Preserving Privacy of Agents in Participatory-Sensing Schemes for Traffic Estimation

Farhad Farokhi and Iman Shames

Abstract— **A measure of privacy infringement for agents (or participants) travelling across a transportation network in participatory-sensing schemes for traffic estimation is introduced. The measure is defined to be the conditional probability that an external observer assigns to the private nodes in the transportation network, e.g., location of home or office, given all the position measurements that it broadcasts over time. An algorithm for finding an optimal trade-off between the measure of privacy infringement and the expected estimation error, captured by the number of the nodes over which the participant stops broadcasting its position, is proposed. The algorithm searches over a family of policies in which an agent stops transmitting its position measurements if its distance (in terms of the number of hops) to the privacy sensitive node is smaller than a prescribed threshold. Employing such symmetric policies are advantageous in terms of the resources required for implementation and the ease of computation. The results are expanded to more general policies. Further, the effect of the heterogeneity of the population density on the optimal policy is explored. Finally, the relationship between the betweenness measure of centrality and the optimal privacy-preserving policy of the agents is numerically explored.**

I. Introduction

A sharp rise in the number of the networked platforms, such as smart phones and wearable gadgets, has enabled new technologies, such as participatory-sensing schemes, to be commercially viable. In these schemes, agents (or participants) and their networked devices act as sensing units to estimate a variable of interest. Waze[1] and Mobile Millennium[2] can be mentioned as examples of commercial and academic products that use participatory-sensing schemes for measuring the traffic flow in real time. These systems often recruit agents that are willing to provide data (by directly providing reports or letting the sensors on their devices to be remotely used). However, gathering data usually reveals some private information about the agents (e.g., the location of their house or office among many other variables), which might make them opt out of the system or switch off their connected devices (more often than needed). One way to alleviate the privacy related anxieties of the agents is to "systematically corrupt" the measurements collected by them. The objective of such corruption is to obfuscate the private information of each agent. A direct link between the intensity of corruption and the quality of the estimate, on the one hand, and the availability of the private information, on the other hand, can be established. This observation is at the core of the literature on differential privacy (see [1]–[5] among other studies), where the responses to statistical queries on random databases are typically corrupted by Laplace noise to protect the privacy of the individuals in the database.

In transportation systems, differential privacy might not be well-suited for preserving the privacy of the agents. To demonstrate this, as an example, consider a scenario in which the networked devices of the agents in a participatory-sensing scheme provide Global Positioning System (GPS) measurements of their position in equidistant intervals of time. From the perspective of a subscribing agents, the privacy infringement is only an issue if the agent is close to a privacy sensitive node[3], such as its home or office. Therefore, when travelling through most of the network, transmitting accurate measurements of its position is not infringing the privacy of the agent (and thus the addition of the noise is conservative). In addition, if an agent is staying in a certain node for long durations of time (e.g., over night at home or during work hours at office), a simple averaging can significantly reduce the effect of any additive noise on the position measurements and, thus, revealing the position of private nodes. Hence, adding noise to the sensor measurements (at least indiscriminately with respect to the position) to protect the privacy of the agent is neither necessary nor beneficial. Therefore, a different approach is required for protecting the privacy of the agents in which they completely stop broadcasting their position based on their proximity to the private nodes. Such a method can form the basis of a software package that gives rise to the democratization of privacy-preserving tools and measures in transportation networks. The software simply asks for the privacy requirements of the agents and automatically turns off their GPS units if they are closer than an optimally selected level to their private nodes. Note that this proposal is not the novel aspect of this paper. Many commercial softwares provide similar (but application-specific) solutions. For instance, Strava[4] enables users (mainly runners and cyclists) to construct a privacy zone (by specifying a position and radius) in which the GPS location of the user is not reported and is thus not revealed to a third-party[5]. Though undoubtedly powerful in preserving privacy of the users over one single trip, such zones might prove ineffective in the

The work of the authors was supported by a McKenzie Fellowship, ARC grant LP130100605, a grant from Melbourne School of Engineering.

The authors are with the Department of Electrical and Electronic Engineering, the University Melbourne, Parkville, VIC 3010, Australia.

[1] https://www.waze.com/

[2] http://traffic.berkeley.edu/

[3] A rational agent, which is privacy-conscious enough to not want its path to be recorded at all, simply does not participate in the scheme.

[4] http://www.strava.com/

[5] Thieves have used such measurements to steal valuable bikes. http://www.telegraph.co.uk/technology/news/11372189/Cycling-apps-put-you-at-risk-from-hi-tech-burglars.html

long run. For instance, imagine that the same user gets out of its privacy zone in three different locations (on three different occasions). Now, equipped with the knowledge that the user's house is in the same distance from these three positions (without even knowing that distance), a well-resourced attacker can simply pinpoint the position of the house of the user by triangulation. Therefore, more careful design strategies are required to protect the privacy of the users. This paper provides a *provably privacy-preserving* method for constructing the privacy zones based on the underlying transportation network.

In this paper, first, a measure of privacy infringement for the agents is introduced. The measure is equal to the conditional probability that an external observer assigns to the private nodes of the agent given all its position measurements over time. By carefully selecting the set of nodes and edges over which the agent does not transmit its position, it can effectively reduce this conditional probability to mislead the maximum likelihood filters that can be used to find the location of the private nodes. However, this can only be achieved at the cost of not reporting the position at some nodes in the network, which would reduce the quality of the estimation provided by the participatory-sensing scheme. An algorithm to find an optimal trade-off between these two competing interests is provided. The algorithm searches over a family of policies in which the agent stops broadcasting its position measurements if its distance in terms of the number of hops to the private nodes is smaller than a level. Employing such symmetric policies are advantageous as they are easy to compute and to implement. This is due to the fact that the agent only needs to count the number of the nodes on its path and does not require a map of the city. However, they have a drawback because an agent that uses a symmetric policy should exclude a larger subgraph from its measurements to achieve the same level of privacy as an agent that does not restrict itself to such policies. The results are subsequently expanded to more general (possibly asymmetric) policies. The effect of the population density is then explored. The impact of the betweenness measure of centrality of the private nodes of an agent on the policy of that agent is numerically explored on a random geometric graph. The betweenness measure of centrality for a node is defined as the number of the shortest paths in the graph that pass through that node divided by the total number of the shortest paths [6, p. 39]. Intuitively, a node with a high betweenness measure connects many nodes to each other (and is thus very central). However, a node with a low betweenness measure can be removed from the graph with no devastating consequences. It is observed numerically that, for the node with the maximum betweenness in a random geometric graph, a fairly large exclusion radius should be selected to reduce the measure of the privacy infringement. In addition, by slightly increasing the radius of the exclusion zone, the number of the nodes and edges over which the agent stop its broadcasting its position measurements rapidly increases. The opposite behaviour is witnessed for the node with the minimum betweenness. Finally, the results are demonstrated on the City of Melbourne (the center of the Melbourne metropolitan area).

The rest of the paper is organized as follows. This section is concluded by presenting some useful notations. Section II introduces the privacy measure and mathematically formulates the problem. Section III provides an algorithm for finding a symmetric policy that balances between the need for privacy and the quality of the estimation. Section IV extend the results to asymmetric policies. The effect of the heterogeneity of the population density is explored in Section V. The numerical examples are illustrated and their implications are discussed in Section VI. Finally, the paper is concluded in Section VII.

A. Notations

Consider an undirected $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set. An undirected edge between nodes $i, j \in \mathcal{V}$ is denoted by $\{i, j\}$. A *walk* is a sequence $v_1, e_1, v_2, e_2, \dots, e_{k-1}, v_k$ of vertices $v_i \in \mathcal{G}$ and edges $e_i \in \mathcal{E}$ such that $e_i = (v_i, v_{i+1})$ for all $i < k$. The vertices v_1 and v_k are called the *initial* and the *terminal* vertices of the walk, respectively. The *length* of the walk is the number of edges in it. A *path* is a walk in which all vertices are distinct. The *distance* $d(i, j)$ between two vertices $i, j \in \mathcal{V}$ is the length of a *shortest* path with initial vertex i and terminal vertex j. Additionally, define the set of nodes with a distance δ from node i by $\mathcal{D}_i(\delta)$ where $\mathcal{D}_i(\delta) = \{j | j \in \mathcal{V}, d(i, j) = \delta\}$. The length of the longest shortest path between any two vertices of a graph is the graph's *diameter*.

II. Problem Formulation

A *transportation network*, or simply a network, is modelled by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the nodes and the edges represent the intersections and the road segments, respectively. Consider a set of mobile agents, denoted by A, that traverse throughout the network along its edges. The agents are participating in a passive[6] participatory-sensing scheme to estimate the state of the traffic (i.e., the number of the vehicles[7] on each edge) in the network. Each agent $a \in A$ is assumed to visit all nodes in $\mathcal{V}$ at least once[8]. Any agent $a \in A$ can choose whether to broadcast its position or not. In case of deciding not to broadcast the position, it can simply turn off the GPS sensor on its connected device or terminate the participatory-sensing application. Assume that agent a does not want an external

[6] The term passive refers to the fact that the agents are not required to report traffic incidents (e.g., as in Waze). The participatory-sensing scheme receives GPS measurements of the agents at different times to construct an estimate of the traffic flow. An example of such a system can be found in [7].

[7] Note that there might be many more vehicles than agents since not everyone is participating in the sensing scheme.

[8] For this assumption to be satisfied, it is only required that the agents visit any node by a non-zero probability (no matter how small). Then, as time goes to infinity, the agents visit all nodes with probability one. Removing this assumption can make the observer's task at determining the private nodes of an agent harder. Therefore, this analysis can be used as a worst-case analysis of the privacy preserving policies for the agents.

observer (with access to all its position measurements) to be able to determine if it has ever visited a prescribed set of nodes denoted by $\mathcal{S}_a \subseteq \mathcal{V}$. This set is assumed to be only known by agent a. The nodes $s_a \in \mathcal{S}_a$ are termed the *private nodes* for agent a. For the sake of the simplicity of exposition, assume that $\mathcal{S}_a = \{s_a\}$ for some $s_a \in \mathcal{V}$. The results of this paper can be easily generalized to the case that $\mathcal{S}_a$ is not a singleton (so long as the elements of $\mathcal{S}_a$ are "far enough" from each other on the graph).

Assume that the observer has a uniform prior on privacy sensitive node s_a of agent a. This assumption is removed later in the paper to account for the heterogeneity of the population density. Given the positions that the agent has broadcast (and the time stamp of those positions), the observer can construct the conditional probability $p_a(v)$ that node v belongs to the set $\mathcal{S}_a$. Agent $a \in A$ wants to keep the conditional probability $\pi_a := p_a(s_a)$ for $s_a \in \mathcal{S}_a$ small. This way, the observer does not have a good chance for correctly inferring that the agent has visited the node s_a. Therefore, this conditional probability can be seen as a measure of privacy infringement for agent a. However, the agents want to achieve this goal with not drastically degrading the performance of the participatory-sensing scheme. Note that, if such a constraint is not enforced, the best policy of an agent is to never broadcast its position.

Throughout this paper, agent a is assumed to follow a policy that instructs it to stop broadcasting its position when it is on a link $\{i, j\}$ if $d(i, s_a) \leq h_a$ or $d(j, s_a) \leq h_a$ for $s_a \in \mathcal{S}_a$, where $h_a \in \mathbb{N}$ is a prescribed integer number that is only known by agent a. The goal is to find h_a optimally. Such a policy is favoured (in comparison to more general asymmetric ones) as it can be easily implemented: it does not require a map of the city. The agent simply stops broadcasting when its hop-distance to its private node, s_a, is smaller than h_a.

As described earlier, the agents do not want to drastically degrade the quality of the participatory-sensing scheme. Let $\mathcal{E}(s_a, h_a)$ denote the set of edges over which agent a stop broadcasting its position. It can be shown that

$$\mathcal{E}(s_a, h_a) = \{\{i, j\} \in \mathcal{E} \,|\, d(i, s_a) \leq h_a \vee d(j, s_a) \leq h_a\}.$$

This is inversely proportional to the quality of the estimation as, in many sensing schemes (e.g., the least mean square method), the covariance of the estimation error reduces by increasing the number of the measurements. Therefore, agent a may wish to keep $r_a := |\mathcal{E}(s_a, h_a)|$ relatively small.

Problem 1: For any agent $a \in A$, find h_a to minimize $\pi_a + \gamma_a r_a$, where the constant $\gamma_a > 0$ is inversely proportional to the importance of privacy for agent a.

For very large constant γ_a, the cost of the agent behaves similarly to r_a. Therefore, the best policy of the agent is to report on all edges and, thus, select $h_a = 0$. However, for very small γ, the cost of the agent is mostly determined by π_a. Thus, the best policy of the agent is to never broadcast its position or, equivalently, set h_a to be larger than the diameter of $\mathcal{G}$. Alternatively, the agent can solve the following problem. This problem formulation avoids the use of a non-intuitive parameter γ_a.

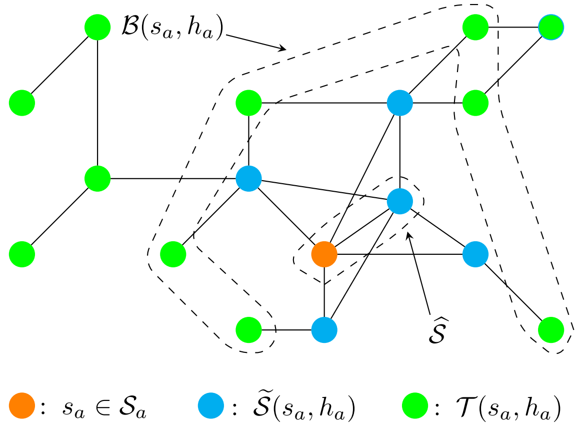
$\mathcal{B}(s_a, h_a)$
$\widehat{\mathcal{S}}$
●: $s_a \in \mathcal{S}_a$
●: $\widetilde{\mathcal{S}}(s_a, h_a)$
●: $\mathcal{T}(s_a, h_a)$

Fig. 1. An illustrative example of the sets $\mathcal{S}_a$, $\mathcal{T}(s_a, h_a)$, $\widetilde{\mathcal{S}}(s_a, h_a)$, $\mathcal{B}(s_a, h_a)$, and $\widehat{\mathcal{S}}$ with $h_a = 1$.

Problem 2: For any agent $a \in A$, find h_a to minimize r_a subject to $\pi_a \leq \xi_a$, where the constant $\xi_a \in [0, 1]$.

If $\xi_a \leq 1/|\mathcal{V}|$, the solution of Problem 2 is to select h_a larger than the diameter of $\mathcal{G}$. This is because, upon providing any position measurements, the conditional probability π_a can only be increased (i.e., only private information can be leaked by providing measurements). For $\xi_a = 1$, the best policy of the agent is to select $h_a = 0$. In this case, the agent has a low sensitivity to privacy infringement. A good aspect of this problem formulation is that the constant ξ_a can be intuitively selected based on the meaning of the maximum tolerable conditional probability π_a.

III. Privacy-Preserving Policy

In the light of the aforementioned problem formulation, an external observer that collects the position broadcasts of agent a can form the set of nodes $\mathcal{T}(s_a, h_a) \subseteq \mathcal{V}$ in which agent a has transmitted its position. This is because it is assumed that each agent visits all the nodes in the network. If the observer waits long enough, the set of all the nodes that agent a has visited converges to $\mathcal{T}(s_a, h_a)$ with probability one. Note that

$$\mathcal{T}(s_a, h_a) = \{i \in \mathcal{V} \,|\, d(i, s_a) \geq h_a + 1\}. \quad (1)$$

See Fig. 1 for an illustrative example of this set. The following immediately follows:

$$s_a \in \widetilde{\mathcal{S}}(s_a, h_a) := \mathcal{V} \setminus \mathcal{T}(s_a, h_a). \quad (2)$$

This is a direct consequence of the selected set of policies. However, due to the nature of the policy of the agents, the observer can further remove points from $\widetilde{\mathcal{S}}(s_a, h_a)$ that cannot possibly belong to $\mathcal{S}_a$. This is investigated in the remainder of this section. The boundary of the set $\mathcal{T}(s_a, h_a)$ interfacing with the set $\widetilde{\mathcal{S}}(s_a, h_a)$ can be defined as

$$\mathcal{B}(s_a, h_a) = \{i \in \mathcal{T}(s_a, h_a) \,|\, \mathcal{D}_i(1) \cap \widetilde{\mathcal{S}}(s_a, h_a) \neq \emptyset\}, \quad (3)$$

where, for any $m \in \mathbb{N}$ and any $i \in \mathcal{V}$, the set $\mathcal{D}_i(m)$ denotes the set of nodes $j \in \mathcal{V}$ such that $d(i,j) = m$ (i.e., that are of distance m to node i). Similarly, define

$$\widehat{\mathcal{S}} := \Bigg\{ s \in \widetilde{\mathcal{S}}(s_a, h_a) \,|\, \exists \delta \in \mathbb{N}_{[1,\ell_a+1]} : \mathcal{D}_s(\delta) = \mathcal{B}(s_a, h_a) \wedge \bigcup_{0 \leq \delta' \leq \delta - 1} \mathcal{D}_s(\delta') = \widetilde{\mathcal{S}}(s_a, h_a) \Bigg\}. \quad (4)$$

where ℓ_a is the diameter of the subgraph induced by $\widetilde{\mathcal{S}}(s_a, h_a)$. Among all the subsets of $\mathcal{V}$ that can be picked by the observer, $\widehat{\mathcal{S}}$ described by (4) is the smallest set that is guaranteed to include s_a. The following results can be proved.

Theorem 1: $\pi_a = 1/|\widehat{\mathcal{S}}|$.

Proof: The Bayes' rule (e.g., see [8]) dictates that

$$\begin{aligned} \mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} &\propto \mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\} \mathbb{P}\{s \in \mathcal{S}_a\} \\ &\propto \mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\}, \end{aligned} \quad (5)$$

where the notation $\propto$ shows that the both sides are proportional to each other (i.e., they are equal to each other if one is multiplied by an appropriate constant). Now, note that

$$\begin{aligned} &\mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\} \\ &\qquad = \mathbb{1}_{\exists h \in \mathbb{N} : \mathcal{T}(s,h) = \mathcal{T}(s_a, h_a)} \\ &\qquad = \mathbb{1}_{\exists h \in \mathbb{N} : \mathcal{B}(s,h) = \mathcal{B}(s_a, h_a) \wedge \widetilde{\mathcal{S}}(s_a, h_a) = \widetilde{\mathcal{S}}(s,h)}, \end{aligned}$$

where $\mathbb{1}$ is a characteristic function, i.e., $\mathbb{1}_p$ is equal to one if the statement p holds true and is equal to zero otherwise. By definition, $\mathcal{B}(s,h) = \mathcal{D}_s(h+1)$ and $\widetilde{\mathcal{S}}(s,h) = \bigcup_{h' \in \mathbb{N} : h' \leq h} \mathcal{D}_s(h')$. Therefore, it can be deduced that

$$\mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\} = \mathbb{1}_{s \in \widehat{\mathcal{S}}}. \quad (6)$$

Substituting (6) into (5) gives

$$\mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} \propto \mathbb{1}_{s \in \widehat{\mathcal{S}}}.$$

Noting that $\sum_{s \in \mathcal{V}} \mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} = 1$ results in

$$\mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} = \frac{1}{|\widehat{\mathcal{S}}|} \mathbb{1}_{s \in \widehat{\mathcal{S}}}.$$

This concludes the proof. ■

In this case, it can be proved that $\mathcal{E}(s_a, h_a) = \{\{i,j\} \in \mathcal{E} \,|\, i \in \widetilde{\mathcal{S}}(s_a, h_a) \vee j \in \widetilde{\mathcal{S}}(s_a, h_a)\}$. Hence, Problem 1 can be cast as

$$\min_{0 \leq h_a \leq d_{\mathcal{G}}} \frac{1}{|\widehat{\mathcal{S}}|} + \gamma_a |\{\{i,j\} \in \mathcal{E} | i \in \widetilde{\mathcal{S}}(s_a, h_a) \vee j \in \widetilde{\mathcal{S}}(s_a, h_a)\}|,$$

where $d_{\mathcal{G}}$ denotes the diameter of the graph $\mathcal{G}$. For each s_a, this problem can be easily solved by calculating the cost function for all applicable h_a and, subsequently, by selecting h_a corresponding to the smallest value. To calculate the cost, sets $\mathcal{D}_i(\delta)$ for all $i \in \mathcal{V}$ and δ need to be calculated. These sets can be determined by calculating Δ^δ with Δ denoting the adjacency matrix, i.e., $\Delta_{ij} = 1$ if there exists an edge $\{i,j\} \in \mathcal{G}$. All these sets for δ upto $d_{\mathcal{G}}$, can be computed by $\mathcal{O}(|\mathcal{V}|^3 d_{\mathcal{G}})$ operations. Noting that, at worst case $d_{\mathcal{G}} = \mathcal{O}(|\mathcal{V}|)$, all these sets can be computed by $\mathcal{O}(|\mathcal{V}|^4)$. Therefore, the set $\widehat{\mathcal{S}}$ can be computed by $\mathcal{O}(|\mathcal{V}|^6)$ operations because, at worst case, $|\widetilde{\mathcal{S}}(s_a, h_a)| = \mathcal{O}(|\mathcal{V}|)$. Finally, to perform all these operations for all h_a, at most $\mathcal{O}(|\mathcal{V}|^7)$ operations are required. These complexity calculations are very conservative and, for most graphs, much fewer operations are required.

Alternatively, Problem 2 can be cast as

$$\begin{aligned} \min_{0 \leq h_a \leq d_{\mathcal{G}}} \; & |\{\{i,j\} \in \mathcal{E} \,|\, i \in \widetilde{\mathcal{S}}(s_a, h_a) \vee j \in \widetilde{\mathcal{S}}(s_a, h_a)\}|, \\ \text{s.t.} \; & |\widehat{\mathcal{S}}| \geq 1/\xi_a \end{aligned}$$

Because $|\{\{i,j\} \in \mathcal{E} \,|\, i \in \widetilde{\mathcal{S}}(s_a, h_a) \vee j \in \widetilde{\mathcal{S}}(s_a, h_a)\}|$ is an increasing function of h_a, the optimal solution is to select h_a to be the smallest element of the set $\{h_a \,|\, |\widehat{\mathcal{S}}| \geq 1/\xi_a\}$. This can also be done by checking all h_a with at most $\mathcal{O}(|\mathcal{V}|^7)$ operations.

IV. Asymmetric Policies

In the previous section, the only viable family of policies for agent a is to stop transmitting its position measurements if its distance in terms of the number of hops to the privacy sensitive node s_a is smaller than or equal to h_a. The symmetry of the policy of the agents was utilized by the observer to further reduce the uncertainty of predicting the privacy sensitive node of the agents. This is clearly a drawback because, for achieving the same level of privacy, an agent that uses a symmetric policy must exclude a larger subgraph from its measurements. However, employing symmetric policies are also advantageous as they can be computed efficiently and their implementation is simpler than the asymmetric ones (the agent only needs to count the number of the nodes on its path).

To formalize this intuition, the set of applicable policies of the agents is temporarily generalized. Agent a is assumed to follow a policy that instruct its connected device to stop broadcasting its position when it is on a link $\{i,j\}$ if $i \in \mathcal{N}_a$ or $j \in \mathcal{N}_a$. It is assumed that $s_a \in \mathcal{N}_a$. Such an assumption is needed because the observer can deduce that a node is of special importance to an agent, if it stops there for a significant amount of time (as it is not only a way point on the path). The goal is to find the set $\mathcal{N}_a$ optimally. Similar to the previous section, an external observer that collects the position broadcasts of agent a can form the set of nodes $\mathcal{T}(s_a, h_a) \subseteq \mathcal{V}$ in which agent a has transmitted its position. Clearly, $\mathcal{N}_a = \mathcal{V} \setminus \mathcal{T}(s_a, h_a)$. The following result can be proved.

Theorem 2: $\pi_a = 1/|\mathcal{N}_a|$.

Proof: Similar to the proof of Theorem 1, the Bayes' rule can be used to show that

$$\begin{aligned} \mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} &\propto \mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\} \mathbb{P}\{s \in \mathcal{S}_a\} \\ &\propto \mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\}. \end{aligned}$$

Now, note that $\mathbb{P}\{\mathcal{T}(s_a, h_a) \,|\, s \in \mathcal{S}_a\} = \mathbb{1}_{s \in \mathcal{N}_a}$. Therefore,

$$\mathbb{P}\{s \in \mathcal{S}_a \,|\, \mathcal{T}(s_a, h_a)\} = \frac{1}{|\mathcal{N}_a|} \mathbb{1}_{s \in \mathcal{N}_a}.$$

This concludes the proof. ■

Problem 1 can then be cast as

$$\min_{\mathcal{N}_a\in 2^{\mathcal{V}}:s_a\in\mathcal{N}_a} \frac{1}{|\mathcal{N}_a|}+\gamma_a|\{\{i,j\}\in\mathcal{E}\,|\,i\in\mathcal{N}_a\vee j\in\mathcal{N}_a\}|.$$

Unfortunately, this optimization problem is very hard solve as the number of all the possible sets $\mathcal{N}_a\in 2^{\mathcal{V}}$ such that $s_a\in\mathcal{N}_a$ grows exponentially with the number of the number of the nodes. A similar argument can also be presented for Problem 2 and is thus removed.

If, in order to reduce the complexity of the problem, the search is conducted on the set of symmetric policies defined to be set of all policies of the form $\mathcal{N}_a=\{s\in\mathcal{V}\,|\,d(s,s_a)\le h_a\}$ for $h_a\in\mathbb{N}$, the results of the previous section can be recovered. In this case, it can be also proved that

$$\frac{1}{|\mathcal{N}_a|}\le\frac{1}{|\widehat{\mathcal{S}}|}$$

because $\widehat{\mathcal{S}}\subseteq\widetilde{\mathcal{S}}(s_a,h_a)=\mathcal{N}_a$. Therefore, it can be inferred that the symmetric polices minimize an upper-bound on the costs over a smaller set of policies, i.e., the set of asymmetrical policies. This results in their superior efficiency in terms of computation and implementation.

V. Varying Density

In most urban areas, the density of the population is not homogeneous throughout the city. Let $\rho:\mathcal{V}\rightarrow\mathbb{R}_{\geq 0}$ be a mapping that determines the density around any node. Following this observation, the observer can no longer assume a uniform prior on the private nodes. Therefore, $\mathbb{P}\{s\in\mathcal{S}_a\}=\rho(s)/\sum_{v\in\mathcal{V}}\rho(v)$. This captures the fact that it is more likely that the privacy sensitive node s_a of agent a to belong to a densely populated area. In this case, the following result can be proved.

Theorem 3: $\pi_a=\rho(s)/\sum_{v\in\widehat{\mathcal{S}}}\rho(v)$.

Proof: Similar to the proof of Theorem 1, it can be shown that

$$\begin{aligned}\mathbb{P}\{s\in\mathcal{S}_a\,|\,\mathcal{T}(s_a,h_a)\}\propto&\mathbb{P}\{\mathcal{T}(s_a,h_a)\,|\,s\in\mathcal{S}_a\}\mathbb{P}\{s\in\mathcal{S}_a\}\\ \propto&\mathbb{P}\{\mathcal{T}(s_a,h_a)\,|\,s\in\mathcal{S}_a\}\rho(s)\\ =&\mathbb{1}_{s\in\widehat{\mathcal{S}}}\rho(s).\end{aligned}$$

As a result,

$$\mathbb{P}\{s\in\mathcal{S}_a\,|\,\mathcal{T}(s_a,h_a)\}=\frac{\rho(s)}{\sum_{v\in\widehat{\mathcal{S}}}\rho(v)}\mathbb{1}_{s\in\widehat{\mathcal{S}}}.$$

This concludes the proof. ■

Following the result of Theorem 3, the problems that the agents need to solve can be adapted and the algorithm in Section III can be used to find an optimal privacy-preserving policy.

Note that a "sensible" agent should also put less emphasize on protecting its privacy if its privacy sensitive node is in a densely populated area. This is because, in densely populated areas, even if the observer pinpoints the location of an agent upto a node, there are many residential and commercial areas in the surrounding that makes identifying the physical location of the agent impossible. Such a behaviour can be reflected in the selection of the term γ_a (if the agent makes such a decision).

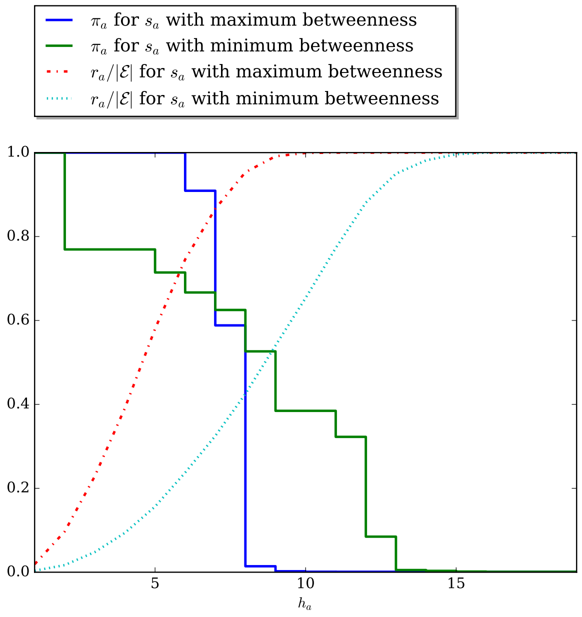

Fig. 2. π_a and r_a as a function of h_a for nodes with the maximum and the minimum betweenness averaged over ten random geometric graphs with 1000 nodes in unit rectangle and connectivity radius of 0.1.

VI. Numerical Example

First, the results of Section III are demonstrated on random graphs. Consider a graph with $|\mathcal{V}|=1000$ nodes. The positions of the nodes are distributed uniformly inside a unit rectangle $[0,1]\times[0,1]$. If the Euclidean distance between two nodes is smaller than or equal to 0.1, the nodes are assumed to be connected. This creates an undirected graph, which is referred to as a random geometric graph; see [9] for more information.

The values of π_a and r_a (as function of h_a) are only illustrated for the nodes with the maximum and the minimum betweenness measures (as two extreme behaviours). The betweenness measure of centrality for a node is defined as the number of the shortest paths in the graph that pass through that node divided by the number of all the shortest paths in the graph. Intuitively, a node with a relatively large betweenness measure connect many nodes to each other. However, a node with a small betweenness measure can be removed from the graph with no devastating consequences. Fig. 2 shows π_a and r_a as a function of h_a for nodes with the maximum and the minimum betweenness averaged over ten random geometric graphs. Based on the illustrated numerical example, it can be seen that, for the node with the maximum betweenness, a fairly large h_a should be selected to reduce π_a initially. This is because for these nodes the set $\mathcal{B}(s_a,h_a)$ is very large and thus $\widehat{\mathcal{S}}$ most often contains only a few nodes. In addition, by increasing h_a, r_a rapidly increases. This is because the central nodes are often very close to all the nodes (as many shortest paths go through them). The reverse behaviour is observed for the node with the minimum betweenness.

Transportation networks are most often highly structured

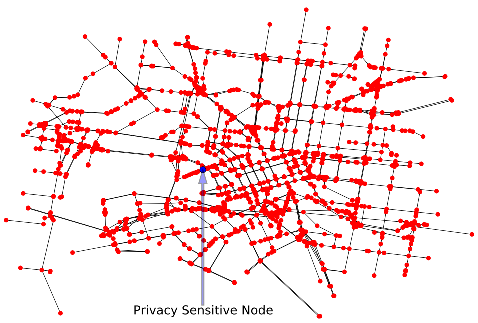

Fig. 3. The roads in the City of Melbourne

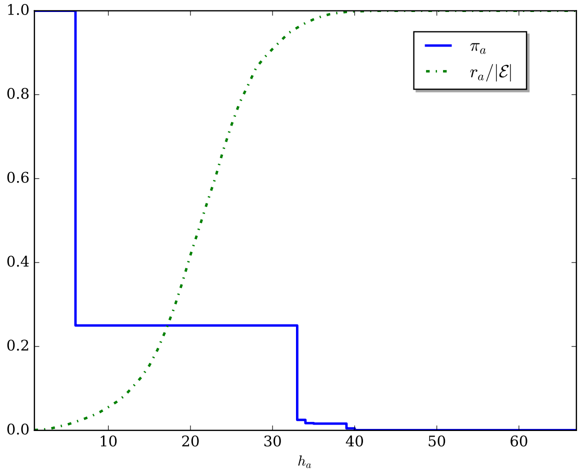

Fig. 4. Different values of π_a versus the ratio of the nodes where the location is not reported for different values of h_a.

graphs as their design is not random (or at least one hopes) and follows careful consideration by the local governments. Therefore, in the next simulation, a scenario in the City of Melbourne. The goal is to ascertain the achievable levels of privacy for an agent starting his travel from the dark blue node[9] in Fig. 3. The possible levels of privacy infringement are depicted in Figure 4. Evidently, with a fairly low h_a, p_a can be reduced drastically. This is also achieved at a relatively low cost because r_a is still small.

VII. Conclusions and Future Work

An algorithm for finding an optimal policy for preserving the privacy of the agents in a participatory-sensing scheme for traffic estimation is presented. The effect of the betweenness measure of centrality on the policy of the agents is numerically explored. In can be seen that, for the node with the maximum betweenness, a fairly large exclusion radius should be selected to enhance the privacy. In addition, by slightly increasing radius, the number of the nodes over which the agent stop its measurement transmission rapidly increases. The reverse behaviour is seen for the node with the minimum betweenness. Finally, note that if the agents want to hide their position on some edges (denoted by the *privacy-sensitive edges*), the same results can be used. To do so, the definition of the graph representing the transportation system should be modified. Specifically, to find the optimal policy, the line graph (see [10, pp. 71–82]) of the transportation network should be constructed. A line graph of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph $L(\mathcal{G})$ with the vertex set $\mathcal{V}_{L(\mathcal{G})} = \mathcal{E}$ and the edge set $\mathcal{E}_L(\mathcal{G}) = \{(e, \bar{e}) \in \mathcal{V}_{L(\mathcal{G})} \times \mathcal{V}_{L(\mathcal{G})} \,|\, i = \bar{i} \vee j = \bar{j} \text{ for } e = \{i, j\}, \bar{e} = \{\bar{i}, \bar{j}\}\}$. In the line graph of the transportation, the roads represent the nodes and two roads are connected to each other by an edge (in the line graph) if they intersect.

The future work can focus on understanding the effect of determined policies using simulators for transportation systems to measure the quality of the participatory-sensing schemes. Also financial incentives can be designed to improve the estimation quality in different areas of the city.

References

[1] C. Dwork, “Differential privacy: A survey of results,” in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi’an, China, April 25-29, 2008. Proceedings*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer, 2008, pp. 1–19.

[2] R. Chen, B. Fung, B. C. Desai, and N. M. Sossou, “Differentially private transit data publication: A case study on the Montreal transportation system,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 213–221.

[3] F. Kargl, A. Friedman, and R. Boreli, “Differential privacy in intelligent transportation systems,” in *Proceedings of the 6th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2013, pp. 107–112.

[4] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, “Entropy-minimizing mechanism for differential privacy of discrete-time linear feedback systems,” in *Proceedings of the IEEE 53rd Annual Conference on Decision and Control*, 2014, pp. 2130–2135.

[5] J. Le Ny and G. J. Pappas, “Differentially private Kalman filtering,” in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1618–1625.

[6] M. O. Jackson, *Social and Economic Networks*, ser. Princeton University Press. Princeton University Press, 2010.

[7] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, “Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment,” *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568 – 583, 2010.

[8] R. Durrett, *Probability: Theory and Examples*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[9] M. Penrose, *Random Geometric Graphs*, ser. Oxford Studies in Probability. Oxford University Press, 2003.

[10] F. Harary, *Graph Theory*, ser. Addison-Wesley Series in Mathematics. Addison-Wesley Publishing Company, 1969.

[9] The interest in this node is not arbitrary as it pinpoints the position of a former residence of one of the authors.