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 enhance privacy. Usually, the addition of artificial traffic is done heuristically with no guarantees that the transmission cost, latency, etc., are optimized in every network topology. Hence, 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 both lossless and lossy global adversaries who can observe node transmissions from the entire network and 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 when the adversary uses the MAP estimation strategy.

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

I. INTRODUCTION

Traffic analysis attacks [1]–[8] 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 [5], [9], [10], (ii) introduce transmission delays at each hop [2], [3], [11] to decorrelate traffic flows, or (iii) introduce dummy traffic [4], [6], [12]–[16] 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.

We focus on hiding the source-destination identities (or unlinkability [5], [17], [18]) of each communication where a global adversary is able to observe node transmissions from the entire network (see Fig. 1). Protecting the privacy of the source-destination identities is important as communicating parties may not want to be linked together by a third party. Hence, our challenge is to decide how to probabilistically route the packets from the source to the destination nodes via carefully chosen phantom (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 phantom receiver nodes (nodes that received the dummy traffic).

The statistical decision-making framework approach extends our earlier work in [4] 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 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.

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.
Our work differs from existing works addressing anonymous routing in wireless networks by rigorously quantifying both the privacy and the overheads incurred, and proposes an optimization formulation for the privacy-utility trade-off. In addition, our linear program can be solved in a distributed fashion by each node in the network under the lossy adversary assumption. A toy example is provided in Appendix-A to illustrate the performance improvement of our optimization approach compared to a Uniform heuristic.

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 maximum-a-posteriori (MAP) estimation strategy. This work significantly extends our earlier work in [4] which did not rigorously study the adversary’s optimal estimation strategy using the statistical decision-making framework. 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.

Organization: The outline of this paper is as follows. First, we review the related work in Section II. Next, we introduce the system model in Section III, and present the problem formulations for the lossy and lossless adversaries in Sections IV and V respectively. Finally, Section VI discusses our experimental results and the conclusions are drawn in Section VII.

II. RELATED WORK

Anonymity enhancing techniques like onion routing [10] and mix-net [9] 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 (e.g., Tor [19]) 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 attacks from a global adversary. In addition, due to the wireless broadcast medium, it may also be 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. [12]. 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 [20]–[22] 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 [7]. Interestingly, the work in [3] used a periodic flooding approach for privacy protection with statistical guarantees.

Subsequently, Deng et al. [1] studied the receiver-location privacy problem in wireless networks and proposed several anonymity enhancing techniques to protect the location of the sink (which is the receiver) from basic traffic analysis attacks. A packet is randomly routed from the source node to the sink and each forwarder node may randomly inject dummy packets to create regions with high communication activity. This makes it harder for an adversary to identify the location of the sink which is assumed to be in an area with high communication activity. Jian et al. [13] devised a protocol to protect the receiver’s location privacy from packet-tracing attacks. They proposed using path diversity (randomized routing and dummy packet injections) to make the incoming and outgoing traffic at each node uniformly distributed.

A stronger global adversary which can observe transmissions in the entire network was considered by [15]. 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 [23], 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 [14], 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 [4]. Lastly, the work in [24] proposed a cloud-based scheme for enhancing the source node privacy and [25] 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 weight it against the additional amount of privacy provided. The work in [26] 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.

The optimization approach allows one to compute the optimal routing paths that minimize (or maximize) an objective function given some constraints. Different from [26], we formulate a statistical decision-making framework and use a more appropriate (non-information-theoretic) privacy metric for our objective function. Our chosen privacy metric directly reflects the adversary’s success probability under his optimal maximum-a-posteriori (MAP) estimation strategy. While the information-theoretic approach can quantify the amount of privacy, they do not directly reveal the detection rates of information.

With this objective, we formulate a statistical decision-making framework and use a probabilistic privacy-preserving routing protocol to minimize the probability of an adversary correctly guessing the source-destination identities. In addition, the privacy-preserving routing scheme (see Fig. 2) should consider the adversary’s observation model $p(y|x)$ while computing a (routing) path distribution $p(x|w)$ that serves the source-destination pair $w$. Intuitively, there should be a many-to-one mapping from $w$ to $y$. And there exists a $h = (s, R) \in x$ such that $v \in R$. Let $\mathcal{Y}$ represent the set of all possible observations $y$.

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 $G$, a probabilistic source-routing protocol selects a path $x \in \mathcal{X}$ according to a path distribution $p(x|w)$ for a given source-destination pair $w \in \mathcal{Y}$.

Motivation for 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 privacy-preserving routing scheme (see Fig. 2) should consider the adversary’s observation model $p(y|x)$ while computing a (routing) path distribution $p(x|w)$ that serves the source-destination pair $w$. Intuitively, there should be a many-to-one mapping from $w$ to the observation $y$ to ensure privacy of $w$ since an observation $y$ could be linked to many $w$’s. A one-to-one mapping from $y$ to $w$ on the other hand, does not offer any privacy for $w$ and should be avoided. Although the simple flooding approach can easily achieve maximum privacy, it is usually considered to be too expensive. Therefore, we formulate a statistical decision-making framework that allows us to compute the optimal path distribution $p(x|w)$ that maximizes the privacy of $w$ given a utility (or cost) constraint. We use the terms anonymity and privacy interchangeably throughout the paper. In addition, we use the term phantom receivers to refer to receiver nodes that are not along the shortest-path from the source node to the destination node but receive the packet transmission. Note that all phantom receivers are actual nodes in the network.

A. Adversary Model

We consider an external, passive, global and informed [3] adversary who observes a (possibly-lossy) sequence of transmissions $y$ from an actual transmission path $x$. The adversary aims to detect the identity of the source-destination pair $w$ for each observation $y$. In other words, he aims to identify which node is talking to which node based on his possibly imperfect observations. We assume that the adversary uses the Bayesian traffic analysis technique described in Section III-A3.
The adversary can identify prior probabilities of the communications, including the packet headers are never, we assume that the external adversary has complete knowledge of the network graph. Suppose that the true source-destination pair is \( x \), and the adversary is potentially able to observe all node transmissions from his set of all possible paths \( x \) in the network. However, we consider the following two heuristics to the lossy nature of the wireless cost (e.g., transmission cost) for using hyperarc \( h \). 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 \( y \) from the actual transmission path \( x \) where we assume that the observation distribution \( p(y|x) \) for observing \( y \) given that \( x \) was transmitted is known. For simplicity, we let \( \alpha \in [0, 0.5] \) be the probability of not observing a given transmission \( h \in x \) (“erasure probability”) and observation of each transmission is independent. In other words, the probability \( p(y|x) \) can be computed using a sequence of \( \|x\|_0 \) independent Bernoulli trials with parameter of success \( (1-\alpha) \), i.e., \( p(y|x) = (1-\alpha)^{\|x\|_0}(1-\alpha)^{1-\|x\|_0} \), where \( \|\cdot\|_0 \) represent the L0-norm which counts the total number of non-zero elements in a vector.

b. **Lossless Observations:** Under the lossless observations assumption, the adversary perfectly observes a sequence of transmissions \( y \) which coincides with the actual transmission path \( x \) (i.e., \( y = x \)).

2) **Adversary’s Capabilities:** We assume that the informed adversary has complete knowledge of the network graph \( G \), prior probabilities \( p(w) \), observation distribution \( p(y|x) \), and path distribution \( p(x|w) \). The actual node transmissions are lossless and only the adversary’s observations may be lossy. The adversary can identify \( w \) from each observed \( y \) by enumerating the entire set of possible observations for each source-destination pair (see example in Appendix-A2). However, we assume that the external adversary 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 and does not manipulate the network traffic by dropping or injecting packets, which can be easily detected.

3) **Optimal Detection of Source-Destination Pair \( w \):** Suppose that the true source-destination pair is \( w \) and the adversary observes \( y \), a successful detection occurs when the adversary’s estimate of the source-destination pair \( \hat{w}(y) \) matches \( w \).

Although, there exists heuristic-based techniques\(^1\) to estimate \( \hat{w} \), the optimal approach to maximize the expected detection probability of the adversary (see example in Appendix-A2) is the Bayesian maximum-a-posteriori (MAP) estimator [27]:

\[
\hat{w}_{MAP} = \arg \max_{w \in \mathbb{V}^2} \max_{y \in \mathbb{Y}} p(w|y),
\]

where the posterior probability is computed using:

\[
p(w|y) = \frac{p(y|w)p(w)}{\sum_{w \in \mathbb{V}^2} p(y|w)p(w)}.
\]

The MAP estimator in (1) allows the adversary to exploit the prior knowledge of \( p(w) \), observation distribution \( p(y|x) \) and the path distribution \( p(x|w) \) 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 transmitted. However, the destination’s identity may still be hidden if there are multiple receivers for each transmission.

For a given observation \( y \), the probability of correctly guessing \( w \) under the MAP approach is given by \( p(W = \hat{w}_{MAP}|y) = \max_{w \in \mathbb{V}^2} p(w|y) \). The (expected) detection probability for all observations \( y \in \mathbb{Y} \) is given by taking the expectation of the conditional detection probabilities:

\[
P_{detect} = \sum_{y \in \mathbb{Y}} p(\text{"detect"}|y)p(y)
\]

\[
= \sum_{y \in \mathbb{Y}} \max_{w \in \mathbb{V}^2} p(w|y)p(y)
\]

\[
= \sum_{y \in \mathbb{Y}} \max_{w \in \mathbb{V}^2} p(w, y).
\]

Suppose the observations are lossy. Let \( p(y|x) \) be the probability of observing \( y \) given that \( x \) was actually transmitted. From (3), the detection probability of the lossy adversary is:

\[
P_{detect}^{lossy} = \sum_{y \in \mathbb{Y}} \max_{w \in \mathbb{V}^2} \sum_{x \in \mathcal{X}} p(w, y, x)
\]

\[
= \sum_{y \in \mathbb{Y}} \max_{w \in \mathbb{V}^2} \sum_{x \in \mathcal{X}} p(y, x|w)p(w)
\]

\[
= \sum_{y \in \mathbb{Y}} \max_{w \in \mathbb{V}^2} \sum_{x \in \mathcal{X}} p(y|x)p(x|w)p(w).
\]

Suppose that the observations are lossless, i.e., the probability \( p(y|x) = 1 \) if \( y = x \), and \( p(y|x) = 0 \) otherwise. From (3), the detection probability of the lossless adversary is:

\[
P_{detect}^{lossless} = \sum_{x \in \mathcal{X}} \max_{w \in \mathbb{V}^2} p(w, x)
\]

\[
= \sum_{x \in \mathcal{X}} \max_{w \in \mathbb{V}^2} p(x|w)p(w).
\]

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.

\(^1\)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.
expressed in terms of \( A \). Privacy Metric for the Paths

We first explain our objective function - minimizing the (OPERA) which solves the following problem statement: compute the optimal path distribution \( p(x|w) \) that minimizes privacy leakage given some user-defined privacy budget \( \eta \).

We first explain our objective function - minimizing the adversary’s detection probability \( P_{\text{detect}} \), followed by the cost of using each path \( x \), 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 \( P_{\text{detect}} \) (see Definition 2) for better privacy. Under our proposed framework, \( P_{\text{detect}} \) can also be expressed in terms of \( p(y|w) \):

\[
P_{\text{detect}} = \sum_{w \in V^2} \sum_{y \in Y} p(y|w)p(w) \max_{w' \in V^2} \frac{p(y|w')p(w')}{p(y)}. \tag{6}
\]

Our chosen privacy metric differs from the works in [28]–[31] which use information-theoretic related metrics, e.g., mutual information to quantify privacy. While the information-theoretic measure is useful for quantifying the loss of anonymity (or information leakage), it may not be optimal under our considered adversarial model. The mutual information (MI) [28]–[34] objective function is given by (see Appendix-E for derivation):

\[
I(W; Y) = \sum_{w \in V^2} \sum_{y \in Y} p(y|w)p(w) \log \frac{p(y|w)}{p(y)}. \tag{7}
\]

Although minimizing both (6) or (7) will decrease \( P_{\text{detect}} \), there are some subtle differences. Minimizing (6) directly minimizes \( P_{\text{detect}} \) by lowering the maximum value of the posterior distribution \( p(w|y) \) while minimizing (7) reduces the amount of information that the adversary has learned from its observations by flattening the posterior distribution \( p(w|y) \), which can potentially reduce \( P_{\text{detect}} \). Note that the optimal solutions for minimizing (6) and (7) may not be unique. For (6), as long as the \( \max p(w|y) \) value remains unchanged, the other \( p(w|y) \) probabilities can take arbitrary values. For (7), as long as the set of \( p(w|y) \) values remains unchanged, their ordering is not important.

A linear program can be formulated to minimize (6) but not for minimizing (7) which is non-linear due to its logarithm term. The linear program is relatively easier to solve than the minimization of (7), which is convex.

B. Cost of Using Privacy-Preserving Paths

For a given source-destination pair \( w \), we define the cost of using a privacy-preserving path \( x \in X^w \) to be the cost difference between the path \( x \) and the minimum-cost path serving \( w \), given by \( \sum_{h \in S} \min_{x' \in X^w} \sum_{h \in S} c_h \). 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:

\[
\mathbb{E}_{w \in V^2} \left[ \mathbb{E}_{x \in X} \left[ \sum_{h \in S} c_h - \min_{x' \in X^w} \sum_{h \in S} c_h \right] \right] = \sum_{w \in V^2} p(w) \left[ \sum_{x \in X^w} p(x|w) \sum_{h \in S} c_h - \min_{x' \in X^w} \sum_{h \in S} c_h \right]. \tag{8}
\]

C. Optimization Formulation

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

(i) Privacy budget for each source node \( u \): The value in (8) should be less than or equal to the budget constraint \( \eta \).

\[
\sum_{v \in V} p(w) \left[ \sum_{x \in X} p(x|w) \sum_{h \in S} c_h - \min_{x' \in X^w} \sum_{h \in S} c_h \right] \leq \eta, \quad \forall u \in V. \tag{9}
\]

Recall that \( w = (u, v) \) and in (9), 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 \( p(x|w) \) over its entire support \( X \) must equal one.

\[
\sum_{x \in X} p(x|w) = 1, \quad \forall w \in V^2. \tag{10}
\]

A valid probability has to be non-zero.

\[
0 \leq p(x|w) \leq 1, \quad \forall x \in X, w \in V^2. \tag{11}
\]

(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 \( x \). In other words, we have

\[
p(x|w) = 0, \quad \forall x \notin X^w, w \in V^2. \tag{12}
\]

This constraint is needed to restrict the valid transmissions that serve the source-destination pair \( w \).

1) General Formulation: Given a network graph \( G = (V, E) \), transmission cost \( \{c_h\}_{h \in H} \), the prior probabilities \( \{p(w)\}_{w \in V^2} \), the adversary’s observation distribution \( \{p(y|x)\}_{y \in Y, x \in X} \), and the privacy budget \( \eta \), find the path distribution \( \{p(x|w)\}_{x \in X, w \in V^2} \) that minimizes the adversary’s detection probability \( P_{\text{detect}} \) in (4) such that the expected cost of the privacy-preserving routes is at most \( \eta \) for each source node \( u \). The solution can be obtained by solving the following minimax optimization in Problem (13):
Using the newly introduced $y$ to match the value of $\text{lem}(13)$ as a linear program by introducing a variable $z_y$.

subject to

$$
\sum_{x \in X} p(x|w) = 1, \quad \forall w \in V^2
$$

$$
0 \leq p(x|w) \leq 1, \quad \forall x \in X, w \in V^2
$$

2) Linear Program Formulation: We can reformulate Problem (13) as a linear program by introducing a variable $z_y$ and the objective function at the optimal solution for each $y \in Y$, along with the following inequality constraint:

$$
z_y = \sum_{x \in X} p(y|x)p(x|w)p(w) \geq 0, \quad \forall y \in Y, w \in V^2.
$$

At the optimal solution, where $\sum_{y \in Y} z_y$ is minimized, we have $z_y = \max_w \sum_{x \in X} p(y|x)p(x|w)p(w)$ for each $y \in Y$. The detection probability of the adversary can then be expressed as:

$$
P_{\text{detect}} = \sum_{y \in Y} \max_{w \in V^2} \sum_{x \in X} p(y|x)p(x|w)p(w) = \sum_{y \in Y} z_y.
$$

Using the newly introduced $\{z_y\}_{y \in Y}$ variables, we arrive at the linear program formulated in Problem (14).

$$
\text{LPPProb}(G, \{c_h\}_{h \in H}, \{p(w)\}_{w \in V^2}, \{p(y|x)\}_{y \in Y, x \in X, \eta}):\text{minimize } \{p(x|w)\}_{x \in X, w \in V^2}, \{z_y\}_{y \in Y} \sum_{y \in Y} z_y
$$

subject to

$$
\sum_{x \in X} p(x|w) = 1, \quad \forall w \in V^2
$$

$$
0 \leq p(x|w) \leq 1, \quad \forall x \in X, w \in V^2
$$

$$
p(x|w) = 0, \quad \forall x \notin X^w, w \in V^2
$$

$$
z_y = \sum_{x \in X} p(y|x)p(x|w)p(w) \geq 0, \quad \forall y \in Y, w \in V^2
$$

$$
\sum_{w \in V^2} \left[ \sum_{x \in X} p(x|w) \sum_{h \in x} c_h \right] - \min_{x' \in X^w} \sum_{h \in x'} c_h \leq \eta, \quad \forall u \in V, \quad (14)
$$

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 $x$, the lossy adversary can observe $(|x||x|_0)$ possible observations with $k$ node transmissions (see example in Appendix-B) where $|x|_0$ is the number of nodes that transmitted in $x$. Given that the adversary may observe $k = 0, \ldots, |x|_0$ number of node transmissions for a path $x$, there are a total of $2^{|x|_0}$ possible observations $y$. The number of possible observations grows exponentially with the dimension of $x$, resulting in a combinatorial explosion. Hence, we propose an approximation method for $P_{\text{detect}}$ in the next subsection IV-C4. In addition, valid paths that contain a minimum spanning tree (MST) can be heuristically pruned to reduce the path search space $X$.

4) Approximation for the Lossy Adversary: We suggest approximating the adversary’s observation model by replacing the observation distribution $\{p(y|x)\}_{x \in X}$ values for observations with more than $n$ transmission losses from a path $x$ with zero. More formally, for each path $x$, we let $p(y|x) = 0$ if $p(y|x) < \epsilon$ where $\epsilon = (1 - \alpha)|x|_0^{-\alpha}n^\alpha$, with $n \in (0, |x|_0)$, and $\alpha \in [0, 0.5]$ is the probability of not observing a given transmission $h \in x$. A smaller parameter $\epsilon$ gives a better approximation of $P_{\text{detect}}$ but offers less computational savings. An example on the approximation method and proof that it gives a lower bound of $P_{\text{detect}}$ can be found in Appendix-B.

V. LOSSLESS ADVERSARIAL OBSERVABILITY (WORST-CASE SCENARIO)

In this section, we consider the lossless adversarial model, which is a special case of Problem (14). The lossless adversary perfectly observes each transmission path $x$ and hence, is the worst-case scenario for the network. As such, the probability of observing $y$ given that $x$ was actually transmitted:

$$
p(y|x) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases}
$$

Considering this, the objective function in the general Problem (14) can be simply replaced by $\sum_{x \in X} \max_{w \in V^2} p(w, x)$, given in (5). Similar to Section IV-C2, we introduce a variable $z_x$.
to match the value of $\max_{w \in \mathcal{V}^2} p(w, x)$, at the optimal solution, along with the following inequality constraint:

$$z_x - p(x|w)p(w) \geq 0, \quad \forall x \in \mathcal{X}, w \in \mathcal{V}^2.$$  

The inequality constraint ensures that the maximum probability, $\max_{w \in \mathcal{V}^2} p(w, x)$, is always greater than or equal to all the other $p(w, x)$ probabilities. Our optimization problem for the lossless adversary can be formulated as the linear program in Problem (15). In addition, the problem can be decomposed into smaller subproblems (see proof of separable property in Appendix-C) for each source node $u$ to solve in a distributed fashion. This allows the optimal solution to be computed in a distributed manner or in parallel for efficiency.

\[
\begin{align*}
\text{minimize} & \quad \sum_{x \in \mathcal{X}} z_x \\
\text{subject to} & \quad z_x - p(x|w)p(w) \geq 0, \quad \forall x \in \mathcal{X}, w \in \mathcal{V}^2 \\
& \quad \sum_{x \in \mathcal{X}} p(x|w) = 1, \quad \forall w \in \mathcal{V}^2 \\
& \quad 0 \leq p(x|w) \leq 1, \quad \forall x \in \mathcal{X}, w \in \mathcal{V}^2 \\
& \quad p(x|w) = 0, \quad \forall x \not\in \mathcal{X}^u, w \in \mathcal{V}^2 \\
& \quad \sum_{w \in \mathcal{V}} p(w) \left( \sum_{x \in \mathcal{X}} p(x|w) \sum_{h \in \mathcal{H}} c_h \right) \leq \eta, \\
& \quad \forall u \in \mathcal{V}. 
\end{align*}
\]

Remark: Note that the difference between Problem (15) and the non-distributed optimization in Problem (14) is that the first constraint for $z_x$ in (15) is localized. Specifically, each $x$ is localized to a specific source node $u$ whereas in (14), each $y$ could be produced by many different source nodes.

VI. SIMULATION RESULTS

In this section, we study the adversary’s detection probability $P_{\text{detect}}$ for the proposed OPERA and compare it against the Greedy and Uniform heuristics (see Appendix-D), a baseline heuristic scheme [15], and the minimization of mutual information (see Appendix-E). We varied the privacy budget $\eta$ and $\alpha$ values for $P_{\text{detect}}$ against the expected cost incurred by the schemes in various (connected) network topologies. We evaluate the schemes using the basic line, binary tree, and grid network topologies (see Fig. 3), the random topology, in addition to two other real-world topologies from the outdoor Roofnet [35] and indoor Indriya [36] testbeds. The line, tree, and grid networks were chosen to evaluate the effect of the number of neighbors per node on the amount of privacy provided by the schemes.

We assume (except for Section VI-F) that the links are symmetric, i.e., for every hyperarc $h = (i, \mathcal{R})$ in the network, there exists $|\mathcal{R}|$ hyperarcs given by $h_k = (k, \mathcal{R}_k)$ where $i \in \mathcal{R}_k$ for every $k \in \mathcal{R}$. We let the cost of each hyperarc be one, i.e., $c_h = 1$ for each $h \in \mathcal{H}$. Single-path routing was assumed from Sections VI-A to VI-D while multipaths were allowed in Sections VI-E to VI-F. Finally, we assume that $w$

![Fig. 3. Example of the used line, binary tree, and grid network topologies for our simulations.](image)

![Fig. 4. Adversary’s detection probability for the lossy observations model in a 10-node line network with different $\alpha$ and $n$ parameters. Recall that $\alpha$ is the probability of not observing a given transmission $h \in \mathcal{H}$ while $n$ is the parameter in our approximation method in Section IV-C4.](image)

A. Lossy Adversarial Observations

We solve Problem (14) to obtain the optimal $P_{\text{detect}}$ values for the proposed OPERA. Figure 4 shows the $P_{\text{detect}}$ 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, $P_{\text{detect}}$ decreases as $\alpha$ 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 $P_{\text{detect}}$ for larger $\alpha$ values. There exists a tradeoff between the value of $n$ and the complexity of the optimization problem in (14) and a higher $n$ value results in a more accurate estimate of the true $P_{\text{detect}}$ at the expense of additional computational costs. More performance degradation is experienced in the tree and grid networks 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 increase even if he uses (4) while the system uses the approximation method.
B. Comparison with Greedy and Uniform Heuristics

From this section onwards, we assume a lossless adversary. The details for the Greedy and Uniform heuristics are given in Algorithms 2 and 3 respectively in the Appendix. In the Uniform heuristic, we uniformly route the packets via all valid paths from $u$ to the $v$ (similar$^2$ to [13]) 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 $G$ 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 $x$ instead of the expected privacy budget for each source node as used in OPERA.

Figures 5a, 5b, and 5c show the $P_{\text{detect}}$ 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. 5a and 5b, 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 necessarily 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, however, does not converge to the maximum achievable privacy even when the privacy budget is slack, unlike the Greedy heuristic. In Fig. 5c, 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 $P_{\text{detect}}$ for the Uniform heuristic even when the expected cost is zero.

Lastly, we plot the average $P_{\text{detect}}$ values of the two heuristics and the proposed OPERA for five randomly generated 80-node networks with the single-path constraint in Fig. 6. The $P_{\text{detect}}$ results for the larger 80-node network have a similar trend to the results from the smaller 20-node line network in Fig. 5a where OPERA outperforms the Uniform and Greedy heuristics.

C. Comparison with Using Sink Simulation and Backbone Flooding for Privacy

We compared our proposed OPERA against an existing protocol proposed by Mehta et al. [15]. Similar to our work, Mehta et al. proposed the sink simulation and backbone flooding techniques in [14, Section 5.2] to provide location privacy for the network sinks under the same global adversary assumption as considered in our work. The sink simulation technique assigns $L$ simulated (fake) sinks, each of which will
receive all traffic sent to the real sink. On the other hand, the backbone flooding method provides privacy protection for the sink by requiring all nodes to flood their packets through a network backbone that connects the data sinks.

As the work in [15] 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 techniques 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, e.g., if the transmission paths from source 1 to destinations 2 and 3 are the same, then the total transmission cost is only counted once for the same path. For the backbone flooding technique, we do not use the proposed approximation algorithm for constructing the backbone network. Instead, we used the minimum spanning tree (which contains the minimal number of nodes) to flood a packet so that the entire network can receive it. 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.

We implemented both the sink simulation and backbone flooding techniques and plot the $P_{\text{detect}}$ values for five randomly generated 80-node networks in Fig. 7 and compared the results with the proposed OPERA using the single-path constraint. The performance of the sink simulation technique is significantly worse (up to five times higher $P_{\text{detect}}$ values) than the proposed 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 technique is slightly better than the proposed 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 Minimizing Mutual Information (MI) for Privacy

In Fig. 8, we plot the $P_{\text{detect}}$ values for the proposed OPERA and the MI minimization problem (7) for the line, binary tree, and grid networks. It is observed that minimizing MI results in a higher $P_{\text{detect}}$ 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 $P_{\text{detect}}$ 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. As the formulation for minimizing MI is non-linear due to its logarithm term, we were unable to scale the network size for minimizing MI as the optimization problem cannot be solved in a reasonable amount of time. Therefore, despite being commonly proposed as a measure for privacy [28]–[34], minimizing MI may not be ideal in every scenario as shown in our case where a MAP adversary was considered.

### E. Comparison of Single-path and Multipath Routing for Privacy

We study the effects of using multipaths $\mathcal{M} = \{x_1, x_2, \ldots\}$ where at least one path $x_i$ will reach the destination node instead of the single-path assumption used in the earlier subsections. In the multipath routing, the routing paths $x \in \mathcal{X}$ in Problem (15) are replaced by a set of paths $\mathcal{M} = \{x_1, x_2, \ldots\}$. We plot the $P_{\text{detect}}$ values for the single and multipath routing in the line, binary tree, and grid networks in Figs. 5d, 5e, and 5f 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 $P_{\text{detect}}$ 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 still be used if the privacy budget is tight. The Greedy heuristic however, performs worse in the multipath scheme when the privacy budget is tight.

F. Using Topologies from Real-World Testbeds

To evaluate the practicality of OPERA in real-world topologies, we used two topologies from the outdoor Roofnet [35] and indoor Indriya [36] testbeds, and plot the $P_{\text{detect}}$ values in Figs. 9 and 10 respectively. The Roofnet project [35] was a real-world outdoor testbed deployed for multi-hop wireless ad hoc networking research. 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 [36] 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 arbitrary selected a node from each room of the network to reduce the computation complexity (for scalability). Similar to other clustering-based approaches, we let the selected nodes (i.e., cluster head) represent all the other nodes in the same room. The selected nodes (18 in total) can then rebroadcast the received message to all the nodes in its respective room. The performance of OPERA in the two mesh-like real-world topologies are similar to our earlier results in the grid topology. There exists little differences in $P_{\text{detect}}$ 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 AND FUTURE WORK

We have developed a statistical decision-making framework to optimally solve the privacy-preserving routing problem in wireless networks given some utility constraints. In addition, we considered a powerful global adversary that uses the optimal maximum-a-posteriori (MAP) estimation strategy. Instead of using heuristic approaches that do not guarantee optimality for arbitrary network topologies and utility functions, our linear program formulations optimally solve the privacy-utility tradeoff problem. We also showed via simulations that our approach is significantly better than the Uniform and Greedy heuristics, and also a baseline scheme, and that minimizing privacy leakage via mutual information may not be optimal under our considered MAP adversary.

The future work could be to study the privacy-utility tradeoff problem for mobile networks. For mobility scenarios, our scheme may use source routing protocols like [38], [39], which extend the well-studied dynamic source routing (DSR) routing protocol for wireless mobile ad hoc networks when the mobility is limited and the mobility patterns are known. The expected amount of transmission overhead may be computed when the mobility patterns are known but the computational complexity may be very costly when there is high mobility among the nodes. Also, note that the use of DSR introduces a large communication overhead as the protocol needs to map the routing information to all other nodes via a route discovery phase, which is basically flooding-based although it uses heuristics to avoid sending duplicate packets.

Another possible future direction would be to provide strong privacy guarantees. There has been a growing trend for privacy-preserving works to adopt privacy frameworks such as $k$-anonymity [40] and differential privacy [34]. However, it is not trivial to enforce $k$-anonymity or differential privacy under our existing linear program formulations as it would require a significant change to our system model. Mainly, to apply $k$-anonymity, we will first need to redefine it for the wireless routing domain, and specify what it means to be indistinguishable from at least $k - 1$ other source-destination pairs. Meanwhile, the typical Laplace and exponential mechanisms in differential privacy cannot be directly applied to our considered routing problem as the observed sequence of node transmissions is neither numerical nor count-based, and more importantly, there is no query-based mechanism in our routing problem. Thus, it would not be applicable to add real-valued noise to the observed sequence of node transmissions. Hence, to apply differential privacy, we have to assume that the adversary is unable to observe all transmissions in the entire network but instead is limited to querying the system.
APPENDIX

A. Motivating Example: Probabilistic Routing For Enhanced Privacy

The idea of probabilistic routing existed since the first paper on location privacy [12] by Ozturk et al. In [12], the authors used a single parameter $P_{\text{forward}}$ to decide if a node forwards a packet to its neighbors (under the probabilistic flooding scheme). Subsequent dummy traffic schemes have also relied on similar parameters to probabilistically route packets. For example, Deng et al. [1] used a parameter $P_t$ to decide the probability that a node forwards a dummy packet to its parent node while Jian et al. [13] used the parameter $P_l$ to decide the probability that a node forwards a dummy packet to another decoy node located away from the destination node. The privacy parameters indirectly determine the path distribution $p(x|w)$ and a larger parameter value generally translates to more privacy at the expense of more dummy traffic. Hence, many authors studied the privacy-utility tradeoffs of their proposed schemes via simulations. Despite this, the amount of privacy provided by the randomized paths may not be optimal for a given privacy budget. Instead of using a fixed parameter, e.g., $P_{\text{forward}}$, to determine the path distribution $p(x|w)$, we let $p(x|w)$ be the optimization decision variables to be optimized. Next, we use a simple example to illustrate the selection of the optimal path distribution $p(x|w)$ and how it compares to a Uniform heuristic.

1) Basic Idea: Optimizing the Routing Paths: For the purpose of illustration, we consider a small 6-node undirected line network. The difference in the adversary’s detection probability $P_{\text{detect}}$ is more significant in a larger network and is investigated in our simulation section. We list the optimized path distribution $p(x|w)$ given a source node $u = 3$ and destinations $v \in \{1, 2, 4, 5, 6\}$ for the proposed OPERA and the Uniform heuristic (explained in Section VI-B) in Table II.

It is clear from the table that for the same amount of dummy traffic incurred as the Uniform heuristic results in a 10% increase in $P_{\text{detect}}$ (which gives lower privacy) compared to OPERA. Hence, the Uniform heuristic does not maximize the amount of privacy achieved for the same amount of dummy traffic incurred. Therefore, in this work, we focus on designing an efficient optimization formulation to select the optimal path distribution that minimizes $P_{\text{detect}}$ for a given privacy budget.

2) Example: Optimal Detection of Source-Destination Pair $w$ from an Observation $y$: Consider the same 6-node network as used in our earlier example in Section A1. Given that the adversary knows $p(w)$, $p(y|x)$, and $p(x|w)$, they are able to compute the posterior probabilities $p(w|y)$ using (2) and obtain the values in Table II. For ease of reading, we only list the source node of each hyperarc $h_i$ in the transmission paths $x$ and $y$ in the table. To illustrate how the MAP estimation in (3) works, we pick the observation $y = (3, 2)$ under the $(x = y)$ column. Suppose that the Uniform heuristic is used. Whenever the adversary observes $(3, 2)$, he can refer to the table (rows 1, 3, and 7) and examine the posterior probabilities: $p(u = 3, v = 1|y = (3, 2)) = \frac{1}{2}$, $p(u = 3, v = 2|y = (3, 2)) = \frac{1}{6}$, and $p(u = 3, v = 4|y = (3, 2)) = \frac{1}{6}$. The probabilities can be interpreted as follows - with $\frac{2}{3}$, $\frac{1}{6}$, and $\frac{1}{6}$ probabilities, the destination node is 1, 2 and 4 respectively. Our MAP adversary always picks node 1 as the destination each time he observes $y = (3, 2)$. The rationale being that he will guess correctly with a higher probability of $\frac{2}{3}$ (the maximum posterior probability) on average compared to other arbitrary approaches.

B. Computation Reduction via Likelihood Truncation

Consider a small undirected line network with only three nodes. Assume that the adversary has lossy observations where the probability of not observing a transmission $\alpha = 0.1$. The possible observations for the lossy adversary are given in Table III for an arbitrary path distribution $p(x|w)$. For ease of reading, we only list the source node of each hyperarc $h_i$ in the transmission paths $x$ and $y$ in the table. The number of possible observations $y$ is much larger compared to that of a lossless adversary who only observes the highlighted rows 1 and 5. This is because we have to consider all the combinations of a path $x$ where $1, 2, 3, \ldots$ transmissions were transmitted but not observed by the adversary. However, we can approximate the adversary’s detection probability $P_{\text{detect}}$ by simply omitting paths with more than $\alpha$ missing transmissions (e.g., the fourth row in Table III) is only 1%. Hence, we obtain a lower bound of $P_{\text{detect}}$ by omitting observations that occur with low probability.

Proposition 1. The approximation method in Section IV-C4 which uses a truncated observation distribution provides a lower bound for $P_{\text{detect}}$ in Problem (14).

Proof. We show that the feasible region in Problem (14) becomes larger in the approximation method which truncates possible observations, leading to a lower $P_{\text{detect}}$ value. Let the truncated observation probability be $q(y|x) = p(y|x)$, if $p(y|x) \geq \epsilon$, and $q(y|x) = 0$, otherwise. Consider the $z_y$ constraint in Problem (14) which can be rewritten as $z_y \geq \sum_{x \in X} p(y|x)p(x|w)p(w), \forall y \in \mathcal{Y}, w \in \mathcal{V}^2$. The set of possible $z_y$ that satisfies the constraint $z_y \geq \sum_{x \in X} p(y|x)p(x|w)p(w)$ is a subset of the set of possible $z_y$ that satisfies the constraint $z_y \geq \sum_{x \in X} q(y|x)p(x|w)p(w)$.
since \( q(y|x) \leq p(y|x) \). Hence, we obtain a larger feasible region when we use the truncated probability \( q(y|x) \). This may lead to a lower objective function value in the minimization problem which serves as a lower bound for \( P_{\text{detect}} \).

\[ \square \]

C. Problem (15) is Block Separable

We show that the linear program in Problem (15) can be decomposed into smaller subproblems for each individual source node \( u \) to solve independently. The key idea in proving the block separable property is that there are no complicating variables in the objective function. Recall that in a lossless observation, the observed source node \( u \) must be the first node to transmit. Hence, two routing paths \( x_1 \) and \( x_2 \) made by two different sources \( u_1 \) and \( u_2 \) cannot be observed to be the same observation, i.e., \( y_1 \neq y_2 \). In other words, given a \( w = (u, v) \) pair, the term \( p(y|w = (u', v)) = 0 \) for all \( u' \in V, u' \neq u \). Let \( a \) and \( b \) represent the column vector of decision variables \( p(x|w) \) and \( z_x \) respectively. The objective function \( \sum_{x \in \mathcal{X}} z_x \) is a function of only \( b \). This allows the variable objective function to be partitioned into \( |V| \) summations of the subvectors \( b_1, b_2, \ldots \) 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 \( b_i \) that correspond to a source node \( u_i \). Therefore, we have shown that Problem (15) is block separable.

D. Example: Greedy and Uniform Heuristics

Consider a 3-node line network, similar to our earlier example in Appendix-B. The path selection under the Greedy and Uniform heuristics is illustrated in Table IV. For ease of reading, we only list the source node of each hyperarc \( h_i \) in the transmission paths \( x \) and \( y \) in the table. Given \( w = (1, 2) \), the Uniform heuristic will select routing paths \( x = (1) \) and \( x = (1, 2) \) with equal probability, see rows 1 and 2. The Greedy heuristic on the other hand, always picks \( x = (1, 2) \) over \( x = (1) \) with probability one. Notice that the Greedy heuristic is deterministic (i.e., it always picks the path containing the most number of receivers) while the Uniform heuristic is probabilistic (i.e., it uniformly picks a path \( x \) from the set of valid paths \( \mathcal{X}^w \)).

\[ \text{Algorithm 2: Greedy routing for preserving source-destination privacy.} \]

1 GreedyRouting( \( G, \{c_h\}_{h \in \mathcal{H}}, \{p(w)\}_{w \in \mathcal{X}^w}, \eta, w \) )

Input : Network graph \( G \), transmission cost \( c_h \), the prior probabilities \( \{p(w)\}_{w \in \mathcal{X}^w} \), privacy budget \( \eta \), and source-destination \( w = (u, v) \).

Output: Path distribution \( p(x|w) \).

2 Compute the cost of the minimum-cost path from \( u \) to \( v \):

\[ c_{\min}(w) = \min_{x \in \mathcal{X}^w} \sum_{h \in x} c_h. \]

3 Compute the set of paths \( \mathcal{X}' \) that satisfy the privacy budget \( \eta \) where

\[ \mathcal{X}' = \{x : x \in \mathcal{X}^w, \sum_{h \in x} c_h - c_{\min}(w) \leq \eta\}. \]

4 Initialize \( p(x|w) = 0 \) for all \( x \in \mathcal{X}' \).

5 Select a path \( x^* \) with the most number of receivers, i.e.,

\[ x^* = \arg\max_{x \in \mathcal{X}'} \{|r : r \in R_1, h_i \in (s_i, R_1), h_i \in x\}. \]

6 Set \( p(x^*|w) = 1 \) and use routing path \( x^* \).

E. Using (Information-Theoretic) Mutual Information to Quantify Privacy

Mutual information [28]–[34] has been used in the literature to measure the information leakage of some anonymous event. Specifically, in our context, it can be used to quantify the source-destination anonymity of a protocol (or to measure the privacy leakage by the routing protocol):
TABLE III
POSSIBLE OBSERVATIONS $y$ FOR $u = 1$ AND THEIR CORRESPONDING LIKELIHOOD $P(Y = y \mid W = w)$, AND POSTERIOR PROBABILITY $P(W = w \mid Y = y)$ FOR $\alpha = 0.1$ IN A 3-NODE LINE NETWORK GIVEN THAT $\pi = (1, 2)$. THE HIGHLIGHTED ROWS (*) REPRESENT THE ACTUAL TRANSMISSION PATH $\pi$ WHICH MAY NOT ALWAYS BE OBSERVED PERFECTLY DUE TO LOSSY OBSERVATIONS. THE EMPTY TUPLE () IN THE OBSERVED PATH COLUMN represents the case where no transmission is observed.

<table>
<thead>
<tr>
<th>Source-dest. pair $w$</th>
<th>Prior prob. $P(W = w)$</th>
<th>Actual path $\pi$</th>
<th>Path distribution $p(X = \pi \mid W = w)$</th>
<th>Observed path $y$</th>
<th>Likelihood $P(Y = y \mid X = \pi)$</th>
<th>Posterior prob. $P(W = w \mid Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>* (1, 2)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(1, 2)</td>
<td>0.81</td>
<td>0.5</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(1)</td>
<td>0.09</td>
<td>0.5</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(2)</td>
<td>0.09</td>
<td>0.0417</td>
</tr>
<tr>
<td>* (1, 3)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(1)</td>
<td>0.01</td>
<td>0.0417</td>
</tr>
<tr>
<td>(1, 3)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(2)</td>
<td>0.09</td>
<td>0.0417</td>
</tr>
<tr>
<td>(1, 3)</td>
<td>1/6</td>
<td>(1, 2)</td>
<td>1</td>
<td>(1)</td>
<td>0.01</td>
<td>0.0417</td>
</tr>
</tbody>
</table>

Definition 3 (Privacy Leakage). Consider a graph $G$. Let $W$ be a random variable representing the source-destination pair $w \in V^2$, and $Y$ be a random variable representing the observed node transmission path $y \in Y$. Let $p(w)$ be the prior probability for a source node $u$ to communicate with the destination node $v$, and $p(w, y)$ be the joint probability of observing both $y$ and $w$. The average amount of information (or privacy) leaked by a protocol is measured by the mutual information, which is given by

$$I(W; Y) = H(W) - H(W|Y)$$

where $H(W)$ is the Shannon entropy (or uncertainty) of $W$ given by $H(W) = \sum_{w \in V^2} p(w) \log p(w)$, and $H(W|Y)$ is the conditional entropy of $W$ given by $H(W|Y) = \sum_{w \in V^2} \sum_{y \in Y} p(w, y) \log \frac{p(w)}{p(w|y)}$.

Entropy $H(W)$ can be interpreted as the amount of information needed (by an adversary) to identify the $w \in V^2$ pair while mutual information $I(W; Y)$ is the reduction in the uncertainty of $W$ due to the knowledge of $Y$. $I(W; Y)$ can be used to quantify the loss of privacy for $W$ (or equivalently, the amount of privacy leakage) of the protocol where a smaller $I(W; Y)$ value indicates better privacy for $W$. The value of $I(W; Y)$ is zero (minimum) when $W$ and $Y$ are independent and equals to $H(W)$ (maximum) when $A$ is a deterministic function of $Y$.

Algorithm 3: Uniform routing for preserving source-destination privacy.

1. UniformRouting ($G, \{c_h\}_{h \in H}, \{p(w)\}_{w \in V^2}, \eta, w$)
2. Compute the cost of the minimum-cost path from $u$ to $v$: $c_{\text{min}}(w) = \min_{w \in X'} \sum_{h \in H} c_h$
3. Compute the set of paths $X'$ that satisfy the privacy budget $\eta$ where $X' = \{\pi : \pi \in X', \sum_{h \in H} c_h - c_{\text{min}}(w) \leq \eta\}$
4. Randomly select a routing path $\pi$ according to the path distribution $p(\pi | w)$.

REFERENCES


