model to guide the obfuscation phase of message propagation to the location based services. *LPAF* utilizes social links between nodes to form an obfuscation path between the source and the LBS to offer best-effort k-anonymity location privacy. We evaluate the proposed protocol using analytical model simulation; and extensive experiments in both city-center map-based heterogeneous mobility scenario, and using real-world data trace. Results show that by adapting the privacy within tolerance to nodes' own connectivity and incorporating location prediction, *LPAF* performs better than existing location-privacy protocols in terms of success ratio with small overhead. *LPAF* is able to maintain around 10% success ratio at high user privacy-level requirement when other protocol fails by adaptively lowering the achievable location privacy to the network conditions.

As future work, we plan to extend the protocol to handle situation where the exact location of the nodes is not known, and examine additional movement models and rural scenarios.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive and detailed review of this paper.

REFERENCES

[1] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, 2009.

[2] I. Parris and T. Henderson, "The impact of location privacy on opportunistic networks," in *WoWMoM*, 2011.

[3] S. Siersdorfer and S. Sizov, "Social recommender systems for web 2.0 folksonomies," in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM, 2009.

[4] M. Dell'Amico and L. Capra, "Sofia: Social filtering for robust recommendations," *Trust Management II*, 2008.

[5] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava, "Biketastic: sensing and mapping for better biking," in *SIGCHI*. ACM, 2010.

[6] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *TIST, ACM*, vol. 2, no. 1, p. 2, 2011.

[7] J. Meyerowitz and R. Roy Choudhury, "Hiding stars with fireworks: location privacy through camouflage," in *MobiCom*. ACM, 2009, pp. 345–356.

[8] A. Mtibaa, M. May, C. Diot, and M. Ammar, "Peoplerank: social opportunistic forwarding," in *INFOCOM*. IEEE, 2010.

[9] I. Parris and T. Henderson, "Privacy-enhanced social-network routing," *Computer Communications*, vol. 35, no. 1, pp. 62–74, 2012.

[10] H. Wei-jen, D. Dutta, and A. Helmy, "Profile-cast: Behavior-aware mobile networking," in *WCNC*. IEEE, 2008.

[11] G. S. Thakur, A. Helmy, and W.-J. Hsu, "Similarity analysis and modeling in mobile societies: the missing link," in *CHANTS*. ACM, 2010.

[12] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick, "A survey on privacy in mobile participatory sensing applications," *Journal of Systems and Software*, vol. 84, no. 11, pp. 1928–1946, 2011.

[13] A. Alganas, X. Lin, and A. Grami, "Evse: An efficient vehicle social evaluation scheme with location privacy preservation for vehicular communications," in *ICC*. IEEE, 2011.

[14] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *Pervasive Computing, IEEE*, vol. 2, no. 1, pp. 46–55, 2003.

[15] M. Duckham and L. Kulik, "Location privacy and location-aware computing," *Dynamic & mobile GIS: investigating change in space and time*, 2006.

[16] S. K. Belle, M. Waldvogel, and O. Haase, "Pathforge: faithful anonymization of movement data," in *MobiHeld*. ACM, 2009.

[17] B. Schneier and P. Sutherland, *Applied cryptography: protocols, algorithms, and source code in C*. John Wiley & Sons, Inc., 1995.

[18] A. Menezes, P. V. Orschot, and S. Vanstone, *Handbook of applied cryptography*. CRC Press, 1996.

[19] S. B. Sassi and N. Le Sommer, "Towards an opportunistic and location-aware service provision in disconnected mobile ad hoc networks," in *MobileWireless Middleware, Operating Systems, and Applications*. Springer, 2009, pp. 393–406. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-01802-2_29

[20] N. Le Sommer and S. Ben Sassi, "Location-based service discovery and delivery in opportunistic networks," in *Networks (ICN), 2010 Ninth International Conference on*. IEEE, 2010, pp. 179–184. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5473971

[21] K. Abrougui, A. Boukerche, and R. W. N. Pazzi, "Design and evaluation of context-aware and location-based service discovery protocols for vehicular networks," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 3, pp. 717–735, 2011. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5951776

[22] K. P. N. Puttaswamy and B. Y. Zhao, "Preserving privacy in location-based mobile social applications," in *HotMobile*. ACM, 2010, pp. 1–6.

[23] L. McNamara, C. Mascolo, and L. Capra, "Media sharing based on colocation prediction in urban transport," in *Proceedings of the 14th ACM international conference on Mobile computing and networking*. ACM, 2008.

[24] A. Mashhadi, B. Mokhtar, and L. Capra, "Habit: Leveraging human mobility and social network for efficient content dissemination in delay tolerant networks," in *WoWMoM*. IEEE, 2009.

[25] W. Xu, C.-Y. Chow, M. L. Yiu, Q. Li, and C. K. Poon, "Mobifeed: a location-aware news feed system for mobile users," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 2012, pp. 538–541. [Online]. Available: http://dl.acm.org/citation.cfm?id=2424409

[26] H. Haddadi, P. Hui, T. Henderson, and I. Brown, "Targeted advertising on the handset: Privacy and security challenges pervasive advertising." Springer London, 2011, pp. 119–137. [Online]. Available: http://dx.doi.org/10.1007/978-0-85729-352-7_6

[27] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *Communications Magazine, IEEE*, vol. 44, no. 11, pp. 134–141, 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4014485

[28] W. Chen, R. K. Guha, T. J. Kwon, J. Lee, and Y.-Y. Hsu, "A survey and challenges in routing and data dissemination in vehicular ad hoc networks," *Wireless Communications and Mobile Computing*, vol. 11, no. 7, pp. 787–795, 2011. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/wcm.862/full

[29] A. Sathiaseelan, J. Crowcroft, M. Goulden, C. Greiffenhagen, R. Mortier, G. Fairhurst, and D. McAuley, "Public access wifi service (paws)," *Digital Economy All Hands Meeting*, 2012.

[30] P. Hui and J. Crowcroft, "How small labels create big improvements," in *PERCOMW*. IEEE, 2007.

[31] M. Musolesi, P. Hui, C. Mascolo, and J. Crowcroft, "Writing on the clean slate: Implementing a socially-aware protocol in haggle," in *WoWMoM*. IEEE, 2008.

[32] T. Small and Z. J. Haas, "The shared wireless infostation model: a new ad hoc networking paradigm(or where there is a whale, there is a way)," in *MobiHoc*. ACM, 2003.

[33] R. Groenevelt., *Stochastic models in mobile ad hoc networks, Ph.D. dissertation*. University of Nice Sophia Antipolis, 2005.

[34] 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 *INFOCOM*. IEEE, 2006.

[35] A. Kottas and B. Sansó, "Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis," *Journal of Statistical Planning and Inference*, vol. 137, no. 10, pp. 3151–3163, 2007.

[36] J. G. Andrews, R. K. Ganti, M. Haenggi, N. Jindal, and S. Weber, "A primer on spatial modeling and analysis in wireless networks," *Communications Magazine, IEEE*, vol. 48, no. 11, pp. 156–163, 2010.

[37] P. Hui, K. Xu, V. O. K. Li, J. Crowcroft, V. Latora, and P. Lio, "Selfishness, altruism and message spreading in mobile social networks," in *INFOCOM*. IEEE, 2009.

[38] G. Zyba, G. M. Voelker, S. Ioannidis, and C. Diot, "Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd," in *INFOCOM*, 2011.

[39] S. Zakhary and M. Radenkovic, "Utilizing social links for location privacy in opportunistic delay-tolerant networks," in *ICC*. IEEE, 2012.

[40] X. Lu, P. Hui, D. Towsley, J. Pu, and Z. Xiong, "Anti-localization anonymous routing for delay tolerant network," *Computer Networks*, vol. 54, no. 11, pp. 1899–1910, 2010.

[41] A. Keränen, J. Ott, and T. Kärkkäinen, "The one simulator for dtn protocol evaluation," in *SIMUTools*. ICST, 2009.

[42] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *SIGCOMM workshop on DTN*. ACM, 2005.

[43] M. X. Yang, S. X. Han, C. Y. Yang, L. Zhang, and D. F. Ye, "Survey on node mobility model for opportunistic network," *Applied Mechanics and Materials*, vol. 52, pp. 1253–1257, 2011.

[44] E. Daly and M. Haahr, "Social network analysis for information flow in disconnected delay-tolerant manets," *IEEE Transactions on Mobile Computing*, 2009.

[45] L. Vu, K. Nahrstedt, S. Retika, and I. Gupta, "Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus," in *MSWIM*. ACM, 2010.

[46] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *SIGCOMM workshop on DTN*. ACM, 2005.

[47] S. J. Murdoch and G. Danezis, "Low-cost traffic analysis of tor," in *Security and Privacy, Symposium on*. IEEE, 2005.

[48] S. K. Belle and M. Waldvogel, *Consistent deniable lying: Privacy in mobile social networks*. Bibliothek der Universität Konstanz, 2008.

[49] J. G. Khuong Vu, Rong Zheng, "Efficient algorithms for k-anonymous location privacy in participatory sensing," in *INFOCOM*, 2012.

[50] M. Wernke, F. Dürr, and K. Rothermel, "Pshare: Position sharing for location privacy based on multi-secret sharing," in *PerCom*. IEEE, 2012.

[51] L. Rongxing, L. Xiaodong, and S. Xuemin, "Spring: A social-based privacy-preserving packet forwarding protocol for vehicular delay tolerant networks," in *INFOCOM*. IEEE, 2010.

[52] M. Mano and Y. Ishikawa, "Anonymizing user location and profile information for privacy-aware mobile services," in *LBSN*. ACM, 2010.

[53] F. Yanfei, C. Jiming, L. Xiaodong, and S. Xuemin, "Preventing traffic explosion and achieving source unobservability in multi-hop wireless networks using network coding," in *GLOBECOM*, 2010.

[54] D. Quercia, I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft, "Spotme if you can: Randomized responses for location obfuscation on mobile phones," in *ICDCS*. IEEE, 2011.

[55] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[56] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," DTIC Document, Tech. Rep., 2004.

[57] M. Srivatsa, A. Iyengar, and L. Liu, "Privacy in voip networks: A k-anonymity approach," in *INFOCOM*. IEEE, 2009.

[58] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *MobiSys*. ACM, 2003.

[59] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[60] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 12, pp. 1633–1649, 2006.

[61] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad hoc networks*, vol. 3, no. 3, pp. 325–349, 2005.

[62] Z. C. Taysi and A. G. Yavuz, "Routing protocols for geonet: A survey," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 2, pp. 939–954, 2012.