A Rate-Distortion Approach to Anonymous Networking

Parvathinathan Venkitasubramaniam and Lang Tong
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853
Email : {pv45, lt35}@cornell.edu

Abstract— **The problem of hiding data routes in a wireless network from eavesdroppers is considered. Using Shannon's equivocation as a measure of anonymity of routes, scheduling and relaying protocols are designed to guarantee anonymity. A duality of this problem to information-theoretic rate-distortion is used to maximize network throughput for any specified level of anonymity. The achievability of this throughput, however, requires each node to have knowledge of all routes in a network session. When each node only has access to local information about routes, a decentralized strategy is proposed, and the achievable throughput is characterized using a constrained distortion-rate optimization.**

Index Terms— **Network Security, Traffic Analysis, Secrecy, Rate-Distortion.**

I. Introduction

Passive monitoring of node transmissions in a network can reveal significant information about network operation even when packets are encrypted. Although contents of communication are encrypted, statistical analysis of packet transmission epochs can reveal critical networking information such as paths of data flow and source-destination pairs. Prevention of transmission time based analysis necessitates a redesign of network protocols so that the communication routes appear obfuscated to eavesdroppers at minimum cost of network performance. In this work, we present a theoretical approach to anonymous networking in the context of multihop wireless networks. In particular we are interested in the tradeoff between "anonymity" of routes and achievable network performance.

Prevention of traffic analysis is a classical problem, and a dominant portion of prior research has centered around Internet applications [1], [2], [3]. In that regard, an important countermeasure was provided by Chaum, through the concept of the traffic mix [4]. A mix is a special relay node or router that batches and reorders packets from multiple sources, thereby decorrelating the timing of incoming and outgoing packets to the mix. In a multihop network, each source picks an arbitrary sequence of mixes to relay its packets and performs layered encryption. Each mix node has access to one layer, and is only aware of its immediate neighbours in the route. This ensures that active compromising of one or two mix nodes is not sufficient to reveal source-destination information of an observed packet.

In wireless networks, the key challenge in countering traffic analysis is adhering to the network constraints such as medium access, latency and stability. Although mixing was well suited to design anonymous remailers and proxy systems for the Internet, the batching strategies were found to be vulnerable to traffic analysis [5] under delay or buffer constraints. An alternative approach, designed primarily for multihop wireless networks is that of deterministic scheduling [6]. In [6], the authors propose a fixed periodic schedule for the entire network, wherein nodes adhered to the schedule by transmitting dummy packets whenever they did not have actual data. While the fixed scheduling idea can be adapted to handle delay constraints, the centralized synchronous implementation renders it impractical for ad hoc wireless networks.

In [7], we proposed asynchronous scheduling and relaying strategies for wireless relay nodes such that the incoming and outgoing streams at the node are uncorrelated, but the relayed packets satisfied tight delay constraints. In particular, we characterized the set of achievable rates for a multiaccess relay when incoming and outgoing schedules are independent Poisson point processes. Independent scheduling ensures that the relay operation is "hidden" from an eavesdropper at the expense of dropped packets and lower relay rates. While a direct extension of the independent scheduling would guarantee perfect secrecy at all times, such a strategy can prove detrimental to throughput, particularly in large networks. In this work, we propose a randomized scheduling strategy where, depending on the active routes, a subset of relay nodes are chosen to perform independent scheduling (as in [7] so that performance loss is minimized for any specified level of anonymity.

A key component in our approach is the analytical model for anonymity of routes. In the context of mix networks, the size or entropy of the anonymity set (set of possible source-destination pairs) of an observed packet has been used to quantify the anonymity of that packet. The use of anonymity sets suffers from two weaknesses. First, hiding source-destination alone may not be sufficient, the direction of data flow could also reveal critical information. Second, the measure of anonymity needs to cater to streams of packets rather than a single packet [5]. The model we propose is based on the information theoretic notion of equivocation, proposed by Shannon [8]. While previous applications of equivocation measured the secrecy of transmitted data on point-to-point channels [9], [10], we use equivocation to

This work is supported in part by the National Science Foundation under awards CCF-0635070 and CCF-0728872, and the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program DAAD19-01-2-0011.

measure the secrecy of routes in a network. Based on the defined metric, we are interested in characterizing the achievable network performance as a function of anonymity. A key insight in characterizing this tradeoff is the duality between anonymous networking and rate distortion, which extends beyond our scheduling strategy, and can be explained using a general intuition.

The objective of the rate-distortion problem is to generate fewest number of codewords for a set of source sequences, such that the corresponding reconstruction sequences satisfy a specified distortion constraint. The idea is to divide the set of source sequences into fewest number of bins such that the distortion between each sequence in a bin and the reconstruction sequence is less than the specified constraint. Alternatively, if the code rate is fixed, then the number of bins is fixed. Then, the sequences are placed optimally within each bin such that the corresponding reconstruction sequences minimize the expected distortion.

In the anonymous networking setup, let the set of active routes at any given time be referred to as a network session. The key idea is to divide the set of all possible network sessions into bins such that, for each bin, there exists a scheduling strategy that would make the sessions within that bin indistinguishable to an eavesdropper. The level of anonymity required determines the number of bins, and the optimal scheduling strategy plays the role of the reconstruction sequence by minimizing the performance loss across sessions within the bin.

In the remainder of this paper, we provide the formal setup of the problem, describe our randomized scheduling strategy and characterize the throughput-anonymity tradeoff using the duality to rate distortion. Further, we also propose a decentralized solution to the anonymous networking problem and characterize the corresponding throughput. The paper is organized as follows. In Section II, we describe the system model and define the analytical measure of anonymity. In Section III, we describe our randomized scheduling strategy to provide anonymity to routes. In Section IV, the duality to rate-distortion is used to characterize the achievable network performance. The decentralized approach and the corresponding performance characterization are presented in Section V.

II. Problem Setup

Let the network be represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes in the network and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of directed links. If (A, B) is an element of $\mathcal{E}$, then node B can receive transmissions from node A. A sequence of nodes $P = (V_1, \cdots, V_n) \in \mathcal{V}^*$ is a *valid path* in $\mathcal{G}$ if $(V_i, V_{i+1}) \in \mathcal{E}$, $\forall i < n$. The set of all possible paths in $\mathcal{G}$ is denoted by $\mathcal{P}(\mathcal{G})$.

We assume that during any network observation by the eavesdropper, a subset of nodes communicate using a fixed set of paths. We call this set of paths $\mathbf{S} \in 2^{\mathcal{P}(\mathcal{G})}$ a network *session*. The set of all possible sessions is denoted by $\mathcal{S}$.

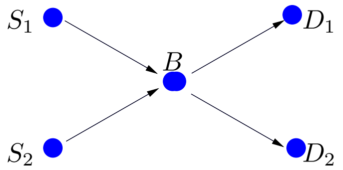

Fig. 1: Two Node Switching Network: $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{S_1, S_2, B, D_1, D_2\}$, $\mathcal{E} = \{(S_1, B), (S_2, B), (B, D_1), (B, D_2)\}$.

Consider the example $\mathcal{G}_1$ shown in Figure 1. Let S_1, S_2 be the sources and D_1, D_2 the destinations. Further, let S_1, S_2 always communicate with distinct destinations. Here,

$$\begin{aligned}\mathcal{P}(\mathcal{G}_1) \;\; = \;\; & \{(S_1, B), (S_1, B, D_1), (S_1, B, D_2), (S_2, B), \\ & (S_2, B, D_1), (S_2, B, D_2), (B, D_1), (B, D_2)\ \}.\end{aligned}$$

However, since destinations are always distinct,

$$\begin{aligned}\mathcal{S} = \;\; \{ \;\; & \{(S_1, B, D_1), (S_2, B, D_2)\} \\ & \{(S_1, B, D_2), (S_2, B, D_1)\}\, \}.\end{aligned}$$

The information that we wish to hide from the eavesdropper is the network session $\mathbf{S}$. We model $\mathbf{S}$ as an i.i.d. random variable $\mathbf{S} \sim p(\mathbf{S})$, where the prior $p(\mathbf{S})$ is obtained using the topology and applications of the particular network. We assume that the prior probabilities are available to the eavesdropper as well. For the purpose of obtaining an analytical characterization of throughput-anonymity tradeoff, we have used a mathematical abstraction that might deviate from the real network operation. We do believe that the insights obtained in this restricted setting will provide design guidelines for real applications.

Transmission Schedules The eavesdroppers' observation comprise of the packet transmission epochs in a session. Since it is not possible to determine the location of the eavesdroppers, we assume that all transmissions are being monitored. Depending on the physical layer model, it may be possible to infer partial information about sender-receiver nodes of packets by merely detecting a transmission. The physical layer we consider is a transmitter directed signaling model.

Transmitter Directed Signaling: All packets transmitted by a particular node are modulated using the same spreading sequence, and each transmitting node is associated with a unique orthogonal spreading sequence. Under this transmission scheme, an eavesdropper would be able to "tune" his detector to a particular spreading sequence and detect the transmission times of packets sent by the corresponding node. Although he knows the transmitting node of each packet, we assume that headers are encrypted, so he would not know the intended recipient of any packet.

Observable Scheduling Let $\mathcal{Y}_A$ represent the transmission epochs of node A. The schedule $\mathcal{Y}_A$ is given by a point process

$$\mathcal{Y}_A = \{Y_A(1), Y_A(2), \cdots\},$$

where $Y_A(i)$ represents the transmission epoch of the i^{th} packet sent by node A. Since we cannot determine which nodes are being monitored, the eavesdroppers' complete observation is assumed to be $\mathcal{Y} = \{\mathcal{Y}_A : A \in \mathcal{V}\}$.

We model $\mathcal{Y}$ as a sequence of random variables with conditional distribution $q(\mathcal{Y}|\mathbf{S})$. The idea is to design $q(\mathcal{Y}|\mathbf{S})$ such that eavesdroppers obtain minimum information about the session $\mathbf{S}$ by observing $\mathcal{Y}$.

A. Anonymity Measure

We define anonymity using the notion of equivocation [8], which measures the uncertainty of the information we wish to hide ($\mathbf{S}$) given the complete observation of the eavesdropper ($\mathcal{Y}$).

Definition 1: A distribution $q(\mathcal{Y}|\mathbf{S})$ is defined to have anonymity α if

$$\frac{H(\mathbf{S}|\mathcal{Y})}{H(\mathbf{S})} \geq \alpha.$$

When $\alpha = 1$, the distribution $q(\mathcal{Y}|\mathbf{S})$ is defined to have *perfect anonymity*. It is easy to see that the schedule $\mathcal{Y}$ does not provide any information about the routes. In other words, $H(\mathbf{S}|\mathcal{Y}) = H(\mathbf{S})$. For a general α, the physical interpretation comes from Fano's Inequality [11]: If the error probability of the eavesdropper in decoding the session $\mathbf{S}$ is P_e, then,

$$P_e \geq \frac{H(\mathbf{S}|\mathcal{Y}) - 1}{\log|\mathcal{S}|} \geq \frac{\alpha H(\mathbf{S}) - 1}{\log|\mathcal{S}|}.$$

Therefore, if $\mathcal{S}$ is a large set with uniform prior, then P_e is lower bounded by α.

B. Network Constraints and Throughput

Wireless networks, due to shared bandwidth and power limitations, pose constraints on transmission rates and latency of packets. The challenge in designing the schedule distribution $q(\mathcal{Y}|\mathbf{S})$ with provable anonymity is to sacrifice minimum performance under these constraints. In this work, we measure performance using network throughput subject to medium access and delay constraints, which are described as follows.

Medium Access Constraints We consider long streams of packets, and measure the packet transmission rate of node A as

$$T_A = \lim_{n\to\infty} \frac{n}{Y_A(n)}. \quad (1)$$

Owing to transmitter directed signaling, transmission rate T_A of each node $A \in \mathcal{V}$ is bounded independently by a constant C_A, which depends on characteristics of the medium and transmission capability of node A. If $T_A \leq C_A$, successful reception is guaranteed at node A. We assume that the network operates in full duplex mode, where every node can transmit and receive packets simultaneously as long as the transmission rates are within the specified bounds. Therefore, $\mathcal{Y}$ is a *valid network schedule* if and only if $T_A \leq C_A$ for every node A.

Latency Constraint: We consider delay sensitive traffic, where the packet delay at an intermediate relay A is bounded by Δ_A. Each relay is allowed to reencrypt packets, reorder arrived packets and transmit dummy packets. However, each received data packet at a node A is required to be forwarded within Δ_A time units of arrival, or otherwise, dropped. Such a delay constraint is of particular importance to time-sensitive network applications such as target tracking in sensor networks or streaming media on peer-to-peer networks. In general, a strict delay constraint would also ensure stability, albeit at the cost of dropped packets.

Relaying Strategy The schedules in $\mathcal{Y}$ only denote when packets are transmitted by each node, and do not specify the routes or indicate which packets actually travel from source to destination on each route of a session. For every schedule, we therefore specify a relaying strategy, which is expressed as a set of subsequences from the schedule $\mathcal{Y}$. This set of subsequences $\mathcal{Z}$ contains the transmission epochs of packets that are relayed from sources to destinations within the delay constraints, and is a function of the routes in the session.

Definition 2: Let a session $\mathbf{S} = (P_1, \cdots, P_{|\mathbf{S}|})$, where each $P_i = (A(i,1), \cdots, A(i,m(i)))$ is a valid path of length $m(i)$, and $A(i,j) \in \mathcal{V}$ represents the j^{th} node in path P_i. A set of subsequences $\mathcal{Z} = \{\mathcal{Z}_{i,j} : i \leq |\mathbf{S}|, 1 \leq j < m(i)\}$ of $\mathcal{Y}$ is a *valid relaying strategy* for $\mathbf{S}$ if:

1) $\forall i \leq |\mathbf{S}|, 1 \leq j < m(i)$, $\mathcal{Z}_{i,j} \subset \mathcal{Y}_{A(i,j)}$.
2) For every $i \leq |\mathbf{S}|$, $\{\mathcal{Z}_{i,j} : j < m(i)\}$ satisfy

$$0 \leq Z_{i,j}(n) - Z_{i,j+1}(n) \leq \Delta_{A(i,j)},\ \forall n. \quad (2)$$

3) If $(A(i,j), A(i,j+1)) = (A(l,m), A(l,m+1))$, then $\mathcal{Z}_{i,j} \cap \mathcal{Z}_{l,m} = \phi$.

In the above definition, condition 2 guarantees that the relayed packets satisfy the delay constraint at every intermediate relay. Condition 3 ensures that, if any pair of nodes is common to multiple routes, the subsequences picked from the transmission schedules are mutually exclusive.

C. Performance Metric

For a given level of anonymity α, we are interested in designing $\mathcal{Y}, \mathcal{Z}$ for every session $\mathbf{S}$, such that the network throughput is maximized. We define network throughput as the mean sum-rate of packets relayed from the sources to the destinations in a session.

It is possible that the set of subsequences $\mathcal{Z}$ are a strict subset of the transmission schedule $\mathcal{Y}$, or in other words, there are epochs in $\mathcal{Y}$ that do not represent any relayed packets. Those transmission epochs would correspond to dropped packets or dummy packet transmissions. Since all epochs in $\mathcal{Y}$ do not represent relayed packets, the network performance is measured by the rates of relayed packets in $\mathcal{Z}$. The rates of relayed packets in session $\mathbf{S}$ are denoted by a vector $\lambda(\mathbf{S}, \mathcal{Z}) = (\lambda_1, \cdots, \lambda_{|\mathbf{S}|})$, where

$$\lambda_i = \lim_{n\to\infty} \frac{n}{Z_{i,1}(n)},\ \forall i.$$

Note that since all the subsequences on any particular route have same length, it is sufficient to use $\mathcal{Z}_{i,1}$ to compute the rate.

Definition 3: R is defined to be an *achievable throughput with anonymity* α if $\exists q(\mathcal{Y}|\mathbf{S})$ with anonymity α such that

1) For every session $\mathbf{S} = \{P_1, \cdots, P_{|\mathbf{S}|}\}$, every realization of $\mathcal{Y}$ given $\mathbf{S}$ is a valid network schedule.
2) For every realization of $(\mathbf{S}, \mathcal{Y})$, there exists a valid relaying strategy $\mathcal{Z}$, such that

$$\mathbb{E}\left(\sum_{i=1}^{|\mathbf{S}|} \lambda_i(\mathcal{Z}, \mathbf{S})\right) \geq R, \tag{3}$$

where the expectation is over the joint pdf of $\mathcal{Y}$ and $\mathbf{S}$.

Note that design of probability distributed function $q(\mathcal{Y}|\mathbf{S})$ has an inherent assumption of centralized scheduling where knowledge of the entire session is used to generate the transmission schedules $\mathcal{Y}$. In Section V, we propose a decentralized scheduling strategy where each node independently decides its transmission schedule based on the local information available, at the cost of lower throughput.

III. Anonymous Scheduling Strategy

Our approach to designing schedules and relay strategies derives its motivation from Mix networks, but differs in several key aspects related to wireless networking. First, owing to encrypted packet headers, if incoming and outgoing schedules at a particular node are uncorrelated, an eavesdropper would not be able to detect the flow of traffic through that node. Therefore, it is not always required to Mix multiple flows to hide the relaying operation. Second, depending on the level of anonymity, it may not be necessary to hide every link of communication. It is possible to reveal certain portions of the routes to the eavesdropper without giving information about the session $\mathbf{S}$.

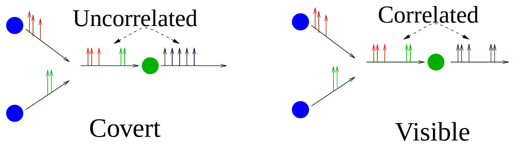

Fig. 2: Visible and Covert Relaying

Specifically, in a session $\mathbf{S} = (P_1, \cdots, P_{|\mathbf{S}|})$, we let each relay node to operate in one of two transmission modes, *covert* and *visible*, which are defined as follows.

Covert Relays: A relay B is *covert*, if its outgoing transmission schedule is statistically independent of the transmission schedules of all nodes occurring previously in the paths that contain B. Since nodes employ transmitter directed scheduling, when a relay is covert, it would be impossible for an eavesdropper to correlate the outgoing transmission schedule of the preceding node in the path and the outgoing schedule of the relay.

Visible Relays: A *visible* relay B generates its schedule depending on the arrival times of packets at B. For every received packet, the relay schedules an epoch after a processing delay (negligible compared to Δ). It is evident that the schedules of streams transmitted by a preceding node in the path and the relay would be highly correlated, and the eavesdropper would detect the relay operation[*]. Note that some of the arriving packets to the relay could be dummy packets, which are also relayed by a visible relay.

By appropriately selecting which relays should be covert in a session, we guarantee the required level of anonymity to the routes. A trivial strategy would be to let all nodes act as covert relays in a session. However, each covert relay incurs a loss in relay rates, which would accrue at every covert relay thereby reducing network performance [12] It is, therefore, necessary to pick the covert relays optimally so that anonymity is guaranteed with minimum loss in throughput. In the following exposition, we first present the achievable rate region for a covert relay (from [7]) and then discuss the randomized strategy to choose covert relays in a session.

A. Covert Relaying

In [7], we had considered a general multiaccess relay (see Figure 3), and provided covert relaying strategies to satisfy a strict delay constraint. We characterized achievable rate regions analytically, when the sources and the relay generate independent Poisson transmission schedules.

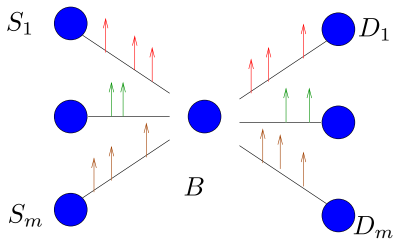

Fig. 3: Two Hop Network: Source S_i transmits packets to D_i through B

Specifically, consider a multiaccess relay as shown in Figure 3, where source nodes $S_1, \cdots, S_m$ transmit to destinations $D_1, \cdots, D_m$ respectively through relay B. Let C_A denote the transmission rate constraint on node A. If the transmission rates of packets from the sources are $T_{S_1}, \cdots, T_{S_m}$ respectively, then let the achievable relay rates be denoted by $\lambda(S_1, B), \cdots, \lambda(S_m, B)$. We restate one of the results in [7], albeit worded a little differently:

Theorem 1: (from [7]) If $\lambda(S_i, B) = T_{S_i}(1 - \epsilon(S_i, B))$, then

1) $\{\lambda(S_i, B)\}$ is achievable if

$$\epsilon(S_i, B) \geq f_e(\sum_j C_{S_j}, C_B) \forall i. \tag{4}$$

where

$$f_e(x, y) = \begin{cases} \frac{C_{D_1} - C_B}{C_{D_1} e^{-\Delta(C_B - C_{D_1})} - C_B} & C_B \neq C_D \\ \frac{C_B^2 \Delta}{1 + C_B \Delta} & C_B = C_{D_1} \end{cases},$$

[*] By tuning the detector to the spreading sequences of successive nodes in a path, the eavesdropper can detect the correlation in schedules to identify the path of traffic flow through the relay.

2) $\{\lambda(S_i, B)\}$ is not achievable if

$$\sum_i \epsilon(S_i, B) \leq f_e(\sum_i C_{S_i}, C_B), \quad \epsilon_i \leq f_e(C_{S_i}, C_B). \tag{5}$$

The results in [7] have been generalized to an average delay constraint in [13], and similar results have been derived for a receiver directed signaling model in [14]. Although the independent Poisson scheduling results in dropped packets, it is possible to achieve the stated relay rates reliably by performing forward error correction on a sufficiently long stream of packets [7].

B. Covert Relay Selection

To optimize the choice of covert relays in a session, we assume that the transmission times of packets by source nodes in a session are generated using independent Poisson processes. Accordingly, the covert relays also generate independent Poisson schedules. Given session $\mathbf{S}$ and set of covert relays $\mathbf{B}$, sequences $\mathcal{Y}, \mathcal{Z}$ can be obtained using the relaying algorithms designed in [7] and the achievable rates at a single covert relay would be given by Theorem 1.

The set of covert relays $\mathbf{B}$ is modeled as a random variable with a conditional probability mass function $\{q(\mathbf{B}|\mathbf{S}) : \mathbf{B} \in 2^V, \mathbf{S} \in \mathcal{S}\}$. We model the choice of relays as a random quantity because randomization increases secrecy [10]. The goal is to optimize the conditional p.m.f $\{q(\mathbf{B}|\mathbf{S})\}$ so that network throughput is maximized for a given level of anonymity α.

Eavesdropper Observation We assume that when a relay is visible, the eavesdropper perfectly correlates the schedules transmitted by a preceding node in a path and that of the relay. As a result, depending on the set of visible relays, the eavesdropper makes a partial detection on the paths of a session. We denote this partial session as a set of paths $\hat{\mathbf{S}} \in 2^{\mathcal{P}(\mathcal{G})}$, which is a function of the actual session $\mathbf{S}$ and the set of covert relays $\mathbf{B}$.

We define function $t : 2^{\mathcal{P}(\mathcal{G})} \times \mathcal{V} \rightarrow 2^{\mathcal{P}(\mathcal{G})}$ to characterize the eavesdropper's observation when at most one relay is covert. For a set of paths $\mathbf{P}$, $t(\mathbf{P}, B)$ contains the observed paths when only node B is covert. If $B = \phi$, then $t(\mathbf{P}, \phi)$ is obtained by removing the destination nodes from every path in $\mathbf{P}$. This is because, even if all relays are visible, receiver directed signaling ensures that it is not possible to detect the final destination in any route. If $B \neq \phi$, then a path $P \in \mathcal{P}(\mathcal{G})$ belongs to $t(\mathbf{P}, B)$ if and only if it satisfies one of the following conditions:

1. $\exists P' = (A_1, \cdots, A_k, B, A_{k+1}, \cdots, A_n) \in \mathbf{P}$, such that $P = (A_1, \cdots, A_k)$ or $P = (B, A_{k+1}, \cdots, A_n)$.
2. $P \in \mathbf{P}$ and $B \notin P$.

Condition 1 states that, when a path in $\mathbf{P}$ contains a covert relay, the eavesdropper would observe two different paths, one terminating before B and the other originating from node B. Condition 2 states that a path that does not contain a covert relay is fully observed. When a subset $\mathbf{B} = (B_1, \cdots, B_m) \subset V$ of relays are covert, then $\hat{\mathbf{S}}$ can be obtained by repeated application of $t()$:

$$\hat{\mathbf{S}} = t(\cdots(t(t(\mathbf{S}, \phi), B_1)\cdots), B_m) \overset{\triangle}{=} \mathbf{T}(\mathbf{S}, \mathbf{B}). \tag{6}$$

For the purpose of optimizing the choice of relays, it is sufficient to use the derived eavesdropper observation $\hat{\mathbf{S}}$, as is evident from the following lemma.

Lemma 1: If $\hat{\mathbf{S}} = \mathbf{T}(\mathbf{S}, \mathbf{B})$, then

1) $\hat{\mathbf{S}}$ is a sufficient statistic for detecting $\mathbf{S}$ using $\mathcal{Y}$.

2) Given $\mathbf{S}$, $\hat{\mathbf{S}}$ is an invertible function of $\mathbf{B}$.

The above lemma shows that, fo an eavesdropper, the information contained in $\mathcal{Y}$ about $\mathbf{S}$ is completely encapsulated in the observed session vector $\hat{\mathbf{S}}$. Further, the pairs $(\mathbf{S}, \mathbf{S})$ and $(\mathbf{S}, \hat{\mathbf{S}})$ are isomorphic, or in other words, there is a one-one correspondence between the two pairs of variable. Therefore, choosing the covert relays $\mathbf{B}$ is equivalent to designing the eavesdropper observation $\hat{\mathbf{S}}$.

C. Throughput Function

In order to obtain the optimal $q(\mathbf{B}|\mathbf{S})$, we need to characterize the throughput under covert relaying. The relaying strategies in [7] were designed to maximize achievable rates at a single covert relay. Extending those results to multihop routes, we can characterize the loss in sum-rate for each session $\mathbf{S}$, when a subset of relays $\mathbf{B}$ are covert.

When anonymity $\alpha = 0$, the maximum sum-rate in a session $\mathbf{S}$ is achieved when all relays are visible. This maximum sum-rate can be characterized using the max-flow in $\mathbf{S}$ that satisfies medium access constraints. Let $\lambda^v(\mathbf{S}) = (\lambda_1^v, \cdots, \lambda_{|\mathbf{S}|}^v)$ represent the vector of achievable relay rates for the paths in session $\mathbf{S}$ with no covert relays, and $\Lambda^v(\mathbf{S})$ be the maximum sum-rate. If $\mathbf{S} = (P_1, \cdots, P_{|\mathbf{S}|})$, then

$$\Lambda^v(\mathbf{S}) \quad = \quad \max(\lambda_1^v + \cdots + \lambda_{|\mathbf{S}|}^v), \tag{7}$$

$$\sum_{i: B \in P_i} \lambda_i^v \quad \leq \quad C_B, \ \forall B \in V. \tag{8}$$

The maximum network throughput when anonymity $\alpha = 0$ is given by the expected sum-rate (expectation over $p(\mathbf{S})$)

$$R(\alpha = 0) = \mathbb{E}(\Lambda^v(\mathbf{S})).$$

When the relays in a subset $\mathbf{B}$ are covert, the loss in sum-rate depends on the delay requirement at each covert relay in $\mathbf{B}$. Let $\lambda^c(\mathbf{S}, \mathbf{B}) = (\lambda_1^c, \cdots, \lambda_{|\mathbf{S}|}^c)$ represent the achievable relay rates from sources to destinations for a session $\mathbf{S} = (P_1, \cdots, P_{|\mathbf{S}|})$, when nodes in $\mathbf{B}$ are covert, and let $\Lambda^c(\mathbf{S}, \mathbf{B}) \overset{\triangle}{=} \sum_{i=1}^{|\mathbf{S}|} \lambda_i^c$ be the achievable sum-rate. If $A(i, j)$ represents the j^{th} node in path P_i, then

$$\lambda_i^c \quad = \quad \lambda_i^0 \prod_{j: A(i,j) \in \mathbf{B} \cap P_i} (1 - \epsilon_i(A(i, j-1), A(i, j))) \tag{9}$$

where $\epsilon_i(A, B)$ represents the fraction of packets transmitted by node A on path P_i, that are dropped by covert relay B. Note that Theorem 1 provides the closed form expression for $\epsilon_i(A, B)$, if B is the first covert relay in path P_i. Since the departure epochs of data packets from a covert relay does not constitute a Poisson process, the expression cannot be applied to subsequent covert relays. The analytical characterization of multiple covert relays is generally cumbersome [13], but can be obtained numerically.

IV. Performance Characterization

A. Throughput-Anonymity Tradeoff

In order to maximize network performance with anonymity α, we need to optimize $\{q(\mathbf{B}|\mathbf{S})\}$ for every session $\mathbf{S}$ using the derived eavesdropper observation and throughput characterization. For a given α, the optimal distribution $q(\mathbf{B}|\mathbf{S})$ can be obtained using a brute force search over a large dimensional probability simplex. Such a procedure would be computationally intensive, and impractical for large networks. The following result, however, proves the duality of this problem to information theoretic rate-distortion, which can then be used to obtain the optimal strategy efficiently and characterize the optimal throughput $R(\alpha)$ analytically.

Theorem 2: Let $d : 2^{\mathcal{P}} \times 2^{\mathcal{P}} \rightarrow \mathcal{R}$ s.t

$$d(\mathbf{S}, \hat{\mathbf{S}}) = \begin{cases} \Lambda^v(\mathbf{S}) - \Lambda^c(\mathbf{S}, \mathbf{B}) & \exists \mathbf{B} \text{ s.t. } \hat{\mathbf{S}} = T(\mathbf{S}, \mathbf{B}) \\ \infty & \text{o.w.} \end{cases} \tag{10}$$

Then, a throughput R is achievable with anonymity α if

$$R(0) - R(\alpha) \geq D\left(H(\mathbf{S})(1-\alpha)\right),$$

where $D(r)$ is the *Distortion-Rate* function defined as

$$D(r) = \min_{q(\hat{\mathbf{S}}|\mathbf{S}): I(\mathbf{S}; \hat{\mathbf{S}}) \leq r} \mathbb{E}(d(\mathbf{S}, \hat{\mathbf{S}})). \tag{11}$$

Proof: Refer to Appendix.

The above theorem characterizes $R(\alpha)$ using the single letter representation of a rate-distortion function. The loss function $d(\mathbf{S}, \hat{\mathbf{S}})$ in (10) represents the throughput reduction due to covert relaying. Although the loss function parameters do not explicitly include the set of covert relays $\mathbf{B}$, from Lemma 1 we know that given $\mathbf{S}, \hat{\mathbf{S}}$, the set of covert relays $\mathbf{B}$ is unique. Therefore, the distribution $q(\mathbf{B}|\mathbf{S})$ to chose covert relays is equivalent to the distortion minimizing distribution in (11). As a result, the Blahut-Arimoto algorithm [15] provides an efficient iterative technique to obtain $q(\mathbf{B}|\mathbf{S})$ and the achievable network throughput $R(\alpha)$. Note that the anonymity α is guaranteed assuming that the eavesdropper is aware of the network topology, the session prior distribution $p(\mathbf{S})$ and the optimal strategy $q(\mathbf{B}|\mathbf{S})$ of choosing covert relays.

The equivalence between anonymous networking and rate distortion is not tied to our strategy of choosing covert relays, as explained in Section I. In our model, the level of anonymity α directly corresponds to the rate of compression and the performance loss function plays the role of distortion. Therefore, obtaining the optimal rate-distortion function is equivalent to obtaining the throughput anonymity relation. We believe that the consequences of this duality extend beyond the characterization of the tradeoff between anonymity and throughput. Rate distortion is a field that has been studied for many decades [11], and the numerous models and techniques developed therein, could serve to design strategies for anonymous networking.

Note that in order to achieve the throughput of Theorem 2, it is necessary that every relay be aware of the entire session $\mathbf{S}$ and use an identical random seed. From a practical perspective, this could be achieved if nodes exchange local messages with their neighbours such that they reach a consensus about the session. Since total number of sessions is finite, perfect convergence can be reached in finite time, assuming no transmission errors. However, in network applications where message exchanges across nodes may not be possible, each node would only have partial information about the session. This is true of Mix networks where layered encryption ensures that each Mix only has knowledge of the neighbouring nodes in the routes. For such situations, a decentralized approach is presented in the following section.

V. Decentralized Approach

Let the information available to a node in any session be represented by function $l : \mathcal{V} \times \mathcal{S} \mapsto 2^{\mathcal{V} \times \mathcal{V}}$, where $l(A, \mathbf{S})$ represents the localized information of node A in session $\mathbf{S}$. If $\mathbf{S} = (P_1, \cdots, P_{|\mathbf{S}|})$ and $A(i,j)$ represent the i^{th} node of path P_j in $\mathbf{S}$, then,

$$l(B, \mathbf{S}) = \{(A(i, j-1), A(i, j+1)) : A(i,j) = B\}.$$

In other words, $l(B, \mathbf{S})$ is the set of node pairs $(A(i, j-1), A(i, j+1))$ such that node B relays packets from $A(i, j-1)$ to $A(i, j+1)$ on route P_i.

Since there are no message exchanges across nodes with regard to the session information, we require that each node makes a decision to be covert based on the local information function only. Further, we do not assume any common randomness available to the nodes, and hence, the decisions of multiple nodes are conditionally independent (conditioned on the session). Accordingly, we define a covert probability function

$$q_c : \mathcal{V} \times 2^{\mathcal{V} \times \mathcal{V}} \mapsto [0, 1],$$

where $q_c(B, l(A, \mathbf{S}))$ is the probability that node B is covert in session $\mathbf{S}$. Owing to conditional independence, the probability that nodes in a subset $\mathbf{B}$ are covert in session $\mathbf{S}$ is given by:

$$q(\mathbf{B}|\mathbf{S}) = \prod_{B \in \mathbf{B}} q_c(B, l(B, \mathbf{S})) \prod_{B \notin \mathbf{B}} (1 - q_c(B, l(B, \mathbf{S}))). \tag{12}$$

Let Q^* represent the set of all conditional probability mass functions $\{q(\mathbf{B}|\mathbf{S}), \mathbf{B} \in \mathcal{V}, \mathbf{S} \in \mathcal{S}\}$, such that there exists covert probability function $q_c(\cdot, \cdot)$ which satisfies (12) for every pair $\mathbf{B}, \mathbf{S}$. From Lemma 1, we know that the pairs of variables $(\mathbf{S}, \mathbf{B})$ and $(\mathbf{S}, \hat{\mathbf{S}})$ have a one-one correspondence. Therefore, Q^* corresponds to an equivalent set Q^{**} of conditional probabilities $\{q(\hat{\mathbf{S}}|\mathbf{S})\}$.

Theorem 3: A throughput $R(\alpha)$ that satisfies

$$R(0) - R(\alpha) \geq D'\left(H(S)(1-\alpha)\right),$$

is achievable with a decentralized strategy where

$$D'(r) = \min_{q(\hat{S}|S) \in Q^{**}: I(S; \hat{S}) \leq r} \mathbb{E}(d(S, \hat{S})). \tag{13}$$

Proof: Since the minimizing distribution $q(\hat{\mathbf{S}}|\mathbf{S})$ is an element of Q^{**}, it corresponds to a conditional distribution $q(\mathbf{B}|\mathbf{S})$ that is expressible in the form (12), which in turn provides the decentralized strategy through the covert probability function $q_c()$. The achievability of $R(\alpha)$ then follows from the proof of Theorem 1. □

Note that the minimization in (13) is over a subset of the probability simplex, and could therefore result in a lower throughput than that of Theorem 2. Even if $l(B, \mathbf{S})$ uniquely identifies the session for all $B, \mathbf{S}$, the throughput may not reach the optimal value of Theorem 1 owing to lack of common randomness. This is illustrated in the following example.

A. Example

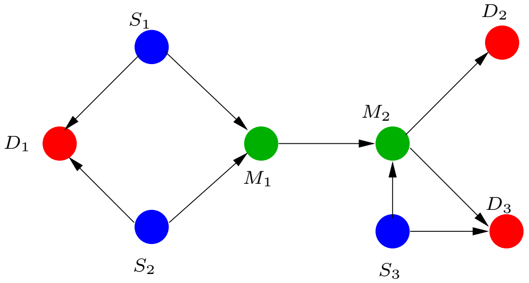

Fig. 4: Example Network: Sources: S_1, S_2, S_3 Destinations: D_1, D_2, D_3 Relays: M_1, M_2. S_1, S_2 can talk to all 3 destinations. S_3 can only talk to destinations S_2, S_3.

Consider the example of a network as shown in Figure 4. Sources S_1, S_2 can talk to all destinations, namely D_1, D_2, D_3 while source S_3 can only talk to destinations D_2, D_3. During any given session, each source picks a distinct destination. Therefore, there are 4 possible sessions in $\mathcal{S}$

$$\mathcal{S} = \begin{array}{c} [\{(S_1, D_1), (S_2, M_1, M_2, D_2), (S_3, D_3)\} \\ \{(S_1, D_1), (S_1, M_1, M_2, D_3), (S_3, M_2, D_2)\} \\ \{(S_1, M_1, M_2, D_2), (S_2, D_1), (S_3, D_3)\} \\ \{(S_1, M_1, M_2, D_3), (S_2, D_1), (S_3, M_2, D_2)\}] \end{array} .$$

For this example, Figure 5 plots the throughput versus anonymity, when the sessions are equally likely. The performance of the centralized approach is a convex function of anonymity, which is a result of the average nature of the metrics, namely equivocation and throughput. As can be seen, the decentralized strategy performs strictly worse than the centralized one. The losses due to local information and common randomness can be clearly observed in the Figure. Even when the relays are provided complete information about the session, the performance of the decentralized strategy is lower than the centralized throughput, particularly for high anonymity. This is because a low value of anonymity can be satisfied by making only one of the relays covert, in which case the lack of common randomness does not affect the performance.

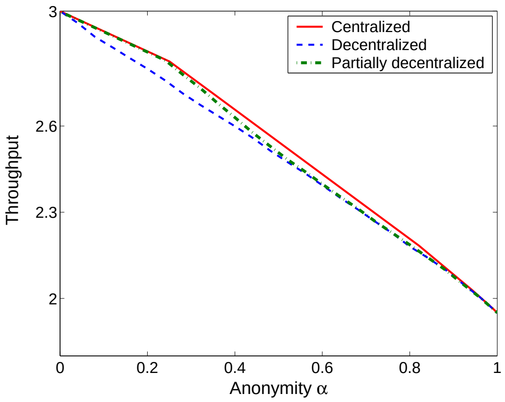

Fig. 5: Throughput vs Anonymity for example network with equally likely sessions. Partial decentralized strategy represents situation when relays have complete information on sessions but no common randomness.

VI. Conclusions and Future Directions

One of our key contributions in this work is the theoretical model for anonymity against traffic analysis. To the best of our knowledge, this is the first analytical metric designed to measure the secrecy of *routes* in an eavesdropped wireless network. Based on the metric, we designed scheduling and relaying strategies to maximize network performance with a guaranteed level of anonymity. Although we consider specific constraints on delay and bandwidth, the ideas of covert relaying and the randomized selection are quite general, and apply to arbitrary multihop wireless networks. The throughput-anonymity tradeoff we obtain reiterates the known paradigm of inverse relationship between communication rate and secrecy in covert channels.

In this work, we used throughput as an indicator of network performance, to optimize the selection strategy. However, the framework we establish extends beyond maximizing throughput. In fact, the loss function we define in (10) can be redefined to represent the loss in any convex function of the achievable relay rates. In our model, we fixed the packet delay and analyzed the loss in relay rates at a covert relay. Alternatively, we could fix the rates of transmission and analyze the increase in latency at every covert relay due to independent scheduling. By optimally designing the loss function to reflect the increase in overall network latency, we would be able to derive the relationship between latency and level of anonymity.

Our current model of independent sessions of observation may not apply to the scenario where an eavesdropper monitors the network for long periods of time. In that case, we would need a stochastic model to account for session changes, depending on when nodes start or stop communication. In this regard, if we adopt a Markovian model for the sessions, we believe that techniques in causal source coding [16] would apply to our problem as an extension of the proven duality.

References

[1] C. Gulcu and G. Tsudik, “Mixing e-mail with babel,” in *Proceedings of the Symposium on Network and Distributed System Security*, pp. 2–19, February 1996.

[2] M. K. Reiter and A. D. Rubin, “Crowds: anonymity for Web transactions,” *ACM Transactions on Information and System Security*, vol. 1, no. 1, pp. 66–92, 1998.

[3] G. Danezis, R. Dingledine, and N. Mathewson, “Mixminion: design of a type iii anonymous remailer protocol,” in *Proceedings of 2003 Symposium on Security and Privacy*, pp. 2–15, May 2003.

[4] D. Chaum, “Untraceable electronic mail, return addresses and digital pseudonyms,” *Communications of the ACM*, vol. 24, pp. 84–88, February 1981.

[5] Y. Zhu, X. Fu, B. Graham, R.Bettati, and W. Zhao, “On flow correlation attacks and countermeasures in mix networks,” in *Proceedings of Privacy Enhancing Technologies workshop*, May 26-28 2004.

[6] B.Radosavljevic and B. Hajek, “Hiding traffic flow in communication networks,” in *Military Communications Conference*, 1992.

[7] P. Venkitasubramaniam, T. He, and L. Tong, “Relay Secrecy in Wireless Networks with Eavesdroppers,” in *Proc. of 2006 Allerton Conference on Communication, Control and Computing*, (Monticello, IL), Sep. 2006.

[8] C. E. Shannon, “Communication theory of secrecy systems,” *Bell System Technical Journal*, 1949.

[9] A. Wyner, “The wiretap channel,” *Bell Syst. Tech. J.*, vol. 54, pp. 1355–1387, 1975.

[10] I. Csiszár and J. Korner, “Broadcast channels with confidential messages,” *IEEE Trans. on Information Theory*, vol. 24, pp. 339–348, May 1978.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[12] T. He, P. Venkitasubramaniam, and L. Tong, “Packet scheduling against stepping-stone attacks with chaff,” in *Proc. IEEE Military Communications Conference*, (Washington,DC), October 2006.

[13] P. Venkitasubramaniam, T. He, and L. Tong, “Anonymous Networking amidst Eavesdroppers,” *Submitted to IEEE Transactions on Information Theory: Special Issue on Information-Thoeretic Security*, Feb. 2007.

[14] P. Venkitasubramaniam, T. He, and L. Tong, “Networking with secrecy constraints,” in *Proc. of 2006 IEEE Military Communications Conference*, (Washington D.C), Oct. 2006.

[15] R. Blahut, “Computation of Channel Capacity and Rate-Distortion Functions,” *IEEE Trans. Infor. Theory*, vol. IT-18, July 1972.

[16] D. Neuhoff and L. Gilbert, “Causal Source Codes,” *IEEE Trans. on Information Theory*, vol. 28, pp. 701–713, Sep. 1982.

Appendix

A. Proof of Lemma 1

1. Let $\hat{\mathcal{Y}}$ be the schedules generated assuming $\hat{\mathbf{S}}$ were a session and none of the nodes were covert. The transmission rates of nodes in $\hat{\mathcal{Y}}$ are assumed identical to $\mathcal{Y}$. For the nodes that are the sources in $\mathbf{S}$, the schedules are independent in $\mathcal{Y}$ and $\hat{\mathcal{Y}}$. Session $\hat{\mathbf{S}}$ has additional sources due to the broken paths which also generate independent transmission schedules. The set of these additional sources is identical to the set of covert relays in $\mathbf{S}$. Therefore, the schedules are independent in $\mathcal{Y}$ as well. Since the remaining nodes relay all received packets within negligible processing delay, $q(\mathcal{Y}|\mathbf{S}) = q(\hat{\mathcal{Y}}|\mathbf{S})$. Therefore, using the data processing inequality ($\mathbf{S} - \hat{\mathbf{S}} - \hat{\mathcal{Y}}$)

$$H(\mathbf{S}|\mathcal{Y}) = H(\mathbf{S}|\hat{\mathcal{Y}}) \leq H(\mathbf{S}|\hat{\mathbf{S}}).$$

2. Suppose $\exists \mathbf{B}_1 \neq \mathbf{B}_2$ such that $\hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_1) = \hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_2)$. Then, we can write $\mathbf{B}_1 = (\mathbf{B}, \mathbf{B}_1'), \mathbf{B}_2 = (\mathbf{B}, \mathbf{B}_2')$ where $\mathbf{B}_1' = (B_{11}, \cdots, B_{1m})$, $\mathbf{B}_2' = (B_{21}, \cdots, B_{2n})$ and $\mathbf{B}_1' \bigcap \mathbf{B}_2' = \phi$. We know that

$$\begin{aligned}\hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_1) &= t(\cdots t(\mathbf{T}(\mathbf{S}, \mathbf{B}), B_{11}), \cdots), B_{1m}) \\ &= t(\cdots t(\mathbf{T}(\mathbf{S}, \mathbf{B}), B_{21}), \cdots), B_{2n}) = \hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_2).\end{aligned}$$

Suppose none of the paths in $\mathbf{T}(\mathbf{S}, \mathbf{B})$ contain $\mathbf{B}_1' \bigcup \mathbf{B}_2'$, then it does not matter if those relays are covert or not, in which case the subset of covert relays would be $\mathbf{B}$. If $\exists P \in \mathbf{T}(\mathbf{S}, \mathbf{B})$ that contains (w.l.o.g) B_{11}, then $\hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_1)$ would contain a path that ends in B_{11}, whereas $\hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_2)$ cannot contain such a path. Therefore, we have a contradiction. □

B. Proof of Theorem 2

Consider the optimal solution $q^*(\hat{\mathbf{S}}|\mathbf{S})$ of the distortion rate problem,

$$D = \min_{q(\hat{\mathbf{S}}|\mathbf{S}): I(\mathbf{S};\hat{\mathbf{S}}) \leq (1-\alpha)H(\mathbf{S})} \mathbb{E}(d(\mathbf{S}, \hat{\mathbf{S}})).$$

From the definition of $d(\mathbf{S}, \hat{\mathbf{S}})$, it is easy to see that if $\nexists \mathbf{B}\ s.t.\ \hat{\mathbf{S}} \neq \mathbf{T}(\mathbf{S}, \mathbf{B})$, then $q^*(\hat{\mathbf{S}}|\mathbf{S}) = 0$. Given $\mathbf{S}, \hat{\mathbf{S}}$, we know from Lemma 1 that the set of covert relays $\mathbf{B}$ are uniquely determined. In other words, we can equivalently write $q^*(\hat{\mathbf{S}}|\mathbf{S}) = q^*(\mathbf{B}|\mathbf{S})$. Therefore, q^* specifies a valid selection strategy. Since $H(\mathbf{S})$ is fixed apriori, $I(\mathbf{S}; \hat{\mathbf{S}}) \leq (1-\alpha)H(\mathbf{S})$ ensures that an anonymity α is guaranteed. Further, for every $\mathbf{B}$, the function $d(\mathbf{S}, \hat{\mathbf{S}})$ evaluates the difference in achievable sum-rates Λ^0 and $\Lambda^c(\mathbf{S}, B)$. Taking expectation over $q^*(\mathbf{B}|\mathbf{S})$, it is easy to see that the distortion D is achievable with anonymity α. □