This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2698424, IEEE Transactions on Information Forensics and Security

Optimal Privacy-Preserving Probabilistic Routing for Wireless Networks

Jing Yang Koh, Derek Leong, Gareth W. Peters, Ido Nevat, and Wai-Choong Wong

***Abstract*—Privacy-preserving routing protocols in wireless networks frequently utilize additional artificial traffic to hide the source-destination identities of the communicating pair. Usually, the addition of artificial traffic is done heuristically with no guarantees that the transmission cost, latency, etc., are optimized in every network topology. In this paper, we explicitly examine the privacy-utility trade-off problem for wireless networks and develop a novel privacy-preserving routing algorithm called Optimal Privacy Enhancing Routing Algorithm (OPERA). OPERA uses a statistical decision-making framework to optimize the privacy of the routing protocol given a utility (or cost) constraint. We consider global adversaries with both lossless and lossy observations that use the Bayesian maximum-a-posteriori (MAP) estimation strategy. We formulate the privacy-utility trade-off problem as a linear program which can be efficiently solved. Our simulation results demonstrate that OPERA reduces the adversary's detection probability by up to 50% compared to the random Uniform and Greedy heuristics, and up to five times compared to a baseline scheme. In addition, OPERA also outperforms the conventional information-theoretic mutual information approach.**

***Index Terms*—Location privacy, privacy-utility trade-off, probabilistic routing, Bayesian traffic analysis, wireless routing.**

I. Introduction

Traffic analysis attacks [1]–[7] are a serious threat to the privacy of users in a communication system. The analysis attacks can be used to infer sensitive contextual information (e.g., source-destination identities) from observed traffic patterns. More worryingly, they are easily executed without raising suspicions in a multihop wireless network where the node transmissions can be passively observed. Hence, extensive research efforts have been invested in mitigating traffic analysis attacks in wireless networks. Typical traffic analysis techniques exploit features such as packet timings, sizes or counts to *correlate traffic patterns* and compromise user privacy.

Three common approaches to mitigate analysis attempts are to: (i) change the physical appearance of each packet at every hop via hop-by-hop encryptions [4], [8], [9], (ii) introduce transmission delays at each hop [2], [10] to decorrelate traffic flows, or (iii) introduce dummy traffic [3], [5], [11]–[15] to obfuscate traffic patterns. The first two approaches may not be desirable for low-cost or battery-powered wireless networks, e.g., wireless sensor networks as (i) the low-cost nodes may not be able to afford using the computationally expensive encryptions at each hop, and (ii) introducing delays at the intermediate nodes may not be effective when there is little traffic in the network. Therefore, we use the dummy traffic approach to provide privacy by lowering the adversary's detection rates (formally defined in Section III) in a wireless network. Specifically, we consider an adversary that uses the optimal maximum-a-posteriori (MAP) estimation strategy.

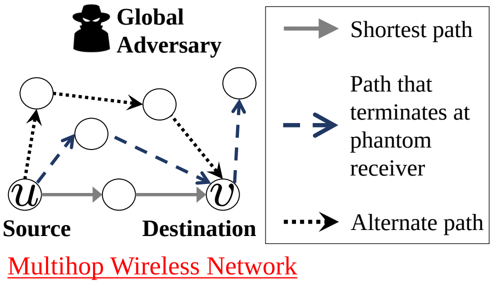


Fig. 1. Suppose there exist three possible routing paths from the source node u to the destination node v. The source has to select a path distribution over the three possible paths to its destination that minimizes the average detection probability of a global adversary who is able to observe the node transmissions.

We focus on hiding the *source-destination identities* (or *unlinkability* [3], [4], [16]) of each communication where a global adversary is able to observe node transmissions from the entire network (see Fig. 1). Our challenge is to decide how to probabilistically route the packets from the source to the destination nodes via carefully chosen *(proxy) receiver* nodes to preserve privacy. For example, consider the network in Fig. 1 where there exist three possible routing paths from the source node u to the destination node v. Even though it is desirable to maximize the amount of privacy for each communicating party, this would usually require a flooding-based solution (e.g., by using all three available paths) which is undesirable due to its high network resource consumption. Hence, we present the Optimal Privacy Enhancing Routing Algorithm (OPERA) which uses a statistical decision-making framework to characterize different network scenarios and select the optimal path distribution that strikes a balance between the *privacy* and *utility* (e.g., in terms of transmission cost) of the routing protocol given some *privacy budget* (e.g., transmission cost constraint). Additional dummy traffic may also be used to extend the routing path to include additional receiver nodes (nodes that received the dummy traffic).

The statistical decision-making framework approach extends our earlier work in [3] where a heuristic probabilistic routing algorithm was proposed to enhance the privacy for the destination node. In this work, we consider a relatively stronger adversary that uses the Bayesian MAP estimation strategy and also consider the case where the adversary has

J. Y. Koh and D. Leong are with the Institute for Infocomm Research (I^{2}R), A*STAR, Singapore (e-mail: kohjy@i2r.a-star.edu.sg; dleong@i2r.a-star.edu.sg). G. W. Peters is with the Department of Statistical Science, University College London (e-mail: garethpeters78@gmail.com). I. Nevat is with TUMCREATE, Singapore (e-mail: idonevat@gmail.com). W.-C. Wong is with the Department of Electrical and Computer Engineering, National University of Singapore (e-mail: elewwcl@nus.edu.sg).

The work of J. Y. Koh was supported in part by the Agency for Science, Technology and Research (A*STAR) Graduate Scholarship.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2698424, IEEE Transactions on Information Forensics and Security

lossy observations. We formulate the selection of the optimal *privacy-preserving* paths for each source-destination pair using a statistical decision-making framework that results in a linear program which is easily solved by many commercial solvers.

A. Main Contributions

Our main contributions are as follows:

- We propose a statistical decision-making framework to optimize the privacy-utility trade-off for routing in wireless networks against a global and informed adversary using the Bayesian maximum-a-posteriori (MAP) estimation strategy. We then formulate linear programs to efficiently compute the optimal privacy-preserving paths under the lossless and lossy adversarial models, given a privacy budget.
- We study the choice of our objective function (minimizing the adversary's detection probability) and how it differs from minimizing mutual information or using the Uniform and Greedy heuristics.
- We propose a low-complexity approximation method to compute the optimal privacy-preserving paths under the lossy adversarial model.
- We demonstrate via simulations that privacy does not necessarily depend on the number of receivers as the communication patterns are more important. We also evaluate our approach in several different network topologies, including two real-world testbeds.

A motivating example for our proposed framework, and other detailed examples, along with the MATLAB code can be found in the extended version of our paper [17].

II. Related Work

Anonymity enhancing techniques like onion routing [9] and mix-net [8] allow users to anonymously communicate over the wired Internet network. These techniques mostly rely on packet encryption and randomized routing from the source to the destination to hide sensitive information (e.g., the nodes' identities) from eavesdropping adversaries. The onion routing offers privacy protection from an adversary with only local observability of the network while the mix-net provides privacy even against adversaries with global observability via special mix nodes. However, the onion routing technique is more prevalent due to its lower latency which makes it practical. Fortunately, the local observability assumption is valid in the large-scale Internet. In contrast, the relatively smaller wireless networks are more vulnerable to traffic analysis from a global adversary. In addition, due to the wireless broadcast medium, it is possible for an adversary to passively eavesdrop on all transmissions from a wireless node without being detected.

To address such problems, the field of location privacy emerged with the first *location privacy* problem (specifically the source-location privacy problem) for wireless networks being studied by Ozturk *et al.* [11]. The authors proposed several flooding-based routing techniques, including the phantom flooding routing to prevent local adversaries from tracing a packet back to its source. Since the flooding-based solution is inherently expensive, several other works [18], [19] have built on the random walk-based routing strategy and improved its effectiveness and efficiency. A thorough survey on source-location privacy can be found in [6]. Interestingly, the work in [2] used a periodic flooding approach for privacy protection with statistical guarantees. Subsequently, Jian *et al.* [12] devised a protocol to protect the receiver's location privacy from packet-tracing attacks by using path diversity to decorrelate the incoming and outgoing traffic at each node.

A stronger global adversary which can observe transmissions in the entire network was considered by [14]. The authors proposed a periodic collection and source simulation (dummy sources) techniques for providing source location privacy and the backbone flooding and sink simulation (dummy sinks) techniques for receiver location privacy. In [20], the authors designed a packet transmission protocol based on random route generation and dummy packet transmissions that is secure against internal adversaries who can view the routing tables of the nodes. In [13], the authors proposed that the destination node randomly forwards some of the packets it receives to a randomly selected neighbor node located M hops away from the destination. A heuristic probabilistic routing algorithm was also used against the global adversary in [3]. Lastly, the work in [21] proposed a cloud-based scheme for enhancing the source node privacy and [22] used symmetric-key-cryptography operations and trapdoor techniques to develop a secure and privacy-preserving communication protocol.

Limitations of Current Heuristic Algorithms: It is evident that the privacy-enhancing schemes do not come for free and there exists a trade-off between the privacy and transmission overheads incurred. Although the above schemes have mainly relied on additional dummy traffic (or/and delays) to mitigate traffic analysis attempts, there is no rigorous quantification of the adversary's detection probability, their optimal attacking strategy, and the overheads incurred by the scheme. Hence, it would be interesting to quantify the loss of utility (or cost) incurred by the privacy-preserving scheme and weigh it against the additional amount of privacy provided. The work in [23] designed an optimal route selection strategy that maximizes the sender anonymity for the Internet and formulated an optimization problem to determine a path length distribution that maximizes the anonymity degree (a function of Shannon's entropy) of a system. Different from [23], we formulate a statistical decision-making framework and use a more direct (non-information-theoretic) privacy metric for our objective function. This paper extends the heuristic probabilistic routing algorithm in our earlier work [3] by using a statistical decision-making framework to compute the optimal privacy-preserving paths for a given privacy budget (or cost constraint).

III. System Model

We consider the scenario where a source node u wants to send packets to a single destination node v in a static wireless network. The source node uses a source routing protocol (e.g., dynamic source routing) and specifies a routing path from itself to the destination (see Definition 1). Due to the wireless broadcast nature of the network, when a node transmits, all its one-hop neighbors are able to receive the transmission.

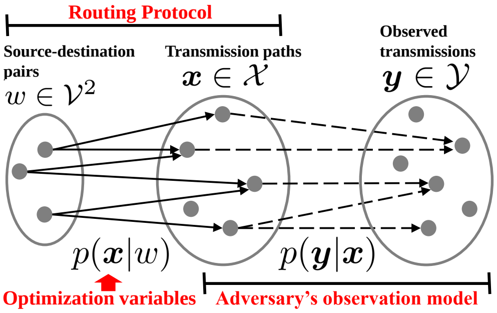


Fig. 2. Illustration of a probabilistic routing scheme that maps a source-destination pair $w \in \mathcal{V}^2$ to a set of transmission paths $\boldsymbol{x} \in \mathcal{X}$. The latter is related to the set of observed node transmissions $\boldsymbol{y} \in \mathcal{Y}$ via the adversary's observation model $p(\boldsymbol{y}|\boldsymbol{x})$. Ideally, knowledge of $\boldsymbol{y}$ should not immediately reveal w, e.g., there should be a many-to-one mapping from w to $\boldsymbol{y}$.

Next, we introduce the graph notations used in the paper:

a. Let the wireless network be modeled as a connected hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{H})$ where $\mathcal{V}$ is the set of nodes and $\mathcal{H}$ is the set of (directed) hyperarcs. A hyperarc $h = (s, \mathcal{R})$ represents a source-receivers pair where $s \in \mathcal{V}$ is the source node and $\mathcal{R} \subseteq \mathcal{V}$ is a non-empty set of receiver nodes adjacent to s.

b. Let $w \triangleq (u, v)$ represent the source-destination pair, where $u \in \mathcal{V}$, $v \in \mathcal{V}$ are the source and destination nodes respectively.

c. Let $\boldsymbol{x} = (h_1, h_2, \ldots)$ be the actual transmission path comprising the distinct hyperarcs h_i and the source node of h_{i+1} must be a receiver node of h_i. Let $\boldsymbol{y}$ be the observed path where $\boldsymbol{y}$ is a subvector of $\boldsymbol{x}$. An observer may not necessarily observe all the hyperarc transmissions in $\boldsymbol{x}$ as some of them may be erased (i.e., lossy observations). The ordering of the observed transmissions, however, remains unchanged.

d. Let $\mathcal{X}$ represent the set of all possible paths $\boldsymbol{x}$ in the network and let $\mathcal{X}^w$ be the set of all paths $\boldsymbol{x} = (h_1, h_2, \ldots)$ that serve the source-destination pair $w = (u, v)$, i.e., $h_1 = (u, \mathcal{R})$ and there exists an $h = (s, \mathcal{R}) \in \boldsymbol{x}$ such that $v \in \mathcal{R}$. Let $\mathcal{Y}$ represent the set of all possible observations $\boldsymbol{y}$.

e. Let $c_h \geq 0$ represent the cost (e.g., transmission cost) for using hyperarc h.

Definition 1 (Routing Protocol)**.** *Given a network graph $\mathcal{G}$, a probabilistic source-routing protocol selects a path $\boldsymbol{x} \in \mathcal{X}^w$ according to a path distribution $p(\boldsymbol{x}|w)$ for a given source-destination pair $w \in \mathcal{V}^2$.*

Optimized Probabilistic Routing: We focus on protecting the privacy of the source-destination identities (see Definition 2) by designing a *probabilistic privacy-preserving* routing protocol to minimize *the probability of an adversary correctly guessing the source-destination identities*. In addition, the routing scheme (see Fig. 2) should consider the adversary's *observation model* $p(\boldsymbol{y}|\boldsymbol{x})$ while computing a (routing) path distribution $p(\boldsymbol{x}|w)$ that serves the source-destination pair w.

A. Adversary Model

We consider an *external, passive, global* and *informed* [2] adversary who observes a (possibly lossy) sequence of transmissions $\boldsymbol{y}$ from an actual transmission path $\boldsymbol{x}$. Using a Bayesian traffic analysis technique, the adversary aims to *detect the identity of the source-destination pair w for each observation $\boldsymbol{y}$*, i.e., he aims to identify which node is talking to which node based on his possibly imperfect observations.

TABLE I
NOTATION

$\mathcal{G}$	connected hypergraph representing the network.
$\mathcal{V}$	set of all nodes in the network.
$\mathcal{H}$	set of all (directed) hyperarcs in the network.
$h = (s, \mathcal{R})$	hyperarc which represents a source-receivers pair where $s \in \mathcal{V}$ is the source node and $\mathcal{R} \subseteq \mathcal{V}$ is a non-empty set of receiver nodes adjacent to s.
$w \triangleq (u, v)$	source-destination pair where $u \in \mathcal{V}$, $v \in \mathcal{V}$ are the source and destination nodes respectively.
$\boldsymbol{x} = (h_1, h_2, \ldots)$	actual transmission path.
$\boldsymbol{y}$	observed path where $\boldsymbol{y}$ is a subvector of $\boldsymbol{x}$.
$\mathcal{X}$	set of all possible paths $\boldsymbol{x}$ in the network.
$\mathcal{X}^w$	set of all possible paths $\boldsymbol{x}$ that serve w.
c_h	cost (e.g., transmission cost) for using hyperarc h.
α	probability of not observing a given transmission $h \in \boldsymbol{x}$.

1) Observation: As the adversary has *global observability*, he is potentially able to observe all node transmissions from the entire network. However, we consider the following two observation models for the adversary:

a. *Lossy Observations:* In practice, the adversary may have lossy observations due to the lossy nature of the wireless channel or some blind spots in his network. Hence, the adversary may observe a subvector $\boldsymbol{y}$ from the actual transmission path $\boldsymbol{x}$ where we assume that the observation distribution $p(\boldsymbol{y}|\boldsymbol{x})$ for observing $\boldsymbol{y}$ given that $\boldsymbol{x}$ was transmitted is known. For simplicity, we let $\alpha \in [0, 0.5]$ be the probability of not observing a given transmission $h \in \boldsymbol{x}$ ("erasure probability") and observation of each transmission is independent. In other words, the probability $p(\boldsymbol{y}|\boldsymbol{x})$ can be computed using a sequence of $\|\boldsymbol{x}\|_0$ independent Bernoulli trials with parameter of success $(1-\alpha)$, i.e., $p(\boldsymbol{y}|\boldsymbol{x}) = (1-\alpha)^{\|\boldsymbol{y}\|_0} \alpha^{(\|\boldsymbol{x}\|_0 - \|\boldsymbol{y}\|_0)}$, where $\|.\|_0$ represent the L0-norm which counts the total number of non-zero elements in a vector.

b. *Lossless Observations:* The lossless observations model is a special case in which the adversary perfectly observes a sequence of transmissions $\boldsymbol{y}$ which coincides with the actual transmission path $\boldsymbol{x}$ (i.e., $\boldsymbol{y} = \boldsymbol{x}$).

2) Adversary's Capabilities: We assume that the adversary is *informed*, in that it has complete knowledge of the network graph $\mathcal{G}$, prior probabilities $p(w)$, observation distribution $p(\boldsymbol{y}|\boldsymbol{x})$, and path distribution $p(\boldsymbol{x}|w)$. The actual node transmissions are lossless and only the adversary's observations may be lossy. We assume that the adversary is *external*, in that it does not have access to the individual nodes in the network, and the contents of the communications, including the packet headers, are protected by encryption and do not leak any information on w. We also assume that the adversary is *passive*, in that it does not manipulate the network traffic by dropping or injecting packets, which can be easily detected. The adversary can identify w from each observed $\boldsymbol{y}$ by enumerating the entire set of possible observations for each source-destination pair.

3) Optimal Detection of Source-Destination Pair w: Suppose that the true source-destination pair is w and the adversary observes $\boldsymbol{y}$. A successful detection occurs when the adversary's estimate of the source-destination pair $\widehat{w}(\boldsymbol{y})$ matches w.

Although there exist heuristic-based techniques[1] to estimate $\widehat{w}$, the optimal approach to maximize the expected detection probability of the adversary is the Bayesian maximum-a-posteriori (MAP) estimator [24]: $\widehat{w}_{\text{MAP}} = \arg\max_{w\in\mathcal{V}^2} p(w|\boldsymbol{y})$, where the posterior probability is computed using: $p(w|\boldsymbol{y}) = \frac{p(w,\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|w)p(w)}{\sum_{w'\in\mathcal{V}^2} p(\boldsymbol{y}|w')p(w')}$. The MAP estimator allows the adversary to exploit the prior knowledge of $p(w)$, observation distribution $p(\boldsymbol{y}|\boldsymbol{x})$ and the path distribution $p(\boldsymbol{x}|w)$ to maximize his expected detection rate. Note that the source's identity is implicitly known if there are lossless (complete) observations since the source is always the first node that transmits. However, the destination's identity may be hidden if there are multiple receivers for the transmission.

For a given observation $\boldsymbol{y}$, the probability of correctly guessing w under the MAP approach is given by $P(W = \widehat{w}_{\text{MAP}}|\boldsymbol{y}) = \max_{w\in\mathcal{V}^2} p(w|\boldsymbol{y})$. Thus, the (expected) detection probability for all observations $\boldsymbol{y} \in \mathcal{Y}$ is given by:

$$\begin{aligned} P_{\text{detect}} &= \sum_{\boldsymbol{y}\in\mathcal{Y}} \max_{w\in\mathcal{V}^2} p(w|\boldsymbol{y})\, p(\boldsymbol{y}) \\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}} \max_{w\in\mathcal{V}^2} p(w,\boldsymbol{y}). \end{aligned} \tag{1}$$

Suppose the observations are lossy. Let $p(\boldsymbol{y}|\boldsymbol{x})$ be the probability of observing $\boldsymbol{y}$ given that $\boldsymbol{x}$ was actually transmitted. From (1), the detection probability of the lossy observations adversary is:

$$\begin{aligned} P^{\text{lossy}}_{\text{detect}} &= \sum_{\boldsymbol{y}\in\mathcal{Y}} \max_{w\in\mathcal{V}^2} \sum_{\boldsymbol{x}\in\mathcal{X}} p(w,\boldsymbol{y},\boldsymbol{x}) \\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}} \max_{w\in\mathcal{V}^2} \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|w)\, p(w). \end{aligned} \tag{2}$$

Suppose that the observations are lossless, i.e., $p(\boldsymbol{y}|\boldsymbol{x}) = 1$ if $\boldsymbol{y} = \boldsymbol{x}$, and $p(\boldsymbol{y}|\boldsymbol{x}) = 0$ otherwise. From (1), the detection probability of the lossless observations adversary is:

$$\begin{aligned} P^{\text{lossless}}_{\text{detect}} &= \sum_{\boldsymbol{x}\in\mathcal{X}} \max_{w\in\mathcal{V}^2} p(w,\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}} \max_{w\in\mathcal{V}^2} p(\boldsymbol{x}|w)\, p(w). \end{aligned} \tag{3}$$

Now that we have quantified the adversary's detection probability P_{detect}, we formulate the optimization problem in the next section to minimize P_{detect} for maximum privacy.

Definition 2 (Detection Probability of Adversary)**.** *The detection probability of the adversary is given by $P_{detect} = \sum_{\boldsymbol{y}\in\mathcal{Y}} \max_{w\in\mathcal{V}^2} p(w,\boldsymbol{y})$ [see (1)]. A lower P_{detect} corresponds to a higher level of privacy and vice versa.*

IV. Optimizing the Privacy-Utility Tradeoff in Wireless Networks

We present the Optimal Privacy Enhancing Routing Algorithm (OPERA) which solves the following problem statement: *compute the optimal path distribution $p(\boldsymbol{x}|w)$ that minimizes privacy leakage given some user-defined privacy budget η.* We first explain our objective — minimizing the adversary's detection probability P_{detect}, followed by the cost of using each path $\boldsymbol{x}$, and finally, the utility and other network constraints in our optimization problem.

A. Privacy Metric for the Paths

Our optimization objective is to *minimize the adversary's detection probability P_{detect} (see Definition 2) for better privacy*, i.e., we minimize the detection probability of the lossy observations adversary given in (2).

B. Cost of Using Privacy-Preserving Paths

For a given source-destination pair w, we define the cost of using a privacy-preserving path $\boldsymbol{x} \in \mathcal{X}^w$ to be the cost difference between the path $\boldsymbol{x}$ and the minimum-cost path serving w, given by $\sum_{h\in\boldsymbol{x}} c_h - \min_{\boldsymbol{x}'\in\mathcal{X}^w} \sum_{h\in\boldsymbol{x}'} c_h$. Next, we define the cost of the privacy-preserving scheme for a given network topology to be given by the *expected amount of additional transmission cost incurred by the network*:

$$\mathbb{E}_{w\in\mathcal{V}^2}\left[\mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}\left[\sum_{h\in\boldsymbol{x}} c_h\right] - \min_{\boldsymbol{x}'\in\mathcal{X}^w} \sum_{h\in\boldsymbol{x}'} c_h\right] \tag{4}$$

C. Optimization Formulation

We first provide a general optimization formulation for the lossy (incomplete observation) adversarial model and examine in Section V the lossless observations adversarial model, which is a special case of this general problem. To correctly specify our problem, our formulation must specify the (i) privacy budget η, (ii) valid probabilities, and (iii) valid routing paths. Consider the following constraints:

(i) Privacy budget for each source node u: The value in (4) should be less than or equal to the budget budget η.

$$\sum_{v\in\mathcal{V}} p(w)\left[\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}|w)\sum_{h\in\boldsymbol{x}} c_h\right] - \min_{\boldsymbol{x}'\in\mathcal{X}^w} \sum_{h\in\boldsymbol{x}'} c_h\right] \le \eta, \quad \forall u \in \mathcal{V}. \tag{5}$$

Recall that $w = (u, v)$ and in (5), we fix the source u while varying the destination v in the outer summation term.

(ii) Sum of probabilities over support and non-negativity of probabilities: The summation of the path distribution $p(\boldsymbol{x}|w)$ over its entire support $\mathcal{X}$ must equal one.

$$\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}|w) = 1, \qquad \forall w \in \mathcal{V}^2. \tag{6}$$

A valid probability has to be non-zero.

$$0 \le p(\boldsymbol{x}|w) \le 1, \qquad \forall \boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2. \tag{7}$$

(iii) Valid transmissions: The source node u, by definition must be the first node to transmit while the destination node v needs to receive the transmission from the sequence of transmissions $\boldsymbol{x}$. In other words, we set the probability of using

[1] For example, given that N nodes have received the transmission, a naive heuristic may assign each node that received the transmission with equal probability ($= \frac{1}{N}$) of being the destination node.

a path $\boldsymbol{x}$ that is not in $\mathcal{X}^w$ (the set of all possible paths that serve w) to zero, i.e., we have

$$p(\boldsymbol{x}|w) = 0, \qquad \forall \boldsymbol{x} \notin \mathcal{X}^w, w \in \mathcal{V}^2. \tag{8}$$

1) General Formulation: Given a network graph $\mathcal{G} = (\mathcal{V}, \mathcal{H})$, transmission cost $\{c_h\}_{h \in \mathcal{H}}$, the prior probabilities $\{p(w)\}_{w \in \mathcal{V}^2}$, the adversary's observation distribution $\{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}$, and the privacy budget η, find the path distribution $\{p(\boldsymbol{x}|w)\}_{\boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2}$ that minimizes the adversary's detection probability P_{detect} in (2) such that the expected cost of the privacy-preserving routes is at most η for each source node u. The solution can be obtained by solving the minimax problem where the objective is to minimize $\sum\limits_{\boldsymbol{y} \in \mathcal{Y}} \max\limits_{w \in \mathcal{V}^2} \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w)\, p(w)$, subject to constraints (5), (6), (7), (8).

2) Linear Program Formulation: We can reformulate the minimax problem in Section IV-C1 as a linear program by introducing a variable $z_{\boldsymbol{y}}$ to match the value of $\max\limits_{w \in \mathcal{V}^2} \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w)$ in the objective function at the optimal solution for each $\boldsymbol{y} \in \mathcal{Y}$, along with the inequality constraint: $z_{\boldsymbol{y}} - \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w) \geq 0$, $\forall \boldsymbol{y} \in \mathcal{Y}, w \in \mathcal{V}^2$. At the optimal solution, where $\sum\limits_{\boldsymbol{y} \in \mathcal{Y}} z_{\boldsymbol{y}}$ is minimized, we have $z_{\boldsymbol{y}} = \max\limits_{w \in \mathcal{V}^2} \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w)$ for each $\boldsymbol{y} \in \mathcal{Y}$. The detection probability of the adversary can then be expressed as: $P_{\text{detect}} = \sum\limits_{\boldsymbol{y} \in \mathcal{Y}} z_{\boldsymbol{y}}$. Using the newly introduced $\{z_{\boldsymbol{y}}\}_{\boldsymbol{y} \in \mathcal{Y}}$ variables, we arrive at the linear program formulated in Problem (9).

<u>LPProb($\mathcal{G}, \{c_h\}_{h \in \mathcal{H}}, \{p(w)\}_{w \in \mathcal{V}^2}, \{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}, \eta$):</u>

$$\begin{aligned}
&\underset{\substack{\{p(\boldsymbol{x}|w)\}_{\boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2},\\ \{z_{\boldsymbol{y}}\}_{\boldsymbol{y} \in \mathcal{Y}}}}{\text{minimize}} && \sum_{\boldsymbol{y} \in \mathcal{Y}} z_{\boldsymbol{y}} \\
&\text{subject to} && \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}|w) = 1, \quad \forall w \in \mathcal{V}^2 \\
& && 0 \leq p(\boldsymbol{x}|w) \leq 1, \quad \forall \boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2 \\
& && p(\boldsymbol{x}|w) = 0, \qquad \forall \boldsymbol{x} \notin \mathcal{X}^w, w \in \mathcal{V}^2 \\
&z_{\boldsymbol{y}} - \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w) \geq 0, && \forall \boldsymbol{y} \in \mathcal{Y}, w \in \mathcal{V}^2 \\
&\sum_{v \in \mathcal{V}} p(w) \left[\Big[\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}|w) \sum_{h \in \boldsymbol{x}} c_h \Big] - \min_{\boldsymbol{x}' \in \mathcal{X}^w} \sum_{h \in \boldsymbol{x}'} c_h \right] \leq \eta, \\
& && \forall u \in \mathcal{V}.
\end{aligned} \tag{9}$$

Note that Problem (9) is a linear program because its objective function is simply the summation of the decision variables $z_{\boldsymbol{y}}$, which is a linear function, and all the constraints are also linear. Algorithm 1 summarizes the proposed Optimal Privacy Enhancing Routing Algorithm (OPERA) which probabilistically selects a path $\boldsymbol{x}$ that serves the source-destination pair w according to an optimized path distribution $p(\boldsymbol{x}|w)$.

3) Computational Complexity: The linear program formulation enables our problem to be solved in polynomial time. However, the search space of the problem grows exponentially according to the network size. For each path $\boldsymbol{x}$, the lossy observations adversary can observe $\binom{\|\boldsymbol{x}\|_0}{k}$ possible observations with k node transmissions (see example in [17, Appendix-B]) where $\|\boldsymbol{x}\|_0$ is the number of nodes that transmitted in $\boldsymbol{x}$. Given that the adversary may observe $k = 0, \ldots, \|\boldsymbol{x}\|_0$ number of node transmissions for a path $\boldsymbol{x}$, there are a total of $2^{\|\boldsymbol{x}\|_0}$ possible observations $\boldsymbol{y}$. The number of possible observations grows exponentially with the dimension of $\boldsymbol{x}$, resulting in a combinatorial explosion. Hence, we propose an approximation method for the adversary's lossy observation distribution in the next Section IV-C4.

Algorithm 1: Compute a privacy-preserving path $\boldsymbol{x}$ for a source-destination pair w.

```
OPERA($\mathcal{G}, \{c_h\}_{h \in \mathcal{H}}, \{p(w)\}_{w \in \mathcal{V}^2}, \{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}, \eta, w$):
  Input : Network graph $\mathcal{G}$, transmission cost $\{c_h\}_{h \in \mathcal{H}}$, the
          prior probabilities $\{p(w)\}_{w \in \mathcal{V}^2}$, the adversary's
          observation distribution $\{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}$, privacy
          budget $\eta$, and source-destination $w = (u, v)$.
  Output: Privacy-preserving path $\boldsymbol{x}$.
  Solve the optimization problem,
  LPProb($\mathcal{G}, \{c_h\}_{h \in \mathcal{H}}, \{p(w)\}_{w \in \mathcal{V}^2}, \{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}, \eta$) in (9)
  to obtain the optimized path distribution $p(\boldsymbol{x}|w)$ for using path
  $\boldsymbol{x}$ given $w$ subjected to the budget constraint $\eta$.
  Randomly select a routing path $\boldsymbol{x}$ according to the path
  distribution $p(\boldsymbol{x}|w)$.
```

4) Approximation for the Lossy Observations Adversary: We suggest approximating the adversary's observation model by replacing the observation distribution $\{p(\boldsymbol{y}|\boldsymbol{x})\}_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}}$ values for observations with more than n transmission losses from a path $\boldsymbol{x}$ with zero. More formally, for each $\boldsymbol{x} \in \mathcal{X}$, we let $p(\boldsymbol{y}|\boldsymbol{x}) = 0$ if $p(\boldsymbol{y}|\boldsymbol{x}) < \epsilon$ where $\epsilon = (1-\alpha)^{\|\boldsymbol{x}\|_0 - n} \alpha^n$, with $n \in (0, \|\boldsymbol{x}\|_0)$, and $\alpha \in [0, 0.5]$ is the probability of not observing a given transmission $h \in \boldsymbol{x}$. A smaller parameter ϵ gives a better approximation of P_{detect} but offers less computational savings. Next, we have the following Proposition 1.

Proposition 1. *The approximation method in Section IV-C4, which uses a truncated observation distribution provides a lower bound for P_{detect} obtained in Problem* (9).

Proof. We show that the feasible region in Problem (9) becomes larger in the approximation method, which leads to a lower P_{detect} value. Let the truncated observation probability be $q(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x})$ if $p(\boldsymbol{y}|\boldsymbol{x}) \geq \epsilon$, and $q(\boldsymbol{y}|\boldsymbol{x}) = 0$ otherwise. Consider the $z_{\boldsymbol{y}}$ constraint in Problem (9), which can be rewritten as $z_{\boldsymbol{y}} \geq \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w)$, $\forall \boldsymbol{y} \in \mathcal{Y}, w \in \mathcal{V}^2$. The set of possible $z_{\boldsymbol{y}}$ that satisfies the constraint $z_{\boldsymbol{y}} \geq \sum\limits_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w)$ is a subset of the set of possible $z_{\boldsymbol{y}}$ that satisfies the constraint $z_{\boldsymbol{y}} \geq \sum\limits_{\boldsymbol{x} \in \mathcal{X}} q(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|w) p(w)$ since $q(\boldsymbol{y}|\boldsymbol{x}) \leq p(\boldsymbol{y}|\boldsymbol{x})$. Hence, we obtain a larger feasible region when we use the truncated probability $q(\boldsymbol{y}|\boldsymbol{x})$. This may lead to a lower objective function value in the minimization problem which serves as a lower bound for P_{detect}. □

V. Lossless Adversarial Observability (Worst-Case Scenario)

In this section, we consider the lossless observations adversarial model, which is a special case of the lossy observations adversarial model. The lossless observations adversary perfectly observes each transmission path $\boldsymbol{x}$ and hence represents a worst-case adversary. The probability of observing $\boldsymbol{y}$ given that $\boldsymbol{x}$ was actually transmitted, $p(\boldsymbol{y}|\boldsymbol{x}) = 1$ if $\boldsymbol{y} = \boldsymbol{x}$, and $p(\boldsymbol{y}|\boldsymbol{x}) = 0$ otherwise. Considering this, the objective function in the general problem in Section IV-C1 can simply be replaced by (3), i.e., $\sum_{\boldsymbol{x}\in\mathcal{X}} \max_{w\in\mathcal{V}^2} p(w, \boldsymbol{x})$. Similar to Section IV-C2, we introduce a variable $z_{\boldsymbol{x}}$ to match the value of $\max_{w\in\mathcal{V}^2} p(w, \boldsymbol{x})$, at the optimal solution, along with the inequality constraint: $z_{\boldsymbol{x}} - p(\boldsymbol{x}|w)p(w) \geq 0,\ \forall \boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2$.

Our optimization problem for the lossless observations adversary can be formulated as the linear program in Problem (10). In addition, the problem can be decomposed into smaller subproblems for each source node u to solve in a distributed fashion (see Proposition 2).

<u>DLPProb($\mathcal{G}, \{c_h\}_{h\in\mathcal{H}}, \{p(w)\}_{w\in\mathcal{V}^2}, \eta$):</u>

$$\begin{aligned}
&\underset{\substack{\{p(\boldsymbol{x}|w)\}_{\boldsymbol{x}\in\mathcal{X}, w\in\mathcal{V}^2},\\ \{z_{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathcal{X}}}}{\text{minimize}} \quad && \sum_{\boldsymbol{x}\in\mathcal{X}} z_{\boldsymbol{x}} \\
&\text{subject to} && \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}|w) = 1, \quad \forall w \in \mathcal{V}^2 \\
& && 0 \leq p(\boldsymbol{x}|w) \leq 1, \quad \forall \boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2 \\
& && p(\boldsymbol{x}|w) = 0, \quad \forall \boldsymbol{x} \notin \mathcal{X}^w, w \in \mathcal{V}^2 \\
& && z_{\boldsymbol{x}} - p(\boldsymbol{x}|w)p(w) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}, w \in \mathcal{V}^2 \\
& \sum_{v\in\mathcal{V}} p(w) \left[\Big[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}|w) \sum_{h\in\boldsymbol{x}} c_h \Big] - \min_{\boldsymbol{x}'\in\mathcal{X}^w} \sum_{h\in\boldsymbol{x}'} c_h \right] \leq \eta, \\
& && \forall u \in \mathcal{V}.
\end{aligned} \tag{10}$$

Proposition 2. *The DLPProb problem in* (10) *is block separable.*

Proof. In a lossless observation, the observed source node u must be the first node to transmit. Hence, two routing paths $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ made by two different sources u_1 and u_2 cannot be observed to be the same observation, i.e., $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. In other words, given a $w = (u, v)$ pair, the term $p(\boldsymbol{y}|w = (u', v)) = 0$ for all $u' \in \mathcal{V}, u' \neq u$. Let $\mathbf{a}$ and $\mathbf{b}$ represent the column vector of decision variables $p(\boldsymbol{x}|w)$ and $z_{\boldsymbol{x}}$ respectively. Since the objective function $\sum_{\boldsymbol{x}\in\mathcal{X}} z_{\boldsymbol{x}}$ is a function of only $\mathbf{b}$, it can be partitioned into $|\mathcal{V}|$ summations of the subvectors $\mathbf{b}_1, \mathbf{b}_2, \ldots$ which corresponds to paths made by source node $u_1, u_2, \ldots$. Similarly, it can be easily shown that the optimization constraints can be partitioned to only include variables from the subvector $\mathbf{b}_i$ that correspond to a source node u_i. Therefore, we conclude that Problem (10) is block separable. □

VI. Simulation Results

In this section, we study the adversary's detection probability P_{detect} under the proposed OPERA vs. other privacy-preserving routing schemes based on the Greedy and Uniform heuristics, a baseline heuristic scheme [14], and the minimization of mutual information. We varied the privacy budget η and compared the P_{detect} values against the expected cost incurred by the schemes in various connected network topologies. We evaluated the schemes using the basic line, binary tree, and grid network topologies (see [17, Fig. 3]), the random topology, in addition to two other real-world topologies from the Roofnet [25] and Indriya [26] testbeds.

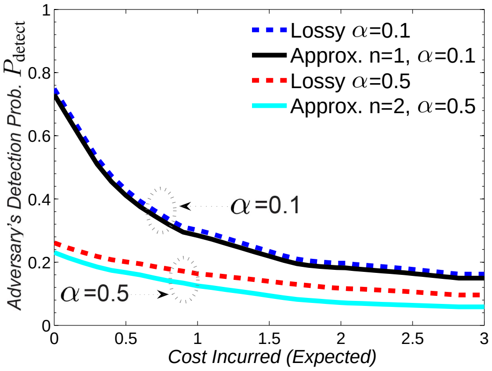


Fig. 3. Adversary's detection probability P_{detect} for the lossy observations model in a 10-node line network with different α and n parameters. Recall that α is the probability of not observing a given transmission $h \in \boldsymbol{x}$ while n is the parameter in our approximation method in Section IV-C4.

We assume (except for Section VI-F) that the links are symmetric, i.e., for each hyperarc $h = (i, \mathcal{R})$ in the network, there exists $|\mathcal{R}|$ hyperarcs given by $h_k = (k, \mathcal{R}_k), k \in \mathcal{R}$, where $i \in \mathcal{R}_k$. We let the cost of each hyperarc $h \in \mathcal{H}$ be one, i.e., $c_h = 1$. We assume single-path routing in Sections VI-A to VI-D and multipath routing in Sections VI-E to VI-F. Finally, we assume that $w = (u, v)$ is chosen uniformly at random from the set of all possible node pairs where $u \neq v$.

Performance Metric: For a fixed cost incurred in expectation, we consider the protocol that achieves higher privacy to be the superior one. Recall that a higher privacy corresponds to a lower P_{detect}.

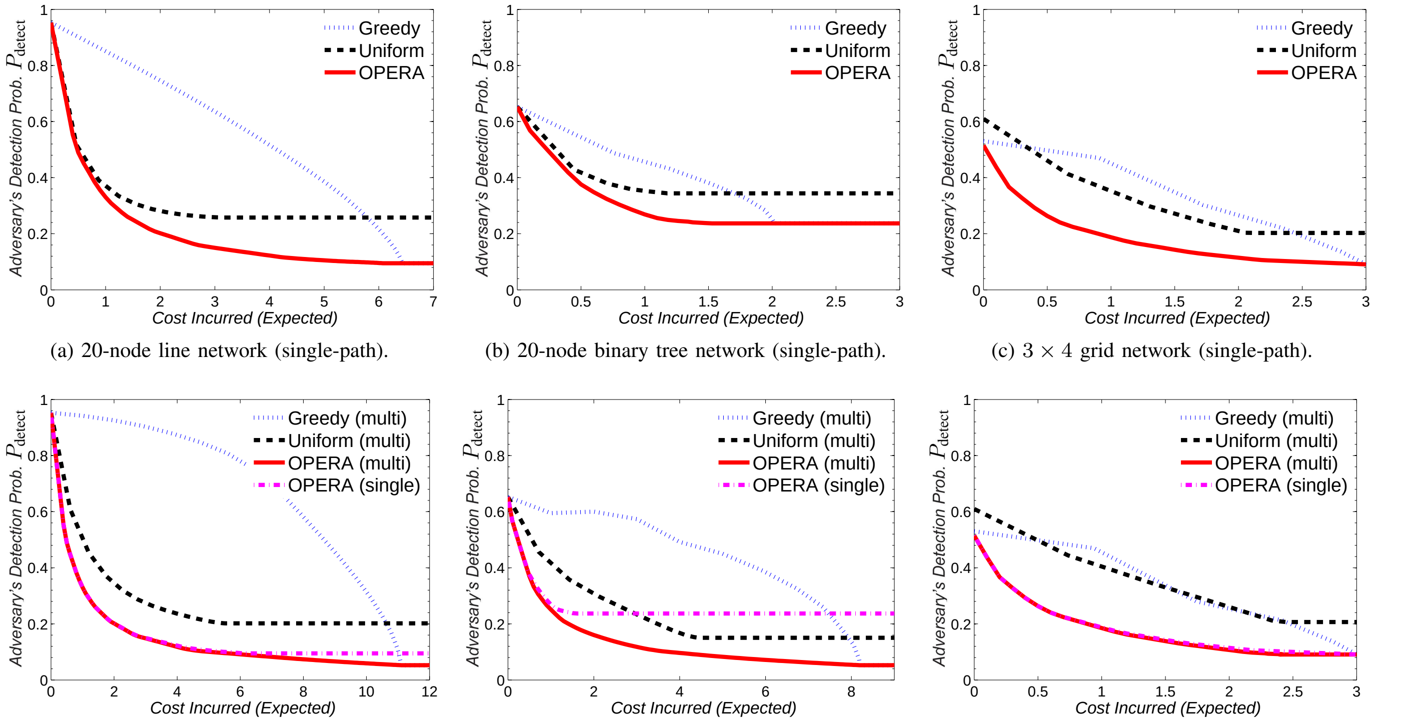
We now discuss our findings under various network settings.

A. Lossy Adversarial Observations

We solve Problem (9) to obtain the optimal P_{detect} values for the proposed OPERA. Fig. 3 shows the P_{detect} values for OPERA and the approximation method in Section IV-C4 for a line network with the erasure probabilities $\alpha = 0.1$ and $\alpha = 0.5$. Generally, P_{detect} decreases as α increases since the unobserved transmissions may belong to a larger set of possible source nodes. Also, a larger n value is needed to better approximate P_{detect} for larger α values. There exists an inverse relationship between the value of n and the complexity of the optimization problem in (9), and a higher n results in a more accurate estimate of the true P_{detect} at the expense of additional computational costs. More performance degradation is experienced in the grid network compared to the line network as the number of possible w pairs increases when less transmissions are observed. Interestingly, the optimized paths from the approximation method coincide with the optimal paths of the original method in our simulation, i.e., the adversary's detection rate does not improve even if he uses (2) while the system uses the approximation method.

B. Comparison with Greedy and Uniform Heuristics

From this section onwards, we assume a lossless observations adversary. The details for the *Greedy* and *Uniform*



(a) 20-node line network (single-path).

(b) 20-node binary tree network (single-path).

(c) 3 × 4 grid network (single-path).

(d) 20-node line network (multipath).

(e) 20-node binary tree network (multipath).

(f) 3 × 4 grid network (multipath).

Fig. 4. Adversary's detection probability P_{detect} under the Greedy and Uniform heuristics and the proposed OPERA, for single vs. multipath routing (rows), and line, binary tree, and grid networks (columns).

heuristics are given in [17, Algorithm 2] and [17, Algorithm 3] respectively. In the Uniform heuristic, we select a path uniformly at random from all valid paths that serve w (similar[2] to [12]) subjected to the privacy budget. In the Greedy heuristic, we always greedily send the packets via the path containing the most number of receivers, subjected to the privacy budget. Similar to OPERA, the two heuristics exploit knowledge of the network graph $\mathcal{G}$ to provide better privacy. As such, they provide an upper bound on the achievable privacy for other heuristics that use only local network topology information. However, the privacy budget constraint applies to each path $\boldsymbol{x}$ instead of the expected privacy budget for each source node as used in OPERA.

Figs. 4a, 4b, and 4c show the P_{detect} values of the two heuristics and the proposed OPERA under different network topologies with the single-path constraint. For most values of the incurred cost, there exists a significant difference (up to 50%) in the performance of the two heuristics compared to OPERA. In Figs. 4a and 4b, the performance of the Greedy heuristic is worse than the Uniform heuristic at lower privacy budgets despite greedily choosing the path with the most number of receivers. This indicates that increasing the number of receiver nodes does not necessary translate to better privacy. In fact, the difference between the Greedy heuristic and OPERA can be quite significant as shown in the figure. The Uniform heuristic does not converge to the maximum achievable privacy even when the privacy budget is slack, unlike the Greedy heuristic. In Fig. 4c, which uses a grid topology, the Uniform heuristic will uniformly pick each valid shortest path that serve w (which leaks information about the destination) while the Greedy and OPERA methods tend to choose a single path. Hence, this results in a higher P_{detect} for the Uniform heuristic even when the expected cost is zero.

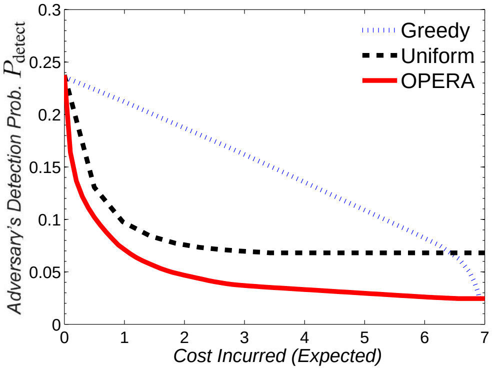


Fig. 5. Adversary's detection probability P_{detect} under the Greedy and Uniform heuristics and the proposed OPERA, averaged over five randomly generated 80-node network topologies (with single-path routing).

Lastly, Fig. 5 shows the P_{detect} values of the two heuristics and the proposed OPERA (with the single-path constraint), averaged over five randomly generated 80-node networks. The P_{detect} values have a similar trend to the results from the smaller 20-node line network in Fig. 4a where OPERA outperforms the Uniform and Greedy heuristics.

[2] The authors in [12] proposed a dummy packet injection scheme that randomly (uniformly) transmits a dummy packet to a chosen receiver located m hops away from the destination where $m > 1$. However, the scheme was designed for an adversary with local observability. Hence, we used a uniform heuristic that follows the authors' main idea of making the transmission paths "completely random instead of a directed one".

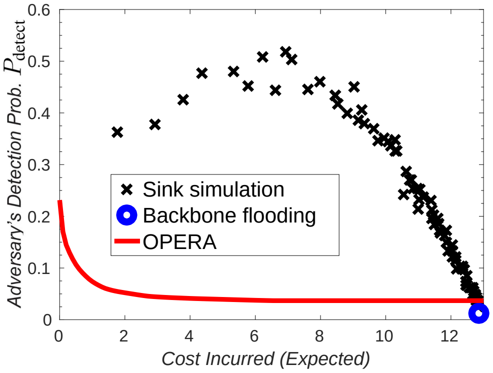


Fig. 6. Adversary's detection probability P_{detect} under the sink simulation (we varied the value of the L parameter from 2-79 and computed the corresponding P_{detect} values for the cost incurred) and backbone flooding schemes proposed in [14] and the proposed OPERA (with single-path routing), averaged over five randomly generated 80-node network topologies.

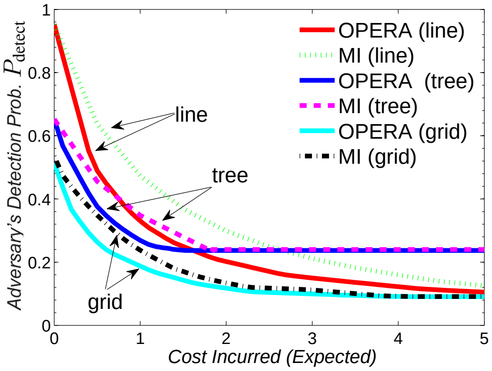


Fig. 7. Adversary's detection probability P_{detect} under the minimizing mutual information (MI) optimization approach and the proposed OPERA for the line, binary tree, and grid networks (with single-path routing).

C. Comparison with the Sink Simulation and Backbone Flooding Schemes

We compared our proposed OPERA against an existing protocol proposed by Mehta *et al.* [14]. Similar to our work, Mehta *et al.* proposed the sink simulation and backbone flooding schemes in [14, Section 5.2] to provide location privacy for the network sinks under the same global adversary assumption as considered in our work. As the work in [14] considered a wireless sensor network setting where all source nodes transmit to a common sink, we have to modify their proposed sink simulation and backbone flooding schemes to suit our setting. Mainly, we arbitrarily assigned the same L simulated (fake) destination nodes for each destination node in the sink simulation technique and let the source node transmit to all the L simulated (and the true) destination nodes using the shortest path routes. To avoid double counting the transmission costs, we allow all transmissions to be piggybacked into a single transmission if the routes overlap. For the backbone flooding scheme, we do not use the proposed approximation algorithm for constructing the backbone network. Instead, we used the minimum spanning tree to flood a packet to the entire network. The minimum spanning tree minimizes the total transmission cost needed for flooding a packet to the entire network, and hence is an ideal backbone network.

Fig. 6 shows the P_{detect} values of the sink simulation and backbone flooding schemes and the proposed OPERA (with the single-path constraint), averaged over five randomly generated 80-node networks. The performance of the sink simulation technique is significantly worse (up to five times higher P_{detect}) than OPERA for the same amount of cost incurred. This is true even for large L values as the privacy of the source-destination pair is not necessary proportional to the number of receiver nodes (simulated sinks). Although the performance of the backbone flooding scheme is slightly better than OPERA, it is not flexible enough to allow users to specify a privacy budget constraint. Hence, depending on the network application, it can result in excessive costs.

D. Comparison with Mutual Information (MI) Minimization

Fig. 7 shows the P_{detect} values of the MI minimization problem (see [17, Appendix-E]) and the proposed OPERA for the line, binary tree, and grid networks. It is observed that minimizing MI results in a higher P_{detect} value (and hence, less privacy) compared to OPERA when the privacy budget is tight. Interestingly, we observed that different MI values may correspond to the same P_{detect} value when the privacy budget is slack. However, the converse is not true in our simulations. For the same number of nodes, the privacy difference is largest in the line network (up to 15%) and smallest in the grid network (up to 6%). However, minimizing MI is still superior to the Greedy and Uniform heuristics. Therefore, despite being commonly proposed as a measure for privacy [27]–[30], minimizing MI may not be ideal in our case where a MAP adversary was considered.

E. Comparison of Single-path and Multipath Routing

We studied the effects of using multipaths $\mathcal{M} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$ where at least one path $\boldsymbol{x}_i$ will reach the destination node. In the multipath routing, the routing paths $\boldsymbol{x} \in \mathcal{X}$ in Problem (10) are replaced by a set of paths $\mathcal{M} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$. The P_{detect} values for the single and multipath routing in the line, binary tree, and grid networks are given in Figs. 4d, 4e, and 4f respectively.

Generally, for a fixed incurred cost, the multipath variants are able to achieve more privacy compared to single-path at the expense of higher computational cost. The improvement in P_{detect} for the proposed OPERA appears to be mild in the line network and does not have any significant effect in the grid network. However, the improvement is more significant in the binary tree network as the privacy budget becomes slack. This is because the multipath approach can improve privacy in scenarios where a leaf node is communicating with another leaf node in the same subtree. When the route is restricted to only a single path, the destination can be easily linked to the same subtree as the path does not travel to other subtrees. This severely limits the number of receivers and lowers privacy when the privacy budget is slack. In practice, the single-path routing constraint can be used if the privacy budget is tight.

F. Using Network Topologies from Real-World Testbeds

To evaluate the practicality of OPERA in real-world topologies, we used topologies from the outdoor Roofnet [25] and indoor Indriya [26] testbeds, and the corresponding P_{detect} values

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2017.2698424, IEEE Transactions on Information Forensics and Security

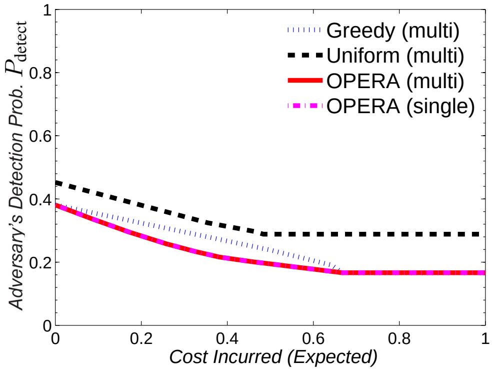


Fig. 8. Adversary's detection probability P_{detect} for the Roofnet network with multipaths under the Greedy and Uniform heuristics and the proposed OPERA.

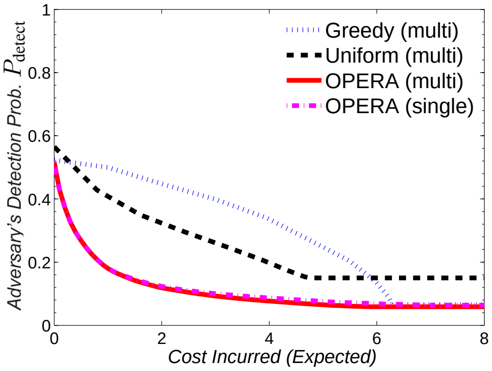


Fig. 9. Adversary's detection probability P_{detect} for the Indriya network with multipaths under the Greedy and Uniform heuristics and the proposed OPERA.

are shown in Figs. 8 and 9 respectively. The Roofnet testbed consisted of nine IEEE 802.11b wireless nodes installed in the apartments of volunteers near Massachusetts Institute of Technology and covers approximately one square kilometer. On the other hand, Indriya [26] is a large-scale indoor wireless sensor network testbed deployed at the National University of Singapore. It consists of 127 TelosB motes and covers 3 floors of a building. For the Roofnet network, we used links with more than 10% delivery rate (includes three non-symmetric links). For the Indriya network, we considered a subset of 18 nodes, arbitrary selected from each room of the network to reduce the computation complexity. The performance of OPERA in the two mesh-like real-world topologies are similar to our earlier results in the grid topology. There exist little differences in P_{detect} between the single-path and multipath routing for the proposed OPERA. Hence, the single-path optimization problem which has lower computational complexity may be used for such real-world networks.

VII. Conclusion

We have developed a statistical decision-making framework to optimally solve the privacy-preserving routing problem in wireless networks given some utility constraints assuming a powerful global adversary that uses the optimal maximum-a-posteriori (MAP) estimation strategy. We also showed via simulations that our approach is significantly better than the Uniform and Greedy heuristics, a baseline scheme, and the mutual information minimization scheme. For future work, it would be interesting to study the privacy-utility trade-off problem for mobile networks and to provide stricter privacy constraints for the communicating parties.

References

[1] J. Deng, R. Han, and S. Mishra, “Countermeasures against traffic analysis attacks in wireless sensor networks,” in *Proc. Int. Conf. Security and Privacy for Emerging Areas in Commun. Networks*, pp. 113–126, 2005.

[2] M. Shao, Y. Yang, S. Zhu, and G. Cao, “Towards statistically strong source anonymity for sensor networks,” in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2008.

[3] J. Y. Koh, J. Teo, D. Leong, and W.-C. Wong, “Reliable privacy-preserving communications for wireless ad hoc networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 6271–6276, Jun. 2015.

[4] P. Zhang, C. Lin, Y. Jiang, P. Lee, and J. Lui, “ANOC: Anonymous network-coding-based communication with efficient cooperation,” *IEEE J. Sel. Areas Commun.*, vol. 30, pp. 1738–1745, Oct. 2012.

[5] H. Shen and L. Zhao, “ALERT: An anonymous location-based efficient routing protocol in MANETs,” *IEEE Trans. Mobile Comput.*, vol. 12, pp. 1079–1093, Jun. 2013.

[6] M. Conti, J. Willemsen, and B. Crispo, “Providing source location privacy in wireless sensor networks: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 15, pp. 1238–1280, Jan. 2013.

[7] Z. Ling, J. Luo, W. Yu, X. Fu, D. Xuan, and W. Jia, “A new cell-counting-based attack against tor,” *IEEE/ACM Trans. Networking*, vol. 20, pp. 1245–1261, Aug 2012.

[8] D. L. Chaum, “Untraceable electronic mail, return addresses, and digital pseudonyms,” *Commun. ACM*, vol. 24, pp. 84–90, Feb. 1981.

[9] M. Reed, P. Syverson, and D. Goldschlag, “Anonymous connections and onion routing,” *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 482–494, May 1998.

[10] S. Mathur and W. Trappe, “BIT-TRAPS: building information-theoretic traffic privacy into packet streams,” *IEEE Trans. Inf. Forens. Security*, vol. 6, pp. 752–762, Sep. 2011.

[11] C. Ozturk, Y. Zhang, and W. Trappe, “Source-location privacy in energy-constrained sensor network routing,” in *Proc. ACM Workshop on Security of Ad Hoc and Sensor Networks*, pp. 88–93, 2004.

[12] Y. Jian, S. Chen, Z. Zhang, and L. Zhang, “A novel scheme for protecting receiver's location privacy in wireless sensor networks,” *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3769–3779, Oct. 2008.

[13] U. Acharya and M. Younis, “Increasing base-station anonymity in wireless sensor networks,” *Elsevier Ad Hoc Networks*, vol. 8, pp. 791–809, Nov. 2010.

[14] K. Mehta, D. Liu, and M. Wright, “Protecting location privacy in sensor networks against a global eavesdropper,” *IEEE Trans. Mobile Comput.*, vol. 11, pp. 320–336, Feb. 2012.

[15] A. Diyanat, A. Khonsari, and S. P. Shariatpanahi, “A dummy-based approach for preserving source rate privacy,” *IEEE Trans. Inf. Forens. Security*, vol. 11, pp. 1321–1332, Jun. 2016.

[16] A. Pfitzmann and M. Hansen, “A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management.” (v0.34). tech. rep., TU Dresden and ULD Kiel, Aug. 2010.

[17] J. Y. Koh, D. Leong, G. W. Peters, I. Nevat, and W.-C. Wong, *Optimal Privacy-Preserving Probabilistic Routing for Wireless Networks*, 2017, [Online]. Available: http://www.purl.org/privacypreservingrouting, accessed on Apr. 18, 2017.

[18] J. Yao and G. Wen, “Preserving source-location privacy in energy-constrained wireless sensor networks,” in *Proc. Int. Conf. Distrib. Comput. Syst. Workshops*, pp. 412–416, Jun. 2008.

[19] Y. Li, J. Ren, and J. Wu, “Quantitative measurement and design of source-location privacy schemes for wireless sensor networks,” *IEEE Trans. Parallel and Distr. Syst.*, vol. 23, pp. 1302–1311, Jul. 2012.

[20] R. Rios, J. Cuellar, and J. Lopez, “Probabilistic receiver-location privacy protection in wireless sensor networks,” *Elsevier Inf. Sciences*, vol. 321, pp. 205–223, Nov. 2015.

[21] M. M. E. A. Mahmoud and X. Shen, “A cloud-based scheme for protecting source-location privacy against hotspot-locating attack in wireless sensor networks,” *IEEE Trans. Parallel and Distr. Syst.*, vol. 23, pp. 1805–1818, Oct 2012.

[22] M. M. E. A. Mahmoud, S. Taha, J. Misic, and X. Shen, “Lightweight privacy-preserving and secure communication protocol for hybrid ad

hoc wireless networks," *IEEE Trans. Parallel and Distr. Syst.*, vol. 25, pp. 2077–2090, Aug 2014.

[23] Y. Guan, X. Fu, R. Bettati, and W. Zhao, "An optimal strategy for anonymous communication protocols," in *Proc. Int. Conf. Distrib. Comput. Syst.*, pp. 257–266, 2002.

[24] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[25] B. A. Chambers, "The grid roofnet: a rooftop ad hoc wireless network," in *Master's thesis, Massachusetts Institute of Technology*, May 2002.

[26] M. Doddavenkatappa, M. Chan, and A. Ananda, "Indriya: A low-cost, 3D wireless sensor network testbed," in *Proc. Int. Conf. Testbeds and Research Infrastructures for the Development of Networks & Communities*, 2011.

[27] S. Zhioua, "A geometric view of mutual information: Application to anonymity protocols," in *Proc. Inf. Theory and its Applications (ISITA)*, pp. 60–65, Oct. 2010.

[28] L. Sankar, S. Rajagopalan, and H. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forens. Security*, vol. 8, pp. 838–852, Jun. 2013.

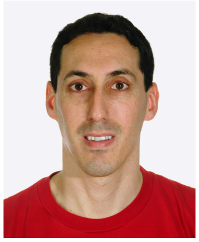
[29] A. R. Coble, "Anonymity, information, and machine-assisted proof," in *Technical Report UCAM-CL-TR-785*, 2010.

[30] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *J. ACM*, vol. 61, pp. 1–57, Dec. 2014.

Jing Yang Koh received the B.Eng. degree in computer engineering from Nanyang Technological University, Singapore in 2013. He is currently a Ph.D. candidate in the National University of Singapore, Singapore and is attached to the Institute for Infocomm Research (I^{2}R), A*STAR, Singapore. His research mainly focuses on using statistical signal processing, machine learning, and optimization techniques for wireless network security and privacy.

Derek Leong received the B.S. degree in electrical and computer engineering from Carnegie Mellon University in 2005, and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology in 2008 and 2013, respectively. He is a scientist with the Smart Energy and Environment cluster at the Institute for Infocomm Research (I^{2}R), A*STAR, Singapore. His research and development interests include distributed systems, sensor networks, smart cities, and the Internet of Things.

Gareth W. Peters is an Assistant Professor in the Department of Statistical Science in University College London. He is a Principle Investigator in CSML , University College London (UCL) and an Academic Member of the UK PhD Center in Financial Computing (UCL). He has published in excess of 100 peer reviewed articles on risk and insurance modelling, 2 research text books on Operational Risk and Insurance as well as being the editor and contributor to 3 edited text books on spatial statistics and Monte Carlo methods. He holds positions as an Adjunct Scientist in the Mathematics, Informatics and Statistics, Commonwealth Scientific and Industrial Research Organisation (CSIRO) since 2009 as well being an Associate Member Oxford-Man Institute in Oxford University since 2012, an Associate Member Systemic Risk Center in London School of Economics since 2014; an Affiliated Prof. School of Earth and Space Sciences, Peking University PKU, Beijing, China since 2015 and a Visiting Prof. in the Institute of Statistical Mathematics, Tokyo, Japan each year since 2010.



Ido Nevat received the B.Sc. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel in 1998 and the Ph.D. degree in Electrical Engineering from the University of NSW, Sydney, Australia, in 2010. Between 2010-2013 he was a postdoctoral research fellow with the Wireless and Networking Technologies Laboratory at CSIRO, Australia. Between 2013-2016 he was a scientist at the Institute for Infocomm Research (I^{2}R), Singapore. Since 2017 Ido has been a team leader at TUMCREATE in Singapore. His main areas of interests include statistical signal processing, machine learning and Bayesian statistics.

Wai-Choong (Lawrence) Wong is Professor in the Department of Electrical and Computer Engineering, National University of Singapore (NUS). Since joining NUS in 1983, he served in various leadership positions at the department, faculty and university levels. Prior to joining NUS in 1983, he was a Member of Technical Staff at AT&T Bell Laboratories, Crawford Hill Lab, NJ, USA, from 1980 to 1983. He received the B.Sc. and Ph.D. degrees in Electronic and Electrical Engineering from Loughborough University, UK. His research interests include wireless and sensor networks and systems, ambient intelligent platforms, localization and source matched transmission techniques with over 280 publications and 5 patents in these areas. He has received several awards including the IEEE Marconi Premium Award in 1989, IEEE Millennium Award in 2000, e-nnovator Awards (Open Category) in 2000, Best Paper Award at the IEEE International Conference on Multimedia and Expo (ICME) 2006, and Best Paper Award at the 2nd International Conference on Ambient Computing, Applications, Services and Technology (AMBIENT), 2012.