Privacy-Aware Message Exchanges for Geographically Routed Human Movement Networks

Adam J. Aviv[1], Micah Sherr[2], Matt Blaze[1], and Jonathan M. Smith[1]

[1] University of Pennsylvania
[2] Georgetown University

Abstract. This paper introduces a privacy-aware geographic routing protocol for *Human Movement Networks* (HumaNets). HumaNets are fully decentralized opportunistic and delay-tolerate networks composed of smartphone devices. Such networks allow participants to exchange messages *phone-to-phone* and have applications where traditional infrastructure is unavailable (*e.g.*, during a disaster) and in totalitarian states where cellular network monitoring and censorship are employed. Our protocol leverages self-determined *location profiles* of smartphone operators' movements as a predictor of future locations, enabling efficient geographic routing over metropolitan-wide areas. Since these profiles contain sensitive information about participants' *prior movements*, our routing protocol is designed to minimize the exposure of sensitive information during a message exchange. We demonstrate via simulation over both synthetic and real-world trace data that our protocol is highly scalable, leaks little information, and balances privacy and efficiency: messages are 30% more likely to be delivered than similar random walk protocols, and the median latency is only 23-28% greater than epidemic protocols while requiring an order of magnitude fewer messages.

1 Introduction

The ubiquity of smartphones enable new communication models beyond those provided by cellular carriers. While standard cellular communication uses a centralized infrastructure that is maintained by the service provider, smartphones have communication interfaces such as ad-hoc WiFi and Bluetooth that allow direct communication between devices. Since smartphone owners often carry their devices, leave them constantly on, and encounter other individuals (and their smartphones) in their daily routines, *smartphones enable fully decentralized store-and-forward networks that completely avoid the cellular infrastructure.*

Human Movement Networks (HumaNets) [2] fit this model and are designed to allow participants to exchange messages phone-to-phone without using any centralized infrastructure. HumaNets' "out-of-band" message passing is applicable when cellular networks are unavailable or if the networks are untrusted (*i.e.*, operated by a totalitarian state that censors, shuts down, or otherwise leverages its communication systems to restrict its citizenry).

Rather than rely on network addresses, HumaNets route messages using *geocast* – an addressing scheme that directs messages towards a particular geographic region. Such a messaging system could be used, for example, to notify a group of people in a targeted area of an upcoming event, or to warn them of some impending crisis. To cope with mobility, HumaNet routing protocols route messages based on message carriers' predicted *future* locations. This is accomplished by leveraging self-determined *location profiles* that approximate the smartphone owners' routine movements. The patterns of human mobility – for example, the daily commute to and from work – serve as predictors of future locations. HumaNets take advantage of this observation by greedily forwarding messages to smartphones whose owners' location profiles indicate that they are good candidates for delivery.

Privacy issues must be central when designing a HumaNet routing protocol since location profiles contain sensitive information about participants' *prior movements*. The disclosure of such information is particularly dangerous when HumaNets are used for covert communication in totalitarian regimes. Existing decentralized routing approaches that do not consider privacy [8,10], rely on trusted third parties [6], or assume *a priori* trust relationships [4] are also unsuitable for HumaNets.

This paper proposes a novel routing protocol for HumaNets that protects participants' location profiles from an adversary who wishes to learn previous movements and/or determine "important" locations of network users (*e.g.*, home, work, or the location of underground activist meetings). Our technique, which we call *Probabilistic Profile-Based Routing* (PPBR), balances performance and privacy by efficiently routing messages in a manner that minimizes the exposure of users' location profiles. We demonstrate through trace-driven simulations using both real-world and synthetic human movement data that our PPBR protocol is highly scalable, efficiently routes messages, and preserves the privacy of profile information. In summary, the contributions of this paper are: (1) The introduction and design of a fully decentralized, privacy-preserving, geographic-based HumaNet message routing protocol for smartphones; (2) An analysis of the privacy and security properties offered by our routing protocol; and, (3) A trace-driven simulation study (using both real-world and synthetic data) that evaluates our method's scalability and efficiency.

2 Network Assumptions and Goals

To achieve reasonable performance, HumaNets leverage humans' tendency to follow *routines*: The locations that people frequented in the past are predictors of their future locations [2]. However, a device's location history may be extremely sensitive, and moreover, combining multiple nodes' location histories may allow an adversary to discover social networks and enumerate participants' movements. Hence, the high-level goal of our PPBR protocol and the central challenge of this paper is to enable *efficient geographic-based messaging that limits the exposure of important location information at message exchanges.*

Importantly, however, our HumaNet routing protocol does not conceal the identities of the network's participants. An adversary who intercepts a PPBR message can reasonably conclude that the sender is participating in a HumaNet. Participating in a HumaNet inherently carries risk if used as an anti-censorship technology: This is unfortunately true of any system that may be deemed "subversive". However, when other means of communication are impossible (either due to global monitoring or blocked connectivity), HumaNets provide a *means* to exchange information in a manner that is efficient, scalable, difficult to surveil, and privacy-aware.

Requirements. HumaNets routing protocols are designed for location-aware mobile devices. We assume that network participants can learn their locations (*e.g.*, via GPS[3]) without relying on the cellular service provider's network, and that devices contain sufficient storage to record their movement histories. We note that current generation smartphones meet HumaNets' modest storage and processing requirements.

We additionally assume that participants have knowledge of the routing area. Since HumaNets enable geocast routing, a message that is targeted at specific receivers requires the sender to have some knowledge about the receivers' likely future locations (*e.g.*, their home or work); this requirement is similar to that imposed by traditional networking where users need knowledge of a service's hostname or IP address. We also assume that participants know some coarse-grain information about general movement statistics over the routing area. In particular, nodes should be capable of estimating the "popularity" of city areas – *e.g.*, that the upper west side of Manhattan is more densely traveled than Far Rockaway, Queens. This information can be obtained from census data, other public source of information, or personal experience. Such information can be shipped with the HumaNets software and is assumed to be known to an adversary.

Threat Model. We envision both passive and active adversaries. A passive adversary may have any number of confederates and is able to observe message exchanges at a fixed number of locations throughout the HumaNet routing area. An active adversary may additionally participate in HumaNets by generating fake messages, accepting messages, and/or dropping or misrouting messages.

We do not provide protection against a *mobile targeting adversary*. An adversary that can physically follow a node can trivially learn about its whereabouts and discover its routine movements. Such a "stalker" adversary is also very costly to deploy. In this paper, we focus on less targeted attackers and assume an adversary who monitors, intercepts, or participates in local exchanges that occur in its presence. The adversary is aware of the participants and their locations at the time of an exchange, and thus we do not claim that our system provides traditional location-privacy [9] for ad hoc networks, although such extensions may be relevant here.

[3] GPS is a unidirectional protocol and requires only the reception of signals from U.S.-operated satellites.

The adversary's goals are as follows:

- Disruption: Inject failures into the network such that messages can no longer be reliably delivered.
- De-anonymization: Determine the originating sender of intercepted messages.
- Profiling: Infer movement patterns of a targeted individual or learn his/her "important" locations (*e.g.*, home, work, underground meeting place).

Performance and Security Goals. The goal of our routing protocol is to provide the following properties in the presence of active and passive adversaries:

- Reliability: Messages should reach their intended destinations with high probability.
- Efficiency: Messages should reach their intended destinations with reasonable latency and overhead.
- Scalability: HumaNets should be able to scale to a large number of participants with many concurrent messages.
- Point-to-Point: Messages should be exchanged only point-to-point and avoid any centralized routing structures.
- Privacy-Preservation: The protocol should not leak the sender's identity, nor should it reveal information about participants' previous locations. We do not distinguish between locations that should or should not remain private (e.g., secret meeting place vs. place of work). The treatment of *all* prior locations as private simplifies our protocol design, and more importantly, improves usability by preventing configuration errors that may lead to accidental exposure of private locations.

At first blush, it may seem that naïve flooding and random walk strategies are sufficient to achieve the above goals. Although these strategies achieve the Point-to-Point and Privacy-Preservation properties, they are lacking with respect to Scalability, Efficiency, and/or Reliability. In particular, flooding achieves optimal latency and delivery rates because all paths are explored, but scales poorly since all transfers that do not occur along the optimal path constitute a wasted effort (and, consequently, wasteful power consumption). Moreover, since several senders may use HumaNets to disseminate their messages, flooding requires that nodes store (and worse, communicate) a large fraction of all messages. At the other extreme, random walk protocols in which messages are transferred (as opposed to copied) upon node contacts scales well but incurs poor Reliability and Efficiency.

It may also seem that traditional cryptographic solutions would be applicable here. However, the decentralized and highly dynamic nature of HumaNets make their deployment difficult. In particular, many cryptographic solutions require centralized services or trusted third parties. Such approaches are problematic in our setting since a strong (*e.g.*, nation-state) adversary could either compromise or prevent access to centralized services. Routing techniques that rely on complex

key distribution schemes or expensive cryptographic operations (for example, SMC [23]) are incompatible with HumaNets' distributed architecture and use of power-constrained devices. A significant advantage of PPBR is that it provides PRIVACY-PRESERVATION using simple probabilistic techniques, and avoids the key management and computation issues present in protocols that provide more traditional cryptographic protections [6,4,21].

Finally, we note that a non-goal of our system is authentication of message senders and message content. PPBR is a content-agnostic service that routes packets, whether they be sent by dissidents trying to organize a rally or a totalitarian state that wishes to provide misinformation. However, as with standard networking protocols, PPBR may be combined with other techniques – for example, the use of pseudoidentities and digital signatures – to provide stronger authenticity guarantees.

3 Privacy-Preserving Routing

At a high level, the *Probabilistic Profile-Based Routing* (PPBR) protocol requires participants (nodes) to *estimate* whether they are good candidates for delivering a message. Upon receiving a message from a *carrier* — *i.e.*, a node that announces a message — the receiving node makes a local determination as to whether it is well positioned to deliver the message to the addressed destination. The node either *accepts* or *discards* the message, and in either case, *does not notify the current carrier as to its choice*. If the message is accepted, the receiving node becomes a carrier and begins to announce the message. However, unlike flooding techniques in which messages are continuously duplicated, leading to an exponential number of message copies, each message carrier in PPBR announces the message to only k contacts, of which only one out of the k receiving nodes should accept it. The main task is thus for a receiver to locally determine whether it is best suited to deliver the message out of the $k-1$ other nodes that received the message.

3.1 HumaNet Preliminaries

Addressing. HumaNets provide a basic addressing primitive, *geocast*, in which messages are addressed to a geographic location (*e.g.*, a city square). Messages are routed to nodes who are likely to travel towards the destination address and are then locally flooded within the confines of the specified destination. We do not consider temporal features in addressing or routing – *i.e.*, addressing a message to a location for a specific time – but the protocol described herein can be easily expanded to meet temporal specifications[4]. Additionally, HumaNets do

[4] One method for delivering messages at a targeted time of day is for nodes to maintain multiple location profiles, each representing movement information collected at different times of day. The message exchange algorithm is as described later; however, each node now uses the location profile most relevant to the addressed time and location.

not provide message confidentiality; however, message payloads can be protected using standard encryption techniques.

HumaNets interpret the routing area as a grid, the dimensions of which are assumed to be known *a priori* to all nodes (for example, based on latitude and longitude). Messages are addressed to a particular grid square. In the remainder of the paper, when describing a message address or destination, we refer to the index of the corresponding grid square.

Finally, HumaNets are fully decentralized, delay tolerate networks, and as such, deliver messages according to a "best-effort" policy. Importantly, PPBR does not utilize message delivery acknowledgments; the omission of ACKs and NACKs *increases* privacy since it prevents an observer from trivially discovering whether or not a message was accepted by the receiver.

Message Exchanges. Messages are exchanged between smartphone devices when they come into wireless contact with one another. We consider a contact to occur when two nodes are within wireless transmission range, *e.g.*, the range of Bluetooth or a point-to-point 802.11 transmission in ad hoc mode. At set time intervals, nodes awaken and begin the routing protocol. If a contact is made, messages can be exchanged. Otherwise, if there are no other participants nearby, the node returns to normal activity.

HumaNets require coarse time synchronization (*i.e.*, within a few seconds) to ensure message exchanges occur at the appropriate times. Such synchronicity could be achieved using NTP servers, but this would require nodes to send messages over centralized networks. Fortunately, smartphone devices are already highly synchronized as a requirement of participating in the centralized cellular network [1,16] (a network which HumaNets do not use to send messages). If cellular services are disabled or are untrusted to provide correct time information, nodes could alternatively obtain the timing information from GPS satellite timestamps.

3.2 Routing Overview and Constructions

PPBR consists of two phases: a *passing phase* and a *holding phase* (see Figure 1). In the passing phase, a carrier of a message attempts to pass the message to the first k nodes that it encounters. A node that receives a message will locally estimate whether it has the highest similarity to the message address (a grid square) out of the $k-1$ other nodes who also received (or will receive) the message. If the node perceives itself to be the best candidate for delivery, it accepts the message, becomes a carrier, and prepares to transition to the passing phase. Otherwise, the message is dropped. A node transitions from the passing phase to the holding phase once it has announced the message to k other neighbors.

The challenge of PPBR is enabling each node to accurately predict whether it is the best of k candidates to accept a message *without conferring with other nodes*. The intuition behind our approach is that a node can compute a *similarity score* to a message's destination using its *location profile* – a compact representation of its movement history. To populate its location profile, a node periodically

records its GPS location and determines the fraction of time spent within each grid square. Using its location profile along with background knowledge of the movement patterns of an "average" node, the node can estimate how well it is positioned to deliver the message relative to the $k-1$ other participants who will receive the message.

An important characteristic of PPBR's passing phase is that message reception is not acknowledged. An eavesdropper therefore cannot determine whether a message was accepted or declined by a nearby node. This makes it difficult for an adversary to conduct PROFILING attacks against a receiver, since it has no information to form a judgment as to whether the receiver's profile is well-suited for delivering the message. (We explore the effectiveness of PROFILING attacks against a carrier who announces a message in Section 5.) To further aggravate PROFILING attacks, if a node accepts a message and becomes a carrier, it does not announce the message until it has moved a distance d away from its current location, preventing the eavesdropper from observing the transition.

After a carrier has performed k message announcements, it transitions to the holding phase. In the holding phase, the carrier maintains the message for some time period, during which the node, hopefully, enters the message's addressed grid square and starts the local flood (restricted to the destination grid square). If the node does not reach the addressed grid square within a *local timeout*, the carrier drops the message. A message also has an associated *global timeout* after which all carriers drop the message.

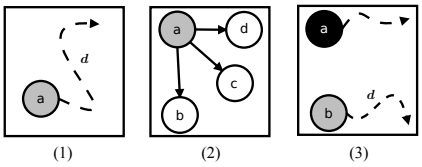

Fig. 1. Overview of PPBR routing. (1) The initial message carrier (node a) enters the passing phase (grey shading). (2) The carrier encounters three nodes. (3) Node b considers itself the best of k candidates and accepts the message, becoming a carrier and initiating its passing phase. After advertising k messages, node a enters the holding phase (black shading).

Location Profiles. Nodes compute *location profiles* based on their movement histories.[5] Although long term collection could be useful in constructing a profile, HumaNets rely on shorter historical windows to minimize the effects from non-repeated movements, *e.g.*, vacations.

Each node periodically polls its location (*e.g.*, via GPS) to update its location profile. The profile is a matrix indexed by geographic grid square such that the value at position $\langle x, y\rangle$ is the normalized number of location readings in which the node was located at position $\langle x, y\rangle$ in the grid. That is, the value at position $\langle x, y\rangle$ in the location profile corresponds to the frequency that the node visited location $\langle x, y\rangle$ in the physical world over some time window. Following our heuristic, we assume that the matrix value at $\langle x, y\rangle$ (which is defined based on past behavior) approximates the node's future likelihood of visiting location $\langle x, y\rangle$ in the physical topology.

[5] News reports suggest that popular smartphones may already collect such information [3].

More formally, consider a current window of location entries $W = (\langle x_i, y_i \rangle, \langle x_j, y_j \rangle \ldots)$ that are already mapped to grid square references. The profile p, indexed by grid squares, contains the values:

$$p[\langle x, y \rangle] = \begin{cases} \frac{|W_{\langle x,y \rangle}|}{|W|} & \text{if } \langle x, y \rangle \in W \\ 0 & \text{otherwise} \end{cases} , \quad (1)$$

where $W_{\langle x,y \rangle}$ is the sub-list containing location entries occurring within the grid square $\langle x, y \rangle$, $p[\cdot]$ is the index function returning the associated value, and $|\cdot|$ indicates the length of the list.

General Node Profile. An advantage of PPBR is that it does not require nodes to share their location profiles. However, the technique assumes some globally shared information which we call the *general node profile*. The general node profile is a model of the "average" node's movement, and has the same structure and features as the standard location profile. Rather than representing the frequented locations of a single node, the general profile expresses the patterns of the general population. We assume that the general node profile is included with HumaNet software.

As we demonstrate in Section 4, the general node profile does not have to be a perfect model and can be based on a rough estimate of population densities. In practice, we posit that a sufficient general node profile could be constructed using public data such as population densities from census data, transportation studies, or common knowledge.

Marginal Similarity. A node determines if it is the best of $k-1$ other message recipients by comparing its similarity with the message's destination to the "average" node's similarity calculated using the general node profile. If the node's similarity is a factor greater, the message is accepted.

More precisely, a node must first be able to calculate the similarity of a location profile to a message address (grid square). We consider not only the value in the profile at the addressed grid-point, but also the values at nearby grid-points, discounted by their square distance. Formally, we define the similarity of a node n to a message m addressed to a_m to be:

$$\mathsf{sim}(p, a_m) = p[a_m] + \sum_{\substack{a_p \in p \\ a_p \neq a_m}} \frac{p[a_p]}{\mathsf{dist}(a_p, a_m)^2} , \quad (2)$$

where p is a location profile and $\mathsf{dist}(a_p, a_m)$ denotes the Euclidean distance between grid-points a_p and a_m. This computation captures the desired property that a node that more frequently visits the message's targeted destination (and nearby areas) will have higher similarity than a node that visits the destination region less often[6].

[6] We have additionally experimented with other decay functions, and found that they produce similar (but slightly degraded) performance.

A similarity score computed with the general node profile, rather than an individual node's profile, represents an estimate of the "average" node's similarity to the message address. We define the relationship between a node n's similarity and that of the general node's similarity as the *marginal similarity* σ. It is calculated as $\sigma = \frac{\mathsf{sim}(p_n, a_m)}{\mathsf{sim}(p_g, a_m)}$, where p_n is the profile of node n and p_g is the general node profile. The marginal similarity speaks to how well a node is suited to become a carrier of a message addressed to a_m as compared to a node on average: higher values indicate the node would make a good message carrier, while lower values indicate a poor carrier. The next challenge is selecting a threshold value for σ at which point only one of the k nodes that received the message will accept it and become a carrier.

Threshold Selection. We define τ as the *threshold marginal similarity score* at which a node accepts a message and becomes a carrier. Intuitively, τ should be the marginal similarity such that $1/k$ marginal similarity calculations are greater than τ. The threshold is calculated locally (and privately) by each node. First, a node computes σ for every grid square in p_g:

$$\bar{\sigma} = \left\langle \frac{\mathsf{sim}(p_n, a)}{\mathsf{sim}(p_g, a)} \;\middle|\; \forall\, a \in p_g \right\rangle \tag{3}$$

The computations are arranged in a sorted list $\bar{\sigma}$, where $\bar{\sigma}_i < \bar{\sigma}_j$ if $i < j$. $\bar{\sigma}$ represents marginal similarity calculations for all likely message addresses, and we wish the node to accept a message for $1/k$ of those addresses. To do this, a node chooses τ such that $1/k$ values in $\bar{\sigma}$ are greater than τ; more precisely, $\tau = \bar{\sigma}_i$ and $i = \lfloor |\bar{\sigma}| * (k-1)/k \rfloor$, where $|\cdot|$ denotes the length function. τ must be updated whenever the node's location profile changes. To conserve battery, such a computation could occur nightly while the device is charging.

It should be noted that the threshold computation assumes a uniform distribution of message addresses. Although this assumption does not likely hold in practice, our experimental results indicate that our approach is sufficiently accurate to cause approximately $1/k$ messages to be accepted by potential carriers. In particular, using our tested datasets (see Section 4) in which messages are addressed non-uniformly, between 8.5%-9.5% of messages are accepted.

4 Performance Evaluation

To evaluate the performance of PPBR, we constructed a discrete event-driven HumaNets simulator. Our simulator takes as input a trace of human (cellphone) movement and overlays the PPBR routing algorithm. In all simulations, we choose k to be 10 and conduct 300 independent runs. Message senders are selected randomly across participants, and message addresses (grid squares) are randomly chosen by selecting a (different) node and addressing the message to its most frequented grid square as defined by its location profile. Our simulation was concerned with measuring the effectiveness of PPBR over metropolitan

Table 1. Characteristics of the movement data sets.

	Nodes	Length	Area	Contact Rate	Waypoints
SLAW [13]	1000	7 days	100 km^2	12.62 per hour	150
Cabspotting [18]	536	20 days	326 km^2	1.17 per hour	n/a

areas, and as such, we did not simulate local flooding. We considered a message successfully delivered if it reaches the destination address. The grid overlay consists of 200 m × 200 m grid squares, roughly the size of a city block, and we chose d — the requisite travel distance of a node before transitioning to the passing phase — to be the size of a grid square (200 m).

Datasets. Due to privacy constraints, the number of realistic datasets that are suited for evaluation is unfortunately small. We require that the data contain not only a large number of nodes, but also that the movement of the nodes should express regular routines over an extended collection time (*i.e.*, many days). To demonstrate the feasibility of PPBR, we utilize a suitable real-world data trace as well as a synthetic trace of human movement (summarized in Table 1):

- **Cabspotting:** The **Cabspotting Dataset** [18] contains GPS coordinates and timestamps of 536 taxicabs in the San Francisco area. The dataset spans 20 days: from May 20, 2008 until June 7, 2008. It should be noted that although the movements of taxis are not representative of the general population (taxis are arguably more mobile than the average person), simulations using this dataset can be interpreted as representing a network composed of the taxi drivers' smartphones.
- **SLAW:** We require a synthetic model that (i) accurately represents human *flight patterns*, (ii) contact rates, (iii) *waypoints* (popular places), and (iv) routines. The closest model to meeting our needs is **Self-similar Least Action Walk** (SLAW) [13]. Based in part on Levy walks [20], SLAW introduces a protocol called *Least Action Trip Planning* (LATP) that produces human-like trips between fractal waypoints, that are themselves determined by finding hotspots in actual GPS traces.

Node Contacts. For two nodes to make contact, they must be in the same location at the same time. However, the periodicity of location entries in the Cabspotting dataset is not consistent across nodes (or for the same node). We consider two nodes to have made contact if they are within 10 meters in a 10 second window. In SLAW, a location entry is generated every 60 seconds consistently across all nodes; we consider a contact to occur if two nodes are within 10 meters at the same minute mark.

Timeouts. We use a 12 hour local timeout with both traces. For the shorter, more dense SLAW movement trace, a three day global timeout is used. The longer, more sparse Cabspotting trace uses a seven day global timeout. Finally,

Table 2. Median and Average Latencies (first and third quartiles in braces) and Delivery Rate.

	Cabspotting [18]		**SLAW** [13]	
	Med/Avg Latency (hrs)	Rate	Med/Avg Latency (hrs)	Rate
PPBR	3.6/6.8 [1.2,4.6]	62.6%	4.2/4.8 [2.6,6.2]	61.8%
Walk-10%	4.4/6.0 [1.6,8.1]	43.4%	5.1/5.5 [2.9,5.2]	48.0%
Flood-10%	2.8/4.1 [1.6,4.4]	99.4%	3.4/3.3 [2.2,4.2]	100.0%

simulations begin after an initial delay so that node profiles can be well seeded; delays of three and seven days are used for SLAW and Cabspotting, respectively.

Location Profiles. Each node constructs its location profile using a three day window of location histories. Location profiles are updated daily, and the current day's profile represents the location history of the three previous days.

To generate the general node profile, we select a 10% sample of nodes from each dataset and use three days worth of movement data. The 10% sample is excluded from all simulation experiments. A visualization of the resulting general node profile are shown in Figures 4 and 5 in the Appendix.

4.1 Simulation Results

To measure the efficiency of PPBR, we compare our strategy against two probabilistic protocols that do not use location information: *probabilistic random walk* and *probabilistic flooding*. The probabilistic random walk routing scheme also has passing and holding phases; however, unlike PPBR, the random walk does not use location profiles. Instead, a node accepts a carrier's advertised message with a fixed probability of $1/k$ (*i.e.*, 10%). We also compare PPBR to a 10% probabilistic flood in which nodes duplicate the message to a contacted node with probability 0.1. The flood provides insight into a worst case for network load – *i.e.*, exponential growth in the number of duplicate messages. The global and local timeouts for both random protocols are identical to those used by PPBR.

Threshold Estimation. As described in Section 3.2, each node computes its threshold marginal similarity score (τ) based on the general node profile and its knowledge of the routing area. To determine if our local, per-node threshold calculations were generating good thresholds, we looked at the variance of thresholds calculated at each node for one day in the simulation. The average value for τ was 1.557 and 1.353 for SLAW and Cabspotting, respectively. We found that there is very low variance among the nodes' thresholds: 0.011 for SLAW and 0.085 for Cabspotting. Further, we observed that thresholds were effectively limiting message acceptance to $1/k$; with $k = 10$ the probability of message retention was 9.5% and 8.5% for SLAW and Cabspotting, respectively.

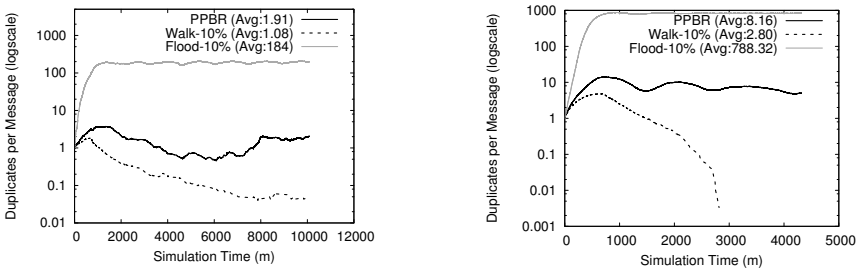

Fig. 2. The number of message copies ("duplicates") of each message for (**left**) Cabspotting and (**right**) SLAW, and inset, the average.

Performance Metrics. We evaluate our routing performance using the following metrics: *delivery rate* is the percentage of messages that reach the destination address (a grid square); *latency* is the amount of time it takes for a message to be delivered; and *network load* is the number of messages in the network at a given time. Ideally, the routing protocol should deliver messages with a high delivery rate, low latency, and low network load.

Delivery Rate and Latency. Table 2 lists the delivery rates and latencies for PPBR, random walk, and probabilistic flooding[7]. Unsurprisingly, flooding offers both the best latency and delivery rates. (As we show later, it also incurs a very high network load, making it impractical for networks of battery-constrained smartphone devices.) PPBR routing outperforms random walk for both median latency and delivery rate. Although the average latency for PPBR using the Cabspotting dataset is 0.8 hours slower, the median latency is nearly an hour faster and within 28% of probabilistic flooding. The skew in the average latency is caused in part by the higher delivery rate, and that some messages were delivered after random walk was no longer delivering messages.

Network Load. The load on the network is measured as the average number of message duplicates in the system across all simulations runs. PPBR does not guarantee that only a single copy of a given message is present in the system. Carriers announce a message to k other nodes; ideally, only one node *should* accept it. If the message is accepted, the carrier retains the message until either it is delivered or a local timeout occurs. Hence, each message could potentially have multiple (or zero) duplicates.

Figure 2 plots the number of messages that persist in the system over time, normalized to the number of senders in the system (which, in our simulation experiments is always 300). The average number of message copies, computed over the entire simulation, is shown in the Figure's key. Note that the number of message duplicates may be less than one if either some messages are not

[7] The delivery rates reported in Table 2 result from single attempted transmissions.

accepted by any of the k encountered nodes, or if all message copies are delivered to their destinations. As expected, flooding incurs significant network load, resulting in approximately two orders of magnitude more message copies than PPBR. Although the number of duplicates is slightly larger for PPBR than our naïve random walk protocol, the load is easily manageable.

5 Security Properties

Profiling. All message exchanges in PPBR occur in the open, and an adversary can observe any exchange in its presence. However, PPBR offers strong privacy protections against PROFILING attacks for both the node announcing a message as well as the node who receives, and possibly accepts, the message announcement.

Message Exchange Carrier Protections: An adversary can determine that a carrier node who advertises a message has a high marginal similarity to the message's address; otherwise, the node would not be advertising the message. The adversary knows that the marginal similarity for the carrier is lower bounded by the threshold τ, and that nodes choose τ such that they should expect to accept messages addressed to $1/k$ of the grid squares. Hence, *the acceptance of a message does not necessarily indicate that the message's address is particularly important to the node that accepted it.* Depending upon the value of k, a node may be expected to accept messages targeted at hundreds of grid squares across the routing area.

Larger values of k decrease privacy since nodes accept messages for fewer locations, and, thus, an adversary could deduce that these locations are more likely relevant to the victim node. Conversely, smaller values of k increase privacy since nodes accept messages to more locations, further obscuring which are important. Smaller values of k also incur higher power consumption and network load as more nodes will likely accept (and transfer) the message. In our simulation studies, we found that $k = 10$ achieves reasonable privacy while restraining the number of message transfers.

To study this tradeoff further, we compared the set of addresses (grid squares) that would result in a node a accepting a message to the node's most frequented locations as defined in the location profile. Although nodes accepted messages addressed to $1/k$, many of those locations correspond to grid square that are uninteresting to an adversary who wishes to learn the most frequented grid squares. This relationship is depicted in Figure 3 (left). The curves represent the averages across all nodes in the Cabspotting and SLAW datasets. The x-axis denotes the number of points an adversary is interested in (*i.e.*, the x grid squares most frequented by the node). The y-axis plots the fraction of the locations that are accepted by the node which are of interest to the adversary. Generally, the more specific the adversary's interest, the more difficult it is for him to distinguish the pertinent message addresses that are announced by a node, and consequently, the more difficult it is to discover the node's most frequented locations.

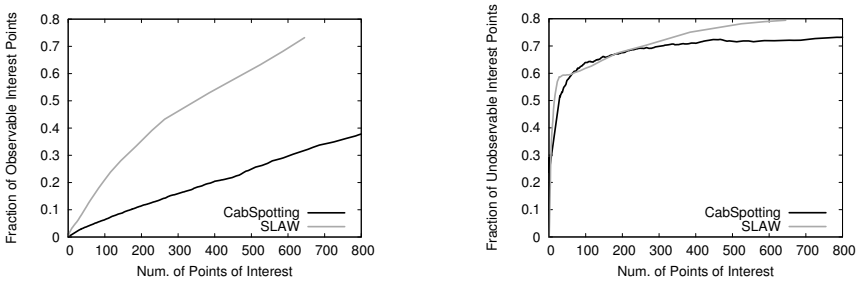

Fig. 3. Fraction of Safe Interest Points (**left**) and Fraction of Interesting Observations (**right**).

The adversary's ability to discern profile information is further diminished due to our algorithm's willingness to discard announcements that are targeted at highly frequented areas. Recall that the marginal similarity is the ratio of the node's similarity score to the general node profile's similarity score. Hence, if a message is addressed to a grid square that is often frequented by the node *but also highly frequented according to the general node profile*, then the ratio will not exceed the τ threshold, and the node will *never* accept a message addressed there. Consequently, such interesting locations are unobservable and *safe* from adversarial analysis. Figure 3 (right) visualizes this relationship. Again, the x-axis considers the number of grid squares an adversary would find interesting for a victim node. The y-axis represents the fraction of those interesting grid squares a node would *never* accept a message for, averaged across all nodes.

Message Exchange Receiver Protections: During the passing phase, receivers do not acknowledge acceptance (or rejection) of a message, and hence an adversary cannot directly determine its similarity to the message's destination address. An adversary who is able to follow the node for a distance of at least d can determine whether the message has been accepted by observing whether or not it is re-advertised by the node. Such a stalking attack inherently leaks the victim's location information regardless of the particular routing protocol being used. Regardless, if the node *is* followed, or if a separate colluding eavesdropper discovers that the node later advertised the message, then the adversary can only conclude that the node accepted the message. In such cases, the effectiveness of a PROFILING attack against the receiver is identical to the effectiveness against a carrier advertising a message (see above).

De-Anonymization. The standard addressing primitive of HumaNets is geocast, and thus all participants at the addressed location at the time of delivery should receive the message. Receiver anonymity is trivially exposed in HumaNets because an adversary located in the address location learns the identities of the message recipients simply by observing them. However, PPBR provides in-transit anonymity for message originators (or senders). An intercepted message,

past the initial hop, cannot be traced to the original sender without completely retracing the message's path. If an adversary is witness to the initial hop of a message, the originating sender may be exposed. We note, however, that this is similar to the level of protection provided by many Internet-based anonymity systems (*e.g.*, Crowds [19]) in which an adversary on the first hop may infer with some probability that it has identified the sender (since the sender may have originated upstream). It is also worth noting that message replay attacks in which an attacker re-injects a message in hopes of discovering its path are also infeasible. It is highly unlikely a message will take the same path due to variability in human movement.

Disruption. PPBR also provides protection against DISRUPTION attacks in which an adversary attempts to intercept messages in the network. If the attacker is able to infiltrate the network and receive a large portion of the k handoffs for each message, then the probability that the message will be transferred to an honest node is reduced. However, such an attack may also be prohibitively expensive for an adversary since message exchanges occur whenever two participants have a chance encounter. Additionally, such an attack may be mitigated by adjusting the number of passing attempts (*i.e.*, k) to compensate.

6 Related Work

The ability to leverage geographic information to efficiently route packets has been well explored in the literature [8,14]. In many instances, these techniques require participants to announce their locations. For example, Last Encounter Routing (LER) [8] and ProPHET [14] expose location information; LER assumes that the network is sufficiently connected to allow stable and longstanding paths. Although these techniques may efficiently route messages, they are not well-suited for settings in which the disclosure of location histories and/or social relationships may be cause for government-imposed punishment.

There are a number of approaches that attempt to preserve *location privacy*. Here, the goal is often to prevent an adversary from either identifying the source of an intercepted communication or tracking a node over time. Several protocols (cf. [7,24]) achieve location privacy by relying on ephemeral pseudoidentities. Such approaches provide *unlinkability* by impeding an adversary's ability to associate different broadcasts with the same node. Although these techniques can be used in conjunction with our PPBR protocol, we assume an adversary who is physically present at various (but not all) locations in the network and can identify individuals and associate broadcasts with their senders (*e.g.*, through physical identification). Similarly, anti-localization techniques [15] that are designed to prevent an adversary from determining a sender's location [11] are ineffective since our adversary can physically observe nodes.

A number of location privacy protocols (cf. [6,22]) are loosely based off of AODV [17], a popular routing protocol for decentralized mobile networks (*e.g.*, MANETs). However, such techniques assume a highly connected and mostly

static network in which messages can be quickly forwarded between nodes. These protocols assume that nodes are mostly stationary, communication can occur with low latency, and anonymous paths can be reused for multiple exchanges, and as such, are therefore not well-suited for networks of mobile smartphones.

There are a number of existing delay tolerant network (DTN) protocols that are similar to HumaNets, but either have limited functionality or lack HumaNets' privacy protections. For instance, Zebranet [12] uses local information to efficiently exchange information between sensor nodes in order to track wildlife. However, the network can route messages only towards fixed basestations. GeoDTN+Nav [5] is a vehicular ad-hoc network routing scheme that, like HumaNets, relies on location profiles to deliver messages in a DTN. However, GeoDTN+Nav requires that at least some nodes follow fixed paths (*e.g.*, bus routes) or provide their destinations before travel (*e.g.*, via a car navigation system). And in previous work, we applied *polygon-intersection algorithm* [2] to HumaNets; however, this protocol does not consider privacy.

The work that perhaps most closely resembles ours is Shifka *et al.*'s protocol [21]. Here, the authors use the heuristic that nodes that share more *contexts* are more likely to encounter one another. Like our approach, participants construct profiles that describe frequented locations, but Shifka *et al.* relies on search able encryption schemes (namely, PEKS) to limit the adversary's ability to enumerate the contents of a profile. Additionally, their approach assumes a trusted third party that assigns attribute values (*e.g.*, a frequented location) to nodes.

7 Conclusion

This paper presents *probabilistic profile based routing* (PPBR), a novel privacy preserving geographic messaging protocol for HumaNets. Designed for networks of smartphone devices, our PPBR routing protocol avoids the use of the cellular network — or any other centralized infrastructure — and is well-suited for environments in which traditional communication is subject to monitoring and/or censorship. PPBR leverages self-determined location profiles to assist routing while minimizing the disclosure of location information to outside observers as well as adversaries who infiltrate the network. In particular, we demonstrate using simulations over real-world and synthetic movement data that PPBR is resistant to disruption, de-anonymization, and location-leakage attacks, while achieving reasonable delivery rates and latency.

Acknowledgments. This work is partially supported by NFS grants CNS-1064986, CNS-1149832 and CNS-0905434, and ONR grant N00014-09-1-0770. This material is based upon work supported by the Defense Advanced Research Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract No. N66001-11-C-4020. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Project Agency and Space and Naval Warfare Systems Center Pacific.

References

1. 3rd Generation Partnership Project. Universal Mobile Telecommunications System (UMTS); Synchronization in (UTRAN) Stage 2. Technical Specification Group Services and System Aspects 3GPP TS25.402 v8.1.0, 3rd Generation Partnership Project, July 2009.
2. A. J. Aviv, M. Sherr, M. Blaze, and J. M. Smith. Evading Cellular Data Monitoring with Human Movement Networks. In *USENIX Workshop on Hot Topics in Security (HotSec)*, August 2010.
3. N. Bilton. Tracking File Found in iPhones. The New York Times, April 20 2011.
4. A. Boukerche, K. El-Khatib, L. Xu, and L. Korba. An Efficient Secure Distributed Anonymous Routing Protocol for Mobile and Wireless Ad Hoc Networks. *Computer Communications*, 28(10):1193–1203, 2005.
5. P. Cheng, J. Weng, L. Tung, K. Lee, M. Gerla, and J. Haerri. GeoDTN+Nav: A Hybrid Geographic and Dtn Routing with Navigation Assistance in Urban Vehicular Networks. In *Symposium on Vehicular Computing Systems*, 2008.
6. K. El Defrawy and G. Tsudik. PRISM: Privacy-friendly Routing in Suspicious MANETs (and VANETs). In *International Conference on Network Protocols (ICNP)*, 2008.
7. J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes. On Non-cooperative Location Privacy: A Game-theoretic Analysis. In *ACM Conference on Computer and Communications Security (CCS)*, 2009.
8. M. Grossglauser and M. Vetterli. Locating mobile nodes with ease: learning efficient routes from encounter histories alone. *IEEE/ACM Trans. Netw.*, 14(3):457–469, 2006.
9. M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2003.
10. P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket Switched Networks and Human Mobility in Conference Environments. In *ACM SIGCOMM Workshop on Delay-tolerant networking (WDTN)*, 2005.
11. N. Husted and S. Myers. Mobile Location Tracking in Metro Areas: Malnets and Others. In *ACM Conference on Computer and Communications Security (CCS)*, 2010.
12. P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet. In *Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X)*, Oct. 2002.
13. K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. SLAW: A New Mobility Model for Human Walks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2009.
14. A. Lindgren, A. Doria, and O. Scheln. Probabilistic routing in intermittently connected networks. In *Service Assurance with Partial and Intermittent Resources*, volume 3126, pages 239–254. 2004.
15. X. Lu, P. Hui, D. Towsley, J. Pu, and Z. Xiong. Anti-localization Anonymous Routing for Delay Tolerant Network. *Computer Networks*, 54(11):1899 – 1910, 2010.
16. P. Mann. Timing Synchronization for 3G Wireless. *EE Times Asia*, December 2004.

17. C. Perkins, E. Belding-Royer, and S. Das. Ad hoc On-Demand Distance Vector (AODV) Routing. RFC 3561, IETF, 2003.
18. M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser. A Parsimonious Model of Mobile Partitioned Networks with Clustering. In *Conference on COMmunication Systems and NETworkS (COMSNETS)*, 2009.
19. M. K. Reiter and A. D. Rubin. Crowds: Anonymity for Web Transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.
20. I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the Levy-Walk Nature of Human Mobility. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2008.
21. A. Shikfa, M. Onen, and R. Molva. Privacy and Confidentiality in Context-Based and Epidemic Forwarding. *Computer Communications*, 33(13):1493–1504, 2010.
22. D. Sy, R. Chen, and L. Bao. ODAR: On-Demand Anonymous Routing in Ad Hoc Networks. In *IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS)*, 2006.
23. A. C. Yao. Protocols for Secure Computations. In *Symposium on Foundations of Computer Science (FOCS)*, 1982.
24. Y. Zhang, W. Liu, W. Lou, and Y. Fang. MASK: Anonymous On-Demand Routing in Mobile Ad Hoc Networks. *IEEE Transactions on Wireless Communications*, 5(9):2376–2385, September 2006.

Appendix: General Node Profile Heatmaps

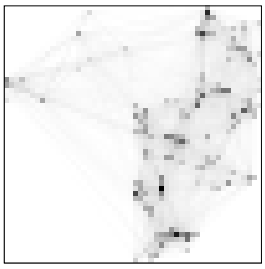

Fig. 4. Heatmap of the General Node Profiles for the SLAW dataset. Darker shades indicate regions with higher node densities.

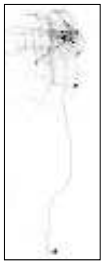

Fig. 5. Heatmap of the General Node Profiles for the Cabspotting dataset. Darker shades indicate regions with higher node densities.