Anonymous Networking Amidst Eavesdroppers

Parvathinathan Venkitasubramaniam, *Member, IEEE*, Ting He, *Member, IEEE*, and Lang Tong, *Fellow, IEEE*

***Abstract*—The problem of security against packet timing based traffic analysis in wireless networks is considered in this work. An analytical measure of "anonymity" of routes in eavesdropped networks is proposed using the information-theoretic equivocation. For a physical layer with orthogonal transmitter directed signaling, scheduling and relaying techniques are designed to maximize achievable network performance for any desired level of anonymity. The network performance is measured by the total rate of packets delivered from the sources to destinations under strict latency and medium access constraints. In particular, analytical results are presented for two scenarios:**

For a single relay that forwards packets from m users, relaying strategies are provided that minimize the packet drops when the source nodes and the relay generate independent transmission schedules. A relay using such an independent scheduling strategy is undetectable by an eavesdropper and is referred to as a covert relay. Achievable rate regions are characterized under strict and average delay constraints on the traffic, when schedules are independent Poisson processes.

For a multihop network with an arbitrary anonymity requirement, the problem of maximizing the sum-rate of flows (network throughput) is considered. A randomized selection strategy to choose covert relays as a function of the routes is designed for this purpose. Using the analytical results for a single covert relay, the strategy is optimized to obtain the maximum achievable throughput as a function of the desired level of anonymity. In particular, the throughput–anonymity relation for the proposed strategy is shown to be equivalent to an information-theoretic rate–distortion function.

***Index Terms*—Anonymity, equivocation, network security, rate–distortion, traffic analysis.**

I. Introduction

EAVESDROPPERS monitoring transmissions in a network can deduce source–destination pairs and paths of data flow by analyzing the timing information in the network traffic. Traffic analysis has played a prominent role in modern warfare [1] and its potential to compromise privacy in Internet communication is well documented in literature [2]–[5]. For example, the weaknesses of protocols for web browsing [4], [6] and SSH [7] have been exposed through traffic analysis.

The primary focus of this work is an analytical approach to security against traffic analysis in wireless networks, and in particular, the design of provably secure countermeasures. Cryptographic techniques can be used to prevent analysis of packet contents (see Section I-B), however, significant information can be inferred by analyzing the correlation across packet transmission schedules of multiple nodes. Furthermore, in wireless networks, due to the open medium, packet timing information is easy to obtain. In this work, we consider the problem of designing scheduling strategies for wireless nodes to prevenet the inference of network routes from packet timing based inference of network routes.

The challenge in designing transmission schedules that hide *networking information* is to minimize the effects on network performance. Wireless networks are subject to constraints on medium access, latency, and stability, which generally result in a high correlation across transmission schedules of nodes in a route. The need for anonymity, however, necessitates that routes are not detectable using the correlation across transmission schedules. These contrasting paradigms result in a tradeoff between anonymity and network performance. For example, consider a simple two-hop setup shown in Fig. 1, where node B relays packets received from nodes S_1 and S_2 subject to a strict delay constraint. Assuming the nodes transmit on orthogonal channels, let the maximum transmission rates allowed by each node be independently bounded. Then the set of rate pairs R_1, R_2 of packets that can be relayed successfully from S_1, S_2 respectively is given by a pentagon (see Fig. 1). Any rate-pair in this region is achieved if the relay transmits each received packet immediately upon reception (assuming processing delays are negligible). It is easy to see that such a strategy would result in a high correlation between the transmission schedules of the sources and the relay. If, in addition to the networking constraints, the relay is forced to transmit packets according to a schedule that is statistically independent of the arrival processes, correlation across schedules no longer provides any information, thus hiding the relaying operation. However, the strict delay constraint would result in packets being dropped or require additional *dummy packet* transmissions by the relay, thereby reducing the achievable relay rates.

The relaying operation of Fig. 1 represents the basic component in wireless networking, and the characterization of achievable rate regions with provable anonymity (of transmission schedules) is one of the contributions of this work. The example highlights that providing anonymity incurs a loss in communication rates. A primary goal of this work is to characterize this tradeoff between anonymity and performance for general multihop networks. For this purpose, we define a quantifiable metric of anonymity using the uncertainty in net-

Manuscript received February 16, 2007; revised October 1, 2007. 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. The material in this paper was presented in part at the 44th and 45th Annual Conferences on Communication, Control and Computing (Allerton 2006 and Allerton 2007), Monticello, IL.

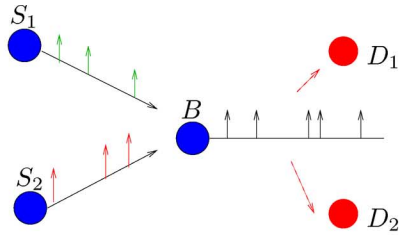
P. Venkitasubramaniam and L. Tong are with the Department of Electrical and Computer Engineering, Cornell University, Ithaca NY 14853 USA (e-mail: pv45@cornell.edu; lt35@cornell.edu).

T. He is with IBM T. J. Watson Research Center, Hawthorne, NY 14853 USA (e-mail: th255@cornell.edu)

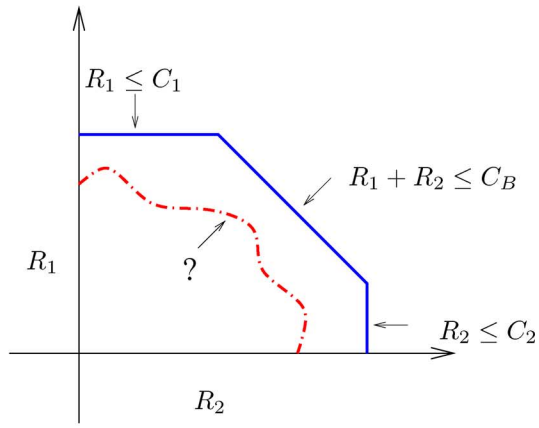
Communicated by Y. Zheng, Guest Editor for Special Issue on Information Theoretic Security.

Color versions of Figures 1, 3–7, and 11 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.921660

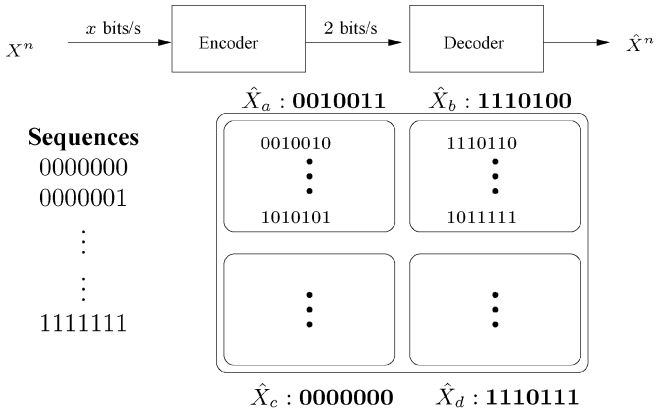


(a) Sources S_1, S_2 transmit packets to Destinations D_1, D_2 through Relay B

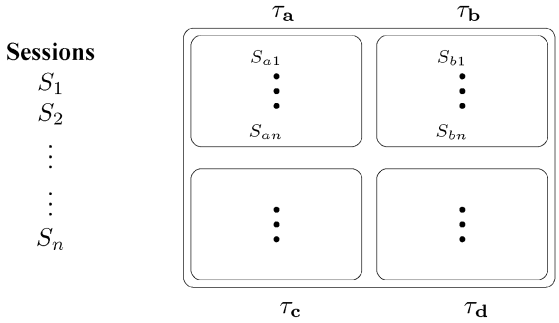


(b) Achievable Rate Region : The horizontal and vertical boundaries are due to rate constraints C_1, C_2 on nodes S_1, S_2. The sum-rate constraint is due to stability at relay B. The inner region (in dotted line) represents the achievable rate region with independent scheduling that we wish to characterize.

Fig. 1. Two-hop relay network.



(a) Rate-Distortion: Any sequence in partition a corresponds to reconstruction sequence $\hat{X}_a$. Reconstruction sequences $\hat{X}$ are chosen to minimize distortion within corresponding partitions



(b) Anonymous Networking: For any network session in partition a, eavesdropper observes τ_a, and cannot decide from $S_{a1} \cdots S_{an}$. $\mathbf{Y}$s are designed to minimize performance loss within corresponding partitions.

Fig. 2. Connection between rate distortion and anonymous networking.

working information (set of active routes in the network) from the perspective of the eavesdropper. Although the example suggests a simple technique to provide maximum anonymity by making the schedules of all nodes statistically independent, the reduction in relay rates at each node could significantly affect the overall network throughput. Our goal is to design transmission strategies that sacrifice minimum network performance while guaranteeing the desired level of anonymity.

A. Main Contributions

We consider a wireless network, where the nodes use orthogonal transmitter directed signaling, and every transmitted packet is subjected to a strict delay constraint at every intermediate relay. Under this model, we define a mathematical notion for anonymity of network routes using Shannon's equivocation [8], assuming a global passive eavesdropper who observes transmission schedules of all nodes in the network. The problem we address is the design of transmission schedules for relay nodes to maximize throughput given a desired level of anonymity. Specifically, we divide relays into two categories: *covert* and *visible*. A visible relay merely forwards packets immediately upon reception, whereas a covert relay transmits packets according to an independently generated transmission schedule. Our key contributions in this regard are divided into two segments: design of covert relaying strategies and the selection of covert relays depending on the network routes.

For a covert relay, we design relaying strategies to minimize the loss in achievable rates due to independence in transmission schedules. Specifically, when the transmission schedules of source nodes and the relay are independent Poisson processes, we characterize the achievable rate region analytically. Although independent Poisson schedules may not be optimal under strict delay constraints, we show that, under certain physical layer conditions, the achievable relay rates are optimal under an average delay constraint.

For the general multihop network, we propose a randomized strategy to choose the set of relays to be covert, given any desired level of anonymity α. Utilizing the results for a single covert relay, we optimize the selection of covert relays in each route, and characterize the resulting network throughput as a function of α. Our key result in this regard shows the equivalence between the throughput–anonymity tradeoff and information-theoretic rate distortion.

The connection between rate distortion and anonymous networking is not tied to our strategy of covert relaying, and can be explained using a general intuition (see Fig. 2). 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 goal is to minimize distortion for a fixed compression rate, then the total number of bins are predetermined. The problem then translates to dividing the sequences optimally into the bins 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 the 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 for the sessions within the bin.

B. Related Work

Prevention of traffic analysis is a classical problem in computer networks, and a dominant portion of prior research has centered around Internet applications. In that regard, an important countermeasure was provided by Chaum through the concept of the traffic mix [9]. A mix is a node or server that collects packets from multiple users and outputs them in a manner that makes it infeasible to correlate an outgoing packet with a unique incoming packet. Specifically, a mix performs re-encryption and packet padding to obfuscate the contents of each packet. It also changes the timing pattern of arrived packets by reordering and batching packets from multiple users together. Subsequent improvements to mixing strategies include maintaining a dynamic pool of packets [10], and random delaying [11].

Mixes have been widely used in designing remailer and proxy systems [12], [13] for the Internet. However, when strict constraints on delay or buffer size are imposed on the traffic, it was shown [14] that known mixing strategies no longer provided anonymity to long streams of traffic. An alternative approach, designed primarily for multihop wireless networks is that of fixed scheduling [15]. In [15], the key idea was that every node transmitted according to a fixed predetermined schedule by transmitting dummy packets whenever data packets were unavailable. Although fixed scheduling prevents any retrieval of information, the strategy results in a large percentage of dummy packets and furthermore, the need for centralized synchronous implementation makes it impractical in *ad hoc* wireless networks.

A key component of our approach is the analytical model for anonymity of routes. In mix networks, anonymity has been measured using the size or entropy of the anonymity set (set of possible source–destination pairs) of an observed packet. In the context of this work, the use of anonymity sets has two disadvantages. First, hiding source–destination pairs alone may not be sufficient, the routes of data flow could also reveal critical information. Second, we require a measure of anonymity that considers continuous streams of packets rather than treat each packet independently [14]. The information-theoretic metric we propose is based on equivocation, which has primarily been used to measure the secrecy of transmitted messages on point-to-point channels [16], [17]. A common theme in [16], [17] and in many subsequent results is the tradeoff between communication rate and level of secrecy, which is also exhibited in our results, albeit from an anonymity perspective.

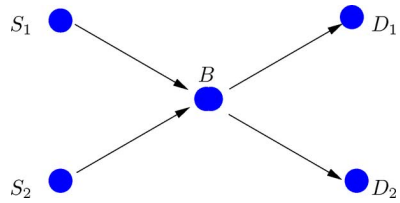


Fig. 3. Two node switching network: $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E})$, $\mathcal{V}_1 = \{S_1, S_2, B, D_1, D_2\}$, $\mathcal{E}_1 = \{(S_1, B), (S_2, B), (B, D_1), (B, D_2)\}$.

Prevention of traffic analysis can be viewed as the complementary problem to intrusion detection [18], which is another important area in network security. Some of the techniques we use to design covert relaying strategies are motivated by prior work on stepping stone detection [19].

II. Problem Setup

A. System Model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph, 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 $P = (V_1, \ldots, V_n)$ is a *valid path* in $\mathcal{G}$ if $(V_i, V_{i+1}) \in \mathcal{E}, \forall i < n$. Let the set of all possible paths in $\mathcal{G}$ be 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. This set of paths $\boldsymbol{S} \in 2^{\mathcal{P}(\mathcal{G})}$ is referred to as a network *session*. The information that we wish to hide from the eavesdropper is the network session $\boldsymbol{S}$. We model $\boldsymbol{S}$ as an independent and identically distributed (i.i.d.) random variable with a probability mass function (pmf)$\{p(\boldsymbol{s}) : \boldsymbol{s} \in 2^{\mathcal{P}(\mathcal{G})}\}$. Therefore, the set of all possible sessions is given by

$$\mathcal{S}(\mathcal{G}) = \{\boldsymbol{s} \in 2^{\mathcal{P}(\mathcal{G})} : p(\boldsymbol{s}) > 0\}.$$

The prior information $p(\boldsymbol{S})$ on sessions, which is obtained using the topology and applications of the particular network, is also available to the eavesdropper.

For the example network $\mathcal{G}_1$ in Fig. 3, let S_1, S_2 be the only allowed sources and D_1, D_2 the allowed destinations. For this network, $\mathcal{P}(\mathcal{G}_1)$, the set of all possible paths, is given by

$$\mathcal{P}(\mathcal{G}_1) = \{(S_1, B), (S_1, B, D_1), (S_1, B, D_2), (S_2, B), (S_2, B, D_1), (S_2, B, D_2), (B, D_1), (B, D_2)\}.$$

If we impose a restriction that the sources always communicate with distinct destinations, then $\mathcal{S}(\mathcal{G}_1)$ contains only two sessions

$$\mathcal{S} = \Big\{\{(S_1, B, D_1), (S_2, B, D_2)\}\, \{(S_1, B, D_2), (S_2, B, D_1)\}\Big\}.$$

Transmission Schedules: The eavesdropper's observation consists of the packet transmission epochs[1] in a session. We consider a global passive eavesdropper who monitors transmissions at all parts of the network. The packet headers are assumed to be encrypted, and hence the contents of transmitted packets do not

[1]"Transmission epoch" in this work refers to the time instant of transmission of a packet. Transmission delays are assumed negligible.

reveal the identity of the transmitting or receiving nodes. However, the physical layer signaling strategy we consider provides the eavesdropper information about the transmitting nodes to the eavesdropper.

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" her detector to a particular spreading sequence and identify the transmission times of packets sent by the corresponding node. Although she knows the transmitting node of each packet, since headers are encrypted, she would not know the intended recipient of any packet. Therefore, in a route involving multiple nodes, even when all transmission schedules are correlated, it may not be possible for an eavesdropper to determine the final destination node.

Eavesdropper Observation: Let $\boldsymbol{\tau}_A$ represent the schedule of packets transmitted by node A. The schedule $\boldsymbol{\tau}_A$ is a sequence of transmission epochs

$$\boldsymbol{\tau}_A = \{T_A(1), T_A(2), \ldots\}$$

where $T_A(i)$ represents the transmission epoch of the ith packet transmitted by node A. By virtue of unique orthogonal codes, the eavesdropper can obtain the transmission schedule of each individual node. The eavesdropper's complete observation is therefore given by $\boldsymbol{\tau} = \{\boldsymbol{\tau}_A : A \in \mathcal{V}\}$. Note that, while $\boldsymbol{\tau}$ represents the schedules of packet transmissions detected by eavesdroppers, it does not specify which packets are relayed from sources to destinations in a session. In fact, some of the epochs in $\boldsymbol{\tau}$ could represent dummy transmissions by nodes.

B. Anonymity Measure

We model $\boldsymbol{\tau}$ as a set of random sequences of epochs with conditional distribution $q(\boldsymbol{\tau}|\boldsymbol{S})$, and use equivocation [8] to define the measure of anonymity. The idea is to design $q(\boldsymbol{\tau}|\boldsymbol{S})$ such that eavesdroppers obtain minimum information about the session $\boldsymbol{S}$ by observing $\boldsymbol{\tau}$.

Definition 1: A distribution $q(\boldsymbol{\tau}|\boldsymbol{S})$ is defined to have *anonymity* α if

$$\frac{H(\boldsymbol{S}|\boldsymbol{\tau})}{H(\boldsymbol{S})} \geq \alpha.$$

When $\alpha = 1$, the distribution $q(\boldsymbol{\tau}|\boldsymbol{S})$ is defined to have *perfect anonymity*. For a distribution with perfect anonymity

$$H(\boldsymbol{S}|\boldsymbol{\tau}) = H(\boldsymbol{S}).$$

In other words, the observed schedules cannot provide the eavesdropper any additional information about the routes than the prior $p(\boldsymbol{S})$.

For a general α, the physical interpretation of anonymity is provided by Fano's inequality [20]: If the error probability of the eavesdropper in decoding the session $\boldsymbol{S}$ is P_e, then

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

In other words, the anonymity provides a lower bound to the probability of error incurred by the eavesdropper in decoding $\boldsymbol{S}$. This notion of anonymity that we consider is different from previous definitions [11], [21], which primarily measured the uncertainty of the source–destination pairs of each individual packet. To the best of our knowledge, this is the first definition of anonymity that deals with multihop routes and considers the timing information in long streams of transmitted packets.

C. Network Constraints and Throughput

The key challenge in designing the schedule distribution $q(\boldsymbol{\tau}|\boldsymbol{S})$ with provable anonymity is to sacrifice minimum performance under the networking constraints. In this work, we measure performance using the total rate of packets delivered from sources to destinations per session subject to the following constraints on medium access and latency.

Medium Access Constraints: The shared medium in wireless network imposes constraints on the maximum rate of packets that can be transmitted by the nodes. Since we consider long streams of packet transmissions, we measure the rate of packets transmitted using the following asymptotic measure:

$$\lambda_A = \lim \inf_{n\to\infty} \frac{n}{T_A(n)} \tag{1}$$

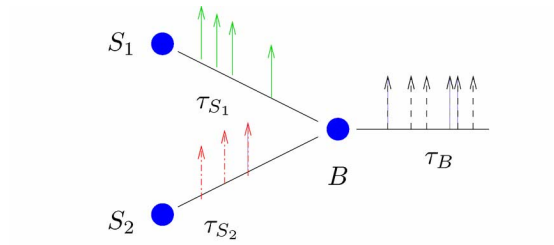
where λ_A denotes the rate of packets transmitted by a node A.

Since each transmitting node is associated with an orthogonal spreading sequence, the constraint on transmission rate of each node is independent. Specifically, we assume the transmission rate λ_A of a node A is upper-bounded by a constant C_A, which depends on the characteristics of the medium and the transmission capability of node A.

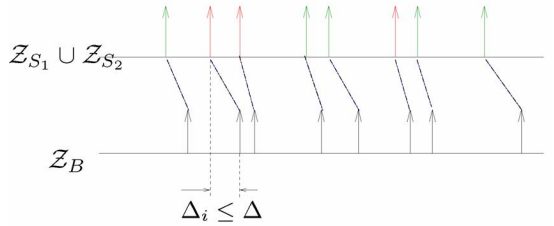
We assume that the network operates in full duplex mode, where every node can transmit and receive packets simultaneously as long as all transmission rates satisfy the specified bounds. In other words, a set of schedules $\boldsymbol{\tau}$ is a valid network schedule if and only if $\lambda_A \leq C_A$ for every node A.

Latency Constraint: We consider a strict delay constraint on the packets, where the packet delay at each intermediate relay in a route is bounded by Δ. In general, each relay is allowed to re-encrypt packets, reorder arrived packets, and transmit dummy packets. However, each received data packet at a relay is required to be forwarded within Δ time units of arrival, or otherwise, dropped. Such a strict delay constraint would apply in practice to time-sensitive applications such as target tracking in sensor networks or streaming media in peer-to-peer networks. In general, a strict delay constraint would prevent congestion in the network and ensure stability, albeit at the cost of dropped packets.

The transmission schedule $\boldsymbol{\tau}$ only specifies when packets are transmitted by each node, and do not indicate which packets actually travel from source to destination on each route in a session. For every schedule $\boldsymbol{\tau}$, we therefore specify a relaying strategy, represented by $\mathcal{Z}$, which is a set of subsequences of $\boldsymbol{\tau}$. The subsequences represent the transmissions epochs of packets that are relayed from sources to destinations and therefore, depend on the routes of the session. we define the validity of a relaying strategy under the delay constraints as follows.



(a) τ_X is the transmission schedule of Node X.



(b) Every packet in the set of subsequences satisfy the strict delay constraint of Δ

Fig. 4. 2×1 relay with strict delay constraint.

Definition 2: Let a session[2] $\boldsymbol{S} = (P(1), \ldots, P(|\boldsymbol{S}|))$, where $P(i) = (A(i,1), \ldots, A(i,|P(i)|))$ is a valid path, and $A(i,j) \in \mathcal{V}$ represents the jth node in path $P(i)$ of session $\boldsymbol{S}$. A set of sequences $\mathcal{Z} = \{\mathcal{Z}_{i,j} : i \leq |\boldsymbol{S}|, j < |P(i)|\}$ is a valid relaying strategy given the schedule $\boldsymbol{\tau}$ and session $\boldsymbol{S}$ iff:

1. $\forall i,j$: $\mathcal{Z}_{i,j} \subseteq \boldsymbol{\tau}_{A(i,j)}$.
2. $\forall i,j,n : 0 \leq Z_{i,j+1}(n) - Z_{i,j}(n) \leq \Delta$.
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 preceding definition, condition 2 ensures that the relayed packets satisfy the delay constraint Δ (see Fig. 4) at every intermediate relay from the sources to the destinations of the session. Condition 3 ensures that, if any pair of nodes is common to multiple routes, the subsequences picked from the transmission schedules are mutually exclusive.

In Section IV-B, we also consider a relaxed version of the delay constraint, where the average delay of packets is bounded at each relay. The corresponding definition of a relaying strategy with average delay constraint is obtained by replacing condition 2 in Definition 2 with the following conditions:

$$\forall i,j,n : Z_{i,j+1}(n) - Z_{i,j}(n) \geq 0 \tag{2}$$

$$\limsup_{n\to\infty} \sum_{m=1}^{n} \frac{Z_{i,j+1}(m) - Z_{i,j}(m)}{n} \leq \bar{\boldsymbol{\Delta}}. \tag{3}$$

It is possible that the set of subsequences $\mathcal{Z}$ are a strict subset of the transmissions schedule $\boldsymbol{\tau}$, or in other words, there are epochs in $\boldsymbol{\tau}$ that do not correspond to any relayed packets. Those transmission epochs in $\boldsymbol{\tau}$ that are not present in $\mathcal{Z}$ would either correspond to packets that are dropped, or represent dummy packet transmissions. Therefore, for a session $\boldsymbol{s} = (P(1), \ldots, P(|\boldsymbol{s}|))$ and relaying schedule $\mathcal{Z}$, the rate of packets relayed from source to destination on route $P(i)$ is given by

$$\lambda_r(\mathcal{Z}, P(i)) = \lim_{n\to\infty} \frac{n}{Z_{i,1}(n)}.$$

Note that, since condition 2 of Definition 2 ensures that all schedules on a route have the same length, it is sufficient to use $\mathcal{Z}_{i,1}$ to compute rate.

D. Performance Metric

For a large network with several possible session, characterization of the set of rates achievable on every path of a session is potentially cumbersome. In order to draw useful inferences on the relationship between anonymity and network performance, we utilize a figure of merit that quantifies overall network performance. Specifically, we consider network throughput as a measure of performance, which is defined as follows.

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

1. for every session $\boldsymbol{s} = \{P(1), \ldots, P(|\boldsymbol{s}|)\}$, every realization of $\boldsymbol{\tau}$ given $\boldsymbol{s}$ is a valid network schedule;
2. for every realization of $(\boldsymbol{S}, \boldsymbol{\tau})$, there exists a valid relaying strategy $\mathcal{Z}$, and
$$\mathbb{E}\left(\sum_{i=1}^{|\boldsymbol{S}|} \lambda_r(\mathcal{Z}, P(i))\right) \geq R \tag{4}$$
where the expectation is over the joint probability density function (pdf) of $\boldsymbol{\tau}$ and $\boldsymbol{S}$.

Note that the throughput as defined above merely represents the rate of packets successfully relayed from sources to destinations. Since the relaying strategy could result in packet drops en route to the destinations, the reliable information rate delivered depends on the specific packet encoding and decoding techniques. We address the issue of forward error correction for reliability in Section IV-C.

III. Covert Relaying

Our approach to designing schedules and relay strategies derives its motivation from mix networks [9] on the Internet, where packets from each source travel through a sequence of special proxy servers or routers called mixes before reaching the destination. A mix collects packets from multiple sources and transmits them in batches, so an eavesdropper would not be able to correlate incoming and outgoing packets at the mix. The batching strategies of mixes, primarily designed for Internet applications such as e-mail and browsing, are, however, not suited to handle tight delay constraints, and their anonymity is compromised when the sources transmit long streams [14].

To provide anonymity for long streams of packets in wireless networks, we adopt the following approach. For every session, we assign a subset of intermediate relays in the routes to generate packet transmission schedules that are statistically independent of the packet arrival schedules to those relays. The remaining relays forward packets depending on the arrival times. In effect, each relay node in a session operates in one of two

[2]The notation $|\cdot|$ refers to number of paths in a session or number of nodes in a path depending on the variable used.

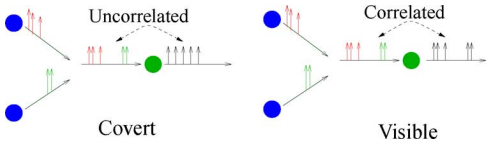


Fig. 5. Visible and covert relaying.

transmission modes (see Fig. 5), *covert* and *visible*, which are defined more precisely as follows.

A. 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. The independence in transmission schedules, owing to strict delay constraints, would result in dropped packets or require dummy packet transmissions. Therefore, the relaying strategy ($\mathcal{Z}$) needs to be optimally designed to minimize packet loss.

An external eavesdropper, who only observes transmission schedules, cannot correlate successive streams (from the previous node to the relay and from the relay to the subsequent node) in the path, and therefore, would not be able to detect the relaying operation of a covert relay.

B. 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 a transmission epoch immediately following the packet arrival (processing delay is assumed 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.[3] 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, since the independent schedules would result in packet loss at every covert relay, network throughput would be reduced significantly. It is, therefore, necessary to choose the covert relays optimally so that anonymity is guaranteed with minimum loss in throughput.

As an example, consider the switching network shown in Fig. 6. Let the maximum transmission rate allowed for each node be C. During any network session, each source S_i transmits packets to a distinct destination D_j, and for each pair S_i, D_j there is a fixed path through the intermediate relays. The set of possible sessions $\mathcal{S}$, therefore, contains 24 elements (all possible pairings) which are assumed equiprobable.

If all relays were visible, then the achievable throughput would be $2C$ (min-cut would be M_1, M_3). Since the transmission schedules of all the relays $(M_1, \ldots, M_4)$ are dependent on the arrival schedules from the sources $(S_1, \ldots, S_4)$, the eavesdropper would be able to determine the paths of flow until, but not including, the destination nodes (by virtue of

[3]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.

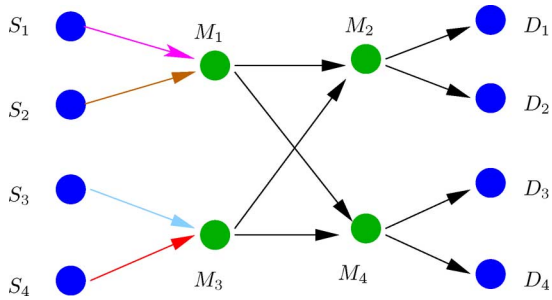


Fig. 6. Switching network: Sources $\{S_i\}$ transmit packets to destinations $\{D_i\}$ through relays $\{M_i\}$. The arrows represent the links in $\mathcal{E}$.

transmitter directed signaling). In this case, it can be shown that the level of anonymity is $\frac{H(\boldsymbol{S}|\boldsymbol{\tau})}{H(\boldsymbol{S})} = 0.436$.

Alternatively, we could let all relays be covert. Such a strategy would provide maximum anonymity $\frac{H(\boldsymbol{S}|\boldsymbol{\tau})}{H(\boldsymbol{S})} = 1$, but would reduce the achievable throughput due to packet drops. Let the fraction of packets dropped at the relays $M_1, \ldots, M_4$ due to the independent schedules be given by $\epsilon_1, \ldots, \epsilon_4$, respectively. The value of ϵ_i depends on the schedules $\boldsymbol{\tau}$ and the designed relaying strategy $\mathcal{Z}$. Due to symmetry, we assume $\epsilon_1 = \epsilon_3$, $\epsilon_2 = \epsilon_4$. Since every path contains two covert relays and all sessions are equally likely, the achievable network throughput is $2C(1-\epsilon_1)(1-\epsilon_2)$.

Suppose, only M_1, M_3 were to be covert, while M_2, M_4 were visible. Then the eavesdropper would be able to observe a portion of the paths, and make an intelligent guess on the actual session $\boldsymbol{S}$. Under such a scenario, it can be shown that the anonymity level $\frac{H(\boldsymbol{S}|\boldsymbol{\tau})}{H(\boldsymbol{S})} = 0.65$. However, since there is only one covert relay in every path, the achievable throughput is increased (from the maximum anonymity strategy) to $2C(1-\epsilon_1)$.

From the example, it is clear that, depending on the level of anonymity required, we need to choose covert relays optimally so that network throughput is maximized. The optimal selection strategy to choose covert relays will be explained in Section V-A. In the remainder of this section, we consider a single covert relay, design relaying strategies for independent schedules, and characterize the minimum packet loss incurred at the covert relay (ϵ_i), as a function of the delay and medium access constraints.

IV. Achievable Rate Regions

One approach to generate independent schedules at a covert relay would be to derive a queuing discipline that forwards packets within the required delay constraints, and yet maintain a statistically independent outgoing schedule. Such a strategy would, however, be vulnerable to active inference methods such as insertion of packets. We, therefore, propose an independent scheduling technique, wherein each node in the network generates a random transmission schedule, statistically independent of the session, and the schedules of other nodes in the network.

Independent scheduling is a particular solution to designing a covert relay schedule. An alternative to independent scheduling would be the fixed scheduling as described in [15]. Under that model, each relay node follows a deterministic schedule irrespective of transmitted data rates or paths of information

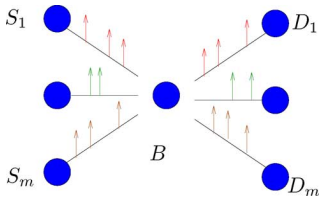


Fig. 7. Multiaccess relay: Source S_i transmits packets to D_i through B.

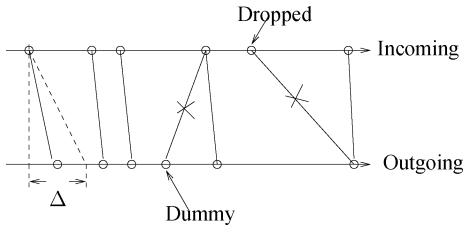


Fig. 8. Bounded Greedy Match: Unmatched packets are dropped, unused epochs have dummy packets.

flow. While the fixed scheduling strategy guarantees maximum anonymity, it would result in a large percentage of dummy packets. Further, a fixed schedule requires a centralized synchronous implementation, which is impractical in large networks.

The relaying algorithms discussed in this section are not specific to the statistics of the particular transmission processes and some of the optimal properties hold for any pair of point processes. However, for the purpose of analytical characterization of relay rates, we have assumed the source transmission schedules to belong to independent Poisson point processes. Poisson processes have typically been used to model the arrival of packets to nodes in a network, due to memoryless interarrival times property. Although independent Poisson schedules for the relay have not been proven optimal under strict delay constraints, under certain conditions on the physical layer they are shown to be optimal for an average delay constraint. Our relaying algorithms can be used on other point processes, such as Pareto-distributed schedules, however, the analytical tractability of achievable rates is not guaranteed.

A. Scheduling Under Strict Delay

Consider the special case of a single source relay (Fig. 7, $m = 1$). We are interested in the achievable relay rate for the route (S_1, B, D_1). The medium access constraints are specified by the bounds $\lambda_{S_1} \leq C_{S_1}$, $\lambda_B \leq C_B$ on the transmission rates. If the delay constraint were absent ($\Delta = \infty$), then each received packet could be relayed by B at the next available epoch in its transmission schedule. Since packets can be held for an indefinitely long time, the achievable relay rate would be $\lambda_r(\mathcal{Z}, (S_1, B, D_1)) = \min\{C_{S_1}, C_B\}$. Note that this is also the maximum possible rate if node B were a visible relay.

When a strict delay constraint of Δ is imposed, we design the relaying strategy using the Bounded Greedy Match (BGM) algorithm proposed in [22] under the context of chaff insertion in stepping stone attacks. The algorithm (Fig. 8) is described in Table I. The basic idea is as follows: When a packet arrives at B, if there exists a departure epoch within Δ of the arrival instant and has not been matched to any previous arrival, it is assigned to the arrived packet. Otherwise, the packet is dropped. If a relay epoch does not have any packet assigned to it, the relay transmits a dummy packet at that epoch.

It was shown in [22] that this greedy algorithm resulted in least packet drops. Based on the algorithm, the following theorem characterizes the best achievable relay rate, when the source node and relay use independent Poisson schedules.

Theorem 1: If the nodes S_1 and B generate independent Poisson transmission schedules, the maximum achievable relay rate from S_1 to D_1 through B is given by $\lambda_r(\mathcal{Z}, (S_1, B, D_1)) = C_{S_1}(1 - \epsilon(S_1, B))$ where

$$\epsilon(S_1, B) = \begin{cases} \frac{C_B - C_{S_1}}{C_B e^{-\Delta(C_{S_1} - C_B)} - C_{S_1}}, & C_{S_1} \neq C_B \\ \frac{1}{1 + C_{S_1}\Delta}, & C_{S_1} = C_B \end{cases} \triangleq f_e(\Delta, C_{S_1}, C_B). \quad (5)$$

Proof: Refer to the Appendix .

TABLE I
BOUNDED GREEDY MATCH ALGORITHM

Let $T_{S_1}(n), T_B(n)$ represent the arrival time of the n^{th} packet from S_1 and departure time of n^{th} packet from B.
1. Initialize $i = 1, j = 1$.
2. Let $t = \min\{T_{S_1}(i), T_B(j)\}$.
3. If $t = T_B(j)$, then
   i. B transmits a dummy packet at time $T_B(j)$.
   ii. $j = j + 1$.
   else if $T_B(j) - T_{S_1}(i) \leq \Delta$
   i. B transmits the i^{th} packet from S_1 at $T_B(j)$.
   ii. $i = i + 1, j = j + 1$.
   else
   i. Drop the i^{th} packet that arrived from S_1.
   ii. $i = i + 1$.
4. Repeat Step 2,3 until the end of the streams.

Theorem 1 expresses the maximum achievable rate in terms of the fraction of packets dropped $\epsilon(S_1, B)$ where $\epsilon(S_1, B)$ represents the fraction of packets dropped at relay B. As the delay constraint Δ increases, it is easy to see that the relay rate converges to $\min\{C_{S_1}, C_B\}$ which is the maximum achievable rate for a visible relay. Furthermore, the convergence of the relay rate to the optimal value is exponential in Δ. The value of $\epsilon(S_1, B)$ given in Theorem 1 is obtained when S_1 uses the maximum transmission rate of C_{S_1} for this particular route. In a general network, S_1 could be simultaneously transmitting to another node, in which case, the rate allocated for the (S_1, B, D_1) route would be strictly less than C_{S_1}. In such a situation, by replacing C_{S_1} in (5) with the allocated rate for the particular route, we can use Theorem 1 to evaluate the corresponding relay rate.

$m \times 1$ *Relay:* Consider the general $m \times 1$ relay as shown in Fig. 7. Using a slight abuse of notation, we denote the vector of achievable rates on the m routes as $\boldsymbol{\lambda_r}(\mathcal{Z}, \{(S_i, B, D_i)\}) = (\lambda_r(1), \ldots, \lambda_r(m))$. In the absence of any delay constraint, the achievable rate region would be identical to that of a visible relay. Using standard min-cut arguments, it is easy to see that any vector

$$\lambda_r(i) \leq C_{S_i}\ \forall i, \quad \sum_i \lambda_r(i) \leq C_B \quad (6)$$

is achievable by a visible relay.

TABLE II
PRIORITY MAPPING ALGORITHM: S_1 HIGHEST PRIORITY

1. Initialize $i = 1, j = 1, k = 1$.
2. If $T_B(k) - T_{S_1}(i) \geq \Delta$
   i. Drop i^{th} packet from S_1.
   ii. $i = i + 1$. Repeat Step 2.
3. Let $t = \min\{T_{S_1}(i), T_B(k)\}$.
4. If $t = T_B(k)$
   i. Let $t' = \min\{T_{S_2}(j), T_B(k)\}$.
   ii. If $t' = T_B(k)$ then B transmit dummy packet at t'. $k = k + 1$.
   else if $T_{S_2}(j) \geq T_B(k) - \Delta$
   B transmits j^{th} packet from S_2. $j = j + 1, k = k + 1$.
   else
   $j = j + 1$. Repeat Step 4.ii.
   else
   B transmits i^{th} packet from S_1. $i = i + 1, k = k + 1$.
5. Repeat Steps 2-4 until end of streams.

For a finite delay constraint, a trivial achievable rate region can be obtained if the relay ignores the originating source of the arriving packets. Specifically, the relay uses the BGM algorithm on the joint incoming schedule $\bigcup \boldsymbol{\tau}_{S_i,B}$ and the generated outgoing schedule $\boldsymbol{\tau}_B$. For this strategy, the single source result in Theorem 1 can be easily extended to characterize an achievable rate region for the $m \times 1$ covert relay, which is given in Corollary 1.

Corollary 1: There exists a relaying strategy for an $m \times 1$ covert relay such that the achievable rates

$$\boldsymbol{\lambda_r}(\mathcal{Z}, \{S_i, B, D_i\}) = (\lambda_r(1), \ldots, \lambda_r(m))$$

satisfy $\lambda_r(i) = \lambda_{S_i}(1 - \epsilon(S_i, B)), \forall i$ where

$$\epsilon(S_i, B) \geq f_e(\Delta, \sum_{j=1}^{m} \lambda_{S_j}, C_B), \quad \forall i \tag{7}$$

$$\lambda_{S_i} \leq C_{S_i}, \quad \forall i. \tag{8}$$

Prioritized Scheduling: Ignoring the source identities and considering a single joint stream is strictly suboptimal. Since the relay observes a distinct stream from each source node (by virtue of transmitter directed signaling), the streams can be prioritized to obtain a larger achievable rate region compared to Corollary 1.

Consider a 2×1 relay and assign the highest priority to S_1. For every departure epoch in $\boldsymbol{\tau}_B$, the relay considers all packets that have arrived within Δ time units before that epoch. If some of those packets arrived from S_1 (highest priority), the relay transmits the earliest of those packets at the chosen epoch. If none of the packets arrived from S_1, then the packet that arrived first (from S_2) is transmitted. Since S_1 is given highest priority, this would provide the maximum rate achievable for the stream from S_1. The priority algorithm is described formally in Table II.

By interchanging the priorities and applying the scheduling algorithm, we can obtain the maximum rate for the stream from S_2. It is easy to see that, when none of the sources are given priority, it is equivalent to considering a single joint stream (Corollary 1). By time-sharing relaying strategies with different priority requirements, a piecewise-linear region of achievable rate vectors is obtained, which is characterized in Theorem 2.

Theorem 2:

1. $\boldsymbol{\lambda_r}(\mathcal{Z}, \{(S_1, B, D_1), (S_2, B, D_2)\}) = (\lambda_r(1), \lambda_r(2))$ is achievable if it lies in a hexagonal region whose endpoints are given by $(0, 0)$, $(\lambda_r(1)^{\max}, 0)$, $(\lambda_r(1)^{\max}, \lambda_r(2)^{\min})$, $(\lambda_r(1)^{\text{sum}}, \lambda_r(2)^{\text{sum}})$, $(\lambda_r(1)^{\min}, \lambda_r(2)^{\max})$, $(0, \lambda_r(2)^{\max})$, where

$$\lambda_r(i)^{\text{sum}} = C_{S_i}(1 - f_e(\Delta, C_{S_1} + C_{S_2}, C_B) \tag{9}$$

$$\lambda_r(i)^{\max} = C_{S_i}(1 - f_e(\Delta, C_{S_i}, C_B) \tag{10}$$

$$\lambda_r(i)^{\min} = \frac{(C_B - \lambda_r(j)^{\max}) C_{S_i}(C_{S_j} + C_B)}{C_{S_i} + C_B^2 + C_{S_j} e^{-C_B \Delta}} \tag{11}$$

$$i, j \in \{1, 2\}, i \neq j. \tag{12}$$

2. $(\lambda_r(1), \lambda_r(2))$ is not achievable if

$$\sum_{i=1,2} \lambda_r(i) \geq (C_{S_1} + C_{S_2})(1 - f_e(\Delta, C_{S_1} + C_{S_2}, C_B))$$

$$\lambda_r(i) \geq C_{S_i}(1 - f_e(\Delta, C_{S_i}, C_B)), \quad i = 1, 2. \tag{13}$$

Proof: Refer to the Appendix .

The priority scheduling cannot be proven to obtain the optimal achievable rate region, and so Theorem 2 also provides an outer bound to determine the extent of suboptimality. The outer bound is expressed as an upper bound on the sum rate $\lambda_r(1) + \lambda_r(2)$. This is obtained using the optimality of the BGM algorithm and Corollary 1. It can be shown that as $\Delta \to \infty$, the inner and outer bounds coincide and converge exponentially fast. Although the optimality of the achievable rate region is still an open problem, the strategy achieves the maximum possible sum-rate.

The prioritized scheduling can be extended to more than two sources. For an $m \times 1$ relay, every priority assignment corresponds to an ordering of the m sources. When packets from multiple sources contend for a single epoch, the packet chosen to be transmitted belongs to the source with highest priority. Further, by time-sharing strategies for different priority assignments, the complete region can be obtained.

An example region for the 2×1 relay is shown in Fig. 9. As is evident, the time-sharing strategy results in a piecewise-linear and convex region. The two corner points of the polygon in the figure which represent the achievable rate-pairs when S_2, S_1 are, respectively, given full priority, clearly demonstrate the gains due to prioritized scheduling. Even when S_1 is given full priority, the relay rate for S_2 is strictly positive. If no priority is used, however, S_1 can achieve maximum rate only when S_2 transmits at zero rate (region of Corollary 1). The maximum priority rate-pairs can also be viewed as the outcome of successive application of the BGM algorithm on the incoming streams from the two sources, with the order of application determined using the priority assignment.

From Theorems 1 and 2, it is clear that when C_{S_i}, C_B, and Δ are finite, the relay rates are strictly less than the transmission rates, thereby resulting in a nonzero packet drop rate. Therefore, the source needs to employ forward error correction (FEC) in order to deliver information to the destination reliably. It can be shown that for very long streams, the coding does not result in further reduction of achievable information rates (see Section IV-C).

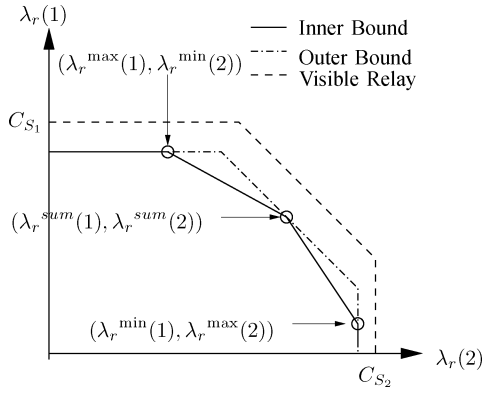


Fig. 9. 2×1 relay rate region. $\mathcal{R}_i$ and $\mathcal{R}_o$ are the inner and outer bounds of Theorem 2. The inner and outer bounds coincide at the maximal sum-rate point. The outermost region is the achievable rate region of a visible relay.

B. Average Delay

Consider the average delay constraint at a relay, as specified by (2) and (3). It is easy to see that a subset of achievable rates under an average delay constraint of $\bar{\mathbf{\Delta}}$ can be trivially obtained by using the algorithms of Section IV-A that assume a strict delay of $\bar{\mathbf{\Delta}}$. This rate region, however, can be significantly improved by using a modified strategy as follows.

Consider the single source relay. Let $m(\Delta, C_{S_1}, C_B)$ represent the mean packet delay obtained when the BGM algorithm is applied with strict delay constraints Δ. Let Δ^* be the strict delay constraint such that the average meandelay of the BGM algorithm satisfies $m(\Delta^*, C_{S_1}, B) = \bar{\mathbf{\Delta}}$. Then applying the BGM algorithm with Δ^*, an improved achievable rate can be obtained.

Theorem 3: $\lambda_r(\mathcal{Z}, (S_1, B, D_1)) = C_{S_1}(1 - \epsilon(S_1, B))$ is an achievable relay rate for an average delay constraint of $\bar{\mathbf{\Delta}}$ if

$$\epsilon(S_1, B) \geq \begin{cases} f_e(\Delta^*, C_{S_1}, C_B), & C_B - C_{S_1} \leq \frac{1}{\bar{\mathbf{\Delta}}} \\ 0, & \text{o.w.} \end{cases}$$

and Δ^* is the solution to $m(\Delta^*, C_{S_1}, C_B) = \bar{\mathbf{\Delta}}$ where

$$m(\Delta^*, C_{S_1}, C_B) = \frac{1 + e^{\Delta^*(C_{S_1} - C_B)}\left[\Delta^*(C_{S_1} - C_B) - 1\right]}{(C_B - C_{S_1})\left[1 - e^{\Delta^*(C_{S_1} - C_B)}\right]}.$$

Proof: Refer to the Appendix

For values of $\bar{\mathbf{\Delta}}$ close to zero, the strict delay constraint required is approximately $2\bar{\mathbf{\Delta}}$. Therefore, for very small delays, relaxation of the strict delay constraint does not provide significant improvement in achievable rate. However, as $\bar{\mathbf{\Delta}}$ increases beyond a certain threshold, the equivalent strict delay Δ^* increases exponentially. In that regime, an achievable rate close to optimal can be obtained even for a finite $\bar{\mathbf{\Delta}}$. Furthermore, as is evident from Theorem 3, when $C_B - C_{S_1} \geq \frac{1}{\bar{\mathbf{\Delta}}}$, the strategy achieves zero packet loss. In other words, every transmitted packet can be relayed successfully within the (average) delay constraint.

Since we consider long streams, this strategy could potentially be improved by dividing the stream into finite number (N) of segments, and implementing the BGM algorithm with different strict delay constraints $\{\Delta_i^*,\ i = 1 \cdots N\}$ in different segments (see Fig. 10). The strict delay constraints

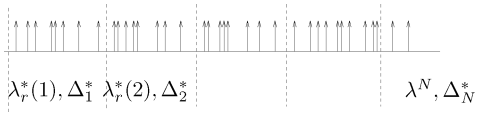


Fig. 10. Delay segmentation: In each segment of the traffic, a different strict delay Δ_i^* is chosen.

should be chosen such that the average delay $\frac{\Sigma_i m(\Delta_i^*, C_{S_1}, B))}{N}$ is less than $\bar{\mathbf{\Delta}}$. As the length of the stream increases, each segment i would provide an achievable relay rate $\lambda_r^*(i) = C_{S_i}(1 - f_e(\Delta_i^*, C_{S_i}, C_B))$ (Theorem 1) and the net achievable rate would be $\frac{\Sigma_i \lambda_r(i)}{N}$. However, for a pair of Poisson processes, $\lambda_r^*(i)$ is a convex $\bigcap$ function of the strict delay Δ_i^*, and hence, this segmentation does not reduce[4] packet loss for a fixed average delay.

Using the relation between the strict delay and average delay in Theorem 3, the achievable region for a general $m \times 1$ relay can also be obtained by appropriately modifying the strict delay constraint in the prioritized scheduling algorithm. The set of transmission rates for which our strategy is optimal in the $m \times 1$ relay case is a straightforward extension of Theorem 3.

Corollary 2: There exists a scheduling strategy that incurs zero packet loss on all incoming streams under average delay constraint $\bar{\mathbf{\Delta}}$, if the medium access constraints satisfy

$$C_B - \sum_i C_{S_i} \geq \frac{1}{\bar{\mathbf{\Delta}}}.$$

From the results presented so far, it is clear that while independent Poisson scheduling generally provides a subset of achievable relay rates for strict delay constraints, under certain conditions on the medium access, it can be optimal for an average delay constraint. An important feature in the algorithms presented is that the relays do not require prior knowledge about transmission schedules of the source nodes. The decision to transmit any packet is based on events occurring between its arrival time and the subsequent departure epoch. This makes it particularly attractive for a decentralized implementation of the scheduling, which is of particular value in *ad hoc* wireless and sensor networks. Note that although the rate expressions derived are for Poisson processes, the algorithms presented are quite general, and can be used on any set of transmission schedules. Furthermore, the optimality of the BGM algorithm also holds for any pair of schedules.

C. Reliability

The independent schedules and relaying algorithms discussed thus far result in strictly nonzero packet drop rate for Poisson processes. Further, since the relay nodes generate schedules in a decentralized manner, it is not possible for the source node to know the identities of packets that would be dropped. This implies that the source nodes must employ FEC techniques to transmit information reliably to the destination. When the traffic is time sensitive such as in media transmission, FEC may not be practical, as it would incur significant coding delay. However,

[4]This convexity may not hold for non-Poisson schedules, in which case, the segmentation could potentially increase the achievable relay rate.

if the strict delay constraint is enforced due to low duty cycles (as in sensor networks) or to maintain stability. FCC ensures reliability at the cost of coding delay.

In order to analyze the reliability of packet transmissions, it is necessary to characterize the channel model between a source and destination. For this purpose, we treat each packet as a binary unit of data, and equate a packet drop to an erasure. Since packets can be indexed, the erasure positions would be known at the destination node.

Consider a relay node forwarding packets from a single source. Let $E(i)$ denote the random variable indicating that packet i was successfully relayed when applying the BGM relay algorithm. Then, using Proposition 4 in [23], it can be shown that the rate specified in Theorem 1 also represents the reliable information rate.

Lemma 1: The capacity C of the binary erasure channel that corresponds to a covert relay using the BGM algorithm is

$$C = 1 - \limsup_n \frac{1}{n} \sum_{i \le n} E(i) = 1 - \epsilon(S_1, B)$$

where $\epsilon(S_1, B)$ is given by (5).

Proof: Refer to the Appendix .

The achievability of this reliable rate, however, requires coding across a long stream of packets. In practice, a packet is not a unit of data and the FEC is different from regular point-to-point communication channels. Specific coding schemes for packet recovery in networks have been discussed in literature [24], [25].

V. Throughput–Anonymity Tradeoff

The achievability results presented in the previous section can be viewed as the basic building blocks for hiding routes in a network. A trivial extension to multihop networks would be to let all relays generate independent transmission schedules. The rate loss incurred at every node would however result in significant loss in overall network throughput. In fact, Theorem 2 in [26] shows that under certain conditions, for an n–hop path with independent Poisson schedules, the maximum rate of packets that can be relayed to the destination with strict delay constraint decays exponentially as n increases. This reinforces the idea of selecting the set of covert relays optimally depending on the desired level of anonymity α.

A. Covert Relay Selection

The source transmission schedules are assumed to belong to independent Poisson processes in each session. We model the transmission schedules to covert relays to be independent Poisson processes as well. Given a session $\boldsymbol{S}$, let $\boldsymbol{B}$ represent the set of relay nodes that are chosen to be covert. Given $\boldsymbol{S}$, $\boldsymbol{B}$, using the relaying algorithms discussed in the previous section, the schedules $\boldsymbol{\tau}$ and the relaying strategy can be generated for all relay nodes in the network. We model the set of covert relays $\boldsymbol{B}$ as a random variable with a conditional pmf $\{q(\boldsymbol{B}|\boldsymbol{S}) : \boldsymbol{B} \in 2^V\}$. The goal is to optimize the conditional pmf $\{q(\boldsymbol{B}|\boldsymbol{S})\}$ so that network throughput is maximized for a given level of anonymity α.

B. Eavesdropper Observation

We assume that when a relay is visible, the eavesdropper perfectly correlates the schedules transmitted by a preceding node and the relay. As a result, depending on the set of visible relays, the eavesdropper can detect a portion of the routes in the session perfectly. We denote this set of partial routes by $\hat{\boldsymbol{S}} \in 2^{\mathcal{P}(\mathcal{G})}$. Using the observation $\hat{\boldsymbol{S}}$, the eavesdropper would try to infer the actual session $\boldsymbol{S}$. The partial observation $\hat{\boldsymbol{S}}$ can be expressed as a deterministic function of the actual session $\boldsymbol{S}$ and the set of covert relays $\boldsymbol{B}$.

We define function $t : 2^{\mathcal{P}(\mathcal{G})} \times \mathcal{V} \to 2^{\mathcal{P}(\mathcal{G})}$ as follows. If $B \neq \phi$, then

$$t(\boldsymbol{P}, B) = \{P : P \in \mathcal{P}(\mathcal{G}) \text{ such that (A1) or (A2) holds.}$$

A1. $\exists P' = (A_1, \ldots, A_k, B, A_{k+1}, \ldots, A_n) \in \boldsymbol{P}$, such that $P = (A_1, \ldots, A_k)$ or $P = (B, A_{k+1}, \ldots, A_n)$.

A2. $P \in \boldsymbol{P}$ and $B \notin P$.

For a set of paths $\boldsymbol{P}$, $t(\boldsymbol{P}, B)$ represents the eavesdropper's observation when node B is covert. Condition 1 states that, when a path in $\boldsymbol{P}$ 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.

If $B = \phi$, then $t(\boldsymbol{P}, \phi)$ is obtained by removing the destination nodes from every path in $\boldsymbol{P}$. This is because, even if all relays are visible, transmitter directed signaling ensures that it is not possible to detect the final destination in any route.

When a subset $\boldsymbol{B} = (B_1, \ldots, B_m) \subset V$ of relays are covert, then $\hat{\boldsymbol{S}}$ can be obtained by repeated application of t

$$\hat{\boldsymbol{S}} = t(\cdots(t(t(\boldsymbol{S}, \phi), B_1)\cdots), B_m) \triangleq \boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B}). \quad (14)$$

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

Lemma 2: If $\hat{\boldsymbol{S}} = \boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B})$, then

1. $\hat{\boldsymbol{S}}$ is a sufficient statistic for detecting $\boldsymbol{S}$ using $\boldsymbol{\tau}$.
2. Given $\boldsymbol{S}$, $\hat{\boldsymbol{S}}$ is an invertible function of $\boldsymbol{B}$.

Proof: Refer to the Appendix .

The preceding lemma shows that, for an eavesdropper, the information contained in $\boldsymbol{\tau}$ about $\boldsymbol{S}$ is completely encapsulated in the observed session $\hat{\boldsymbol{S}}$. Further, the pairs of variables $(\boldsymbol{S}, \boldsymbol{B})$ and $(\boldsymbol{S}, \hat{\boldsymbol{S}})$ are isomorphic, or in other words, there is a one to one correspondence between the two pairs of variables. Therefore, choosing the set of covert relays $\boldsymbol{B}$ is equivalent to designing the eavesdropper observation $\hat{\boldsymbol{S}}$.

C. Throughput Function

The relaying strategies in Section IV-A were designed to minimize the packet loss at a single covert relay. Extending those results to multihop routes, we characterize the loss in sum-rate of each session $\boldsymbol{S}$, when a subset of relays $\boldsymbol{B}$ are covert.

The maximum throughput in the network is achieved when all relays are visible. In a session $\boldsymbol{S}$, the maximum achievable sum-rate can be characterized using the max-flow under the transmission rate constraints. Specifically, let $\boldsymbol{\lambda_r}(\boldsymbol{S}, \phi) = (\lambda_r^v(1), \ldots, \lambda_r^v(|\boldsymbol{S}|))$ represent the vector of

achievable relay rates for the paths in session $\boldsymbol{S}$ with no covert relays, and $\boldsymbol{\Lambda_r}(\boldsymbol{S},\phi)$ be the maximum achievable sum-rate.

Let $\boldsymbol{S} = (P(1),\ldots,P(|\boldsymbol{S}|))$. The maximum sum-rate is achieved when all relays are visible, which is given by the solution to

$$\boldsymbol{\Lambda_r}(\boldsymbol{S},\phi) = \max(\lambda_r^v(1)+\cdots+\lambda_r^v(|\boldsymbol{S}|)) \quad (15)$$

$$\sum_{i:B\in P(i)} \lambda_r^v(i) \le C_B,\ \forall B \in V. \quad (16)$$

Therefore , the maximum throughput when anonymity $\alpha = 0$ is given by

$$R(\alpha=0) = \mathbb{E}(\boldsymbol{\Lambda_r}(\boldsymbol{S},\phi))$$

where the expectation is over the prior $p(\boldsymbol{S})$.

When a subset of relays are covert, the achievable sum-rate in each session is reduced depending on the fraction of packets dropped at each covert relay. Specifically, let $\boldsymbol{\lambda_r}(\boldsymbol{S},\boldsymbol{B}) = (\lambda_r{}^{\boldsymbol{B}}(1),\ldots,\lambda_r{}^{\boldsymbol{B}}(|\boldsymbol{S}|))$ represent the achievable relay rates from sources to destinations for a session $\boldsymbol{S} = (P(1),\ldots,P(|\boldsymbol{S}|))$, when nodes in $\boldsymbol{B}$ are covert, and let

$$\boldsymbol{\Lambda_r}(\boldsymbol{S},\boldsymbol{B}) \triangleq \sum_{i=1}^{|\boldsymbol{S}|} \lambda_r{}^{\boldsymbol{B}}(i)$$

be the achievable sum-rate. If $A(i,j)$ represents the jth node in path $P(i)$, then

$$\lambda_r{}^{\boldsymbol{B}}(i) = \lambda_r^v(i) \prod_{j:A(i,j)\in\boldsymbol{B}\cap P(i)} (1-\epsilon(A(i,j-1),A(i,j))) \quad (17)$$

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 Theorems 1 and 2 provide closed-form expressions for $\epsilon_i(A,B)$ only if B is the first covert relay in the path $P(i)$. Since the departure epochs of data packets from a covert relay do not constitute a Poisson process, the expression cannot be applied to subsequent covert relays. The analytical characterization of multiple covert relays in a path is generally cumbersome, but can be obtained numerically.

Although the solution of the optimization in (15),(16) specifies a set of transmission rates for the nodes, we know from Theorems 1 and 2 that, increasing the transmission rates of nodes can reduce the packet loss. Therefore, if the relay immediately following a source node is covert, the source node could transmit at the maximum rate possible to minimize packet losses. Since only the source is allowed to perform forward error correction, it does not help to increase transmission rates of subsequent relays (as we would only get additional dummy packets).

VI. Performance Characterization

With the eavesdropper observation of (14) and throughput characterization in (17), we now have all the elements required to maximize throughput with anonymity α. Prior to optimizing the general randomized strategy, to ease understanding, we first discuss a simple deterministic strategy to obtain a smaller region of achievable throughput anonymity pairs. Then, expanding on that idea, we will present the optimal strategy to choose covert relays.

A. Deterministic Covert Scheduling

A direct optimization using (17) provides a deterministic strategy to characterize achievable sum-rates under the anonymity requirements. Specifically, a subset $\boldsymbol{B}$ of relays is chosen to remain covert for all sessions, such that the sum-rates are maximized without violating the anonymity requirement. Using Lemma 2, the following result presents an achievable throughput-anonymity region.

Corollary 3: A throughput R is achievable with anonymity α if

$$R \le \max_{\boldsymbol{B}:H(\boldsymbol{S}|\hat{\boldsymbol{S}})\ge\alpha} \mathbb{E}[\boldsymbol{\Lambda_r}(\boldsymbol{S},\boldsymbol{B})]$$

where $\hat{\boldsymbol{S}} = \boldsymbol{T}(\boldsymbol{S},\boldsymbol{B})$.

Depending on the level of anonymity required, the strategy picks the best subset of nodes to remain covert (for all sessions). Since the number of possible subsets of relays is finite, the achievable sum-rate anonymity region would be constant within intervals of α, with sudden jumps corresponding to a change in the optimal subset (see example in Section VII).

B. Probabilistic Covert Scheduling

The drawback in the deterministic strategy is that the subset $\boldsymbol{B}$ is chosen independent of the session $\boldsymbol{S}$. We consider a general class of strategies, where the set of covert relays are chosen according to a random distribution $\{q(\boldsymbol{B}|\boldsymbol{S})\}$ that depends on $\boldsymbol{S}$. The goal is then to optimize $\{q(\boldsymbol{B}|\boldsymbol{S})\}$ so that achievable throughput is maximized for the desired level of anonymity α. The optimal distribution and the corresponding analytical characterization of the optimal throughput is given in the following theorem using an information-theoretic rate–distortion function.

Theorem 4: Let $d : 2^{\mathcal{P}} \times 2^{\mathcal{P}} \to \mathcal{R}$ so that

$$d(\boldsymbol{S},\hat{\boldsymbol{S}}) = \begin{cases} \boldsymbol{\Lambda_r}(\boldsymbol{S},\phi) - \boldsymbol{\Lambda_r}(\boldsymbol{S},\boldsymbol{B}), & \exists \boldsymbol{B} \text{ s.t. } \hat{\boldsymbol{S}} = T(\boldsymbol{S},\boldsymbol{B}) \\ \infty, & \text{o.w.} \end{cases} \quad (18)$$

Then, a throughput $R(\alpha)$ is achievable with anonymity α if

$$R(0) - R(\alpha) \ge D\left(H(\boldsymbol{S})(1-\alpha)\right)$$

where $D(r)$ is the *distortion–rate* function defined as

$$D(r) = \min_{q(\hat{\boldsymbol{S}}|\boldsymbol{S}):I(\boldsymbol{S};\hat{\boldsymbol{S}})\le r} \mathbb{E}(d(\boldsymbol{S},\hat{\boldsymbol{S}})). \quad (19)$$

Proof: Refer to the Appendix .

According to the preceding theorem, $R(\alpha)$ can be expressed using the single-letter characterization of a rate–distortion function. The function $d(\boldsymbol{S},\hat{\boldsymbol{S}})$ represents the reduction in sum-rate in session $\boldsymbol{S}$ when the observed session is $\hat{\boldsymbol{S}}$. Although the loss function parameters do not explicitly include the set of covert relays $\boldsymbol{B}$, we know from Lemma 2 that given $\boldsymbol{S},\hat{\boldsymbol{S}}$, the set of covert relays $\boldsymbol{B}$ is unique. Therefore, the distribution $q(\boldsymbol{B}|\boldsymbol{S})$ to chose covert relays is equivalent to the distortion minimizing distribution in (19). Note that due to the equivalence to a rate–distortion function, the Blahut–Arimoto algorithm [27] provides an efficient iterative technique to obtain $q(\boldsymbol{B}|\boldsymbol{S})$ and to characterize the achievable throughput $R(\alpha)$. Note that the anonymity

α is guaranteed assuming that the eavesdropper is aware of the network topology, the session prior distribution $p(\boldsymbol{S})$, and the optimal strategy $q(\boldsymbol{B}|\boldsymbol{S})$ 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-A. In our model, the level of anonymity α directly corresponds to the rate of compression and the reduction in throughput models the 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 [20], and the numerous models and techniques developed therein could possibly be utilized in anonymous networking. One example is the use of Blahut–Arimoto algorithm as an efficient iterative technique to obtain the optimal distribution of covert relays in a session.

Presently, we have considered independent sessions of observation, which 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. One approach would be to adapt a Markovian model for the temporal correlation of sessions, in which case, we believe that ideas in causal source coding [28] would provide useful insights.

We currently model the entire session as a single entity (the variable $\boldsymbol{S}$) which may not be practical to analyze in a large-scale network. This model could be broken down to hiding each route independently, depending on the level of anonymity required by that particular route.

VII. Example

Consider the switching example given in the beginning of Section V (Fig. 6). During any network session, each source S_i picks a distinct destination D_i. The set of sessions $\mathcal{S}$, contains 24 elements which are assumed equiprobable. For this example, Fig. 11 plots the sum–rate anonymity region for the deterministic and probabilistic strategies discussed previously.

The sum-rate anonymity relationship is convex as seen in the figure. This is because the performance metrics, namely, anonymity and throughput, are average quantities, which permits the use of time sharing to convexify any set of achievable rates. The figure clearly demonstrates the performance improvement due to the randomized covert scheduling. As can be seen, when all relays are visible, the maximum sum-rate $2C$ is achieved with a strictly positive secrecy level. This is because, given the transmission stream from relay M_2 (or M_4), it is not possible for the eavesdropper to detect which packets are received by each destination node. Another interesting observation is that it suffices to make relays M_2, M_4 covert in order to obtain perfect anonymity. This shows that, although making all relays covert ensures perfect secrecy, it may not be necessary.

VIII. Conclusion

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.

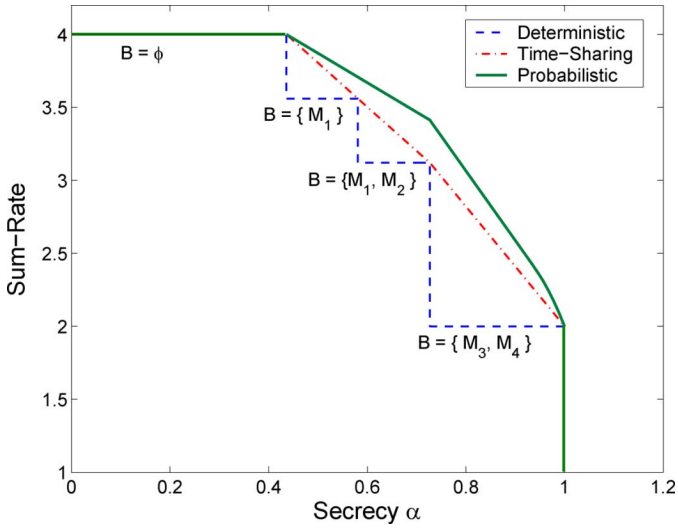


Fig. 11. Throughput anonymity region for 4×4 switching network with $C = 2$. The curve labeled "time-sharing" is achieved by time-sharing multiple deterministic strategies, and results in a convex relationship.

In this work, we used throughput as an indicator of network performance and optimized the selection strategy. However, the framework we establish extends beyond maximizing throughput. In fact, the loss function we define in (18) can be redefined to represent the loss in any convex function of the achievable relay rates. Further, instead of fixing the packet delay and minimizing the loss in sum-rate, we could fix the rates of transmission and analyze the increase in latency at every covert relay. 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. Recently, we provided an approach for decentralized covert relaying in [29], and also presented an achievable latency-anonymity region in [30].

Appendix

Proof of Theorem 1:

To prove the theorem, we adopt the technique used in [19]. Consider the two point processes $\boldsymbol{\tau}_{S_1}, \boldsymbol{\tau}_B$. Let X_j be the jth packet delay, i.e., $X_j = T_B(j) - T_{S_1}(j)$. Define

$$Z_j \triangleq X_j - X_{j-1} = (T_{S_1}(j) - T_B(j-1)) - (T_{S_1}(j) - T_B(j-1)).$$

We see that Z_j's are i.i.d. random variables; each Z_j is the difference between two independent exponential random variables with mean $1/C_B$ and $1/C_{S_1}$, respectively. The process $\{X_j\}_{j=1}^{\infty}$ is a general random walk with step Z_j. Define $X_0 = 0$.

Now for every dummy packet transmitted at t in $\boldsymbol{\tau}_B$, we insert a virtual packet at t in $\boldsymbol{\tau}_{S_1}$; for every packet dropped at time s in $\boldsymbol{\tau}_{S_1}$, we insert a virtual packet at $s+\Delta$ in $\boldsymbol{\tau}_B$. Let the new packet delays after the insertion of virtual packets be $\{X'_j\}_{j=0}^{\infty}$. It can be shown that $\{X'_j\}_{j=0}^{\infty}$ is also a random walk with step Z_j, but it has two absorbing barriers at 0 and Δ, i.e.,

$$X'_j = \min(\max(X'_{j-1} + Z_j,\, 0),\, \Delta).$$

Since it is almost surely impossible for $X'_{j-1} + Z_j$ to be exactly equal to 0 or Δ, each time $X'_j = 0$ corresponds to a dummy transmission in $\boldsymbol{\tau}_B$, and $X'_j = \Delta$ corresponds to a dropped packet in $\boldsymbol{\tau}_{S_1}$. From Example 2.16 in [31], we know that the probability of $X'_j = \Delta$ is given by

$$\Pr\{X'_j = \Delta\} = \frac{1 - \frac{\lambda_{S_1}}{\lambda_B}}{\frac{\lambda_B}{\lambda_{S_1}} e^{-\Delta(\lambda_{S_1} - \lambda_B)} - \frac{\lambda_{S_1}}{\lambda_B}} = \Pr\{X'_j = 0\}.$$

Therefore, the fraction of dropped packets in $\boldsymbol{\tau}_{S_1}$ is

$$\epsilon_A = \frac{\Pr\{X'_i = \Delta\}}{(1 - \Pr\{X'_i = 0\})} = \frac{\lambda_B - \lambda_{S_1}}{\lambda_B e^{-\Delta(\lambda_{S_1} - \lambda_B)} - \lambda_{S_1}}.$$

By replacing the transmission rates λ_{S_i}, λ_B with the maximum values C_{S_i}, C_B, the theorem is proved. In [22], the authors have shown that the BGM algorithm inserts the least chaff fraction for any pair of point processes. Hence, for any $(\lambda_{S_1}, \lambda_B)$, it is impossible to obtain a higher relay rate than (5). This procedure can be extended to multihop by considering multidimensional random walk, but closed-form evaluation of the relay rates is cumbersome, even for a few hops. □

Proof of Theorem 2:

1. Let the zero priority region of Corollary 1 be represented by $\mathcal{R}_0$. Every point on the boundary of $\mathcal{R}_0$ is obtained by letting one node transmit at the highest rate and varying the transmission rate of the other source node from 0 to the maximum value C_{S_i}. The maximum sum-rate point in $\mathcal{R}_0$ is a special case of priority mapping; when each node transmits at full rate and the relay uses BGM on the joint arrival process, it is equivalent to priority mapping with both nodes given equal priority. This corresponds to the vertex $({\lambda_r}^{\text{sum}}(1), {\lambda_r}^{\text{sum}}(2))$. When node 1 is given full priority, the achievable rate pair corresponds to the vertex $({\lambda_r}^{\max}(1), {\lambda_r}^{\min}(i))$. We will present the derivation for this pair of achievable rates. The achievability of the rate-pair when node 2 is given maximum priority can be similarly obtained. By time-sharing across strategies, the complete hexagon of rate-pairs can be shown achievable.

When node S_1 is given max priority, the rate ${\lambda_r}^{\max}(1)$ is obtained using a direct application of Theorem 1. ${\lambda_r}^{\min}(2)$ is obtained using the following derivation for the achievable fraction of packets relayed from source S_2 using the unmarked epochs. Let the stream of dummy packets obtained after application of the BGM algorithm on the stream from source S_1 correspond to times $\{T_1^r, T_2^r, \ldots\}$. We refer to this as the *residual stream*. An epoch T_i^r in the residual stream would correspond to a relayed packet from source S_2 if there exists a packet from S_2 in $[T_{i-1}^r, T_i^r]$ such that it arrived at most Δ time units prior to T_i^r. This does not capture all packets from S_2 as the BGM, but provides a lower bound on the fraction of packets relayed. Let $\lambda_{S_2}^i$ denote the arrival time of the first packet from S_2 after T_{i-1}^r, and let $X_i = T_i^r - T_{i-1}^r$, $Z_i = T_i^r - \lambda_{S_2}^i$. Then the rate of relayed packets from S_2 is given by

$$\begin{aligned}(1 - \epsilon(S_2, B)) &\geq \Pr(Z_i \leq \Delta, X_i \geq Z_i) \\ &= \int_0^{\Delta} C_{S_2} e^{-C_{S_2} x} \Pr(X_i \geq Z_i) \\ &\geq \int_0^{\Delta} C_{S_2} e^{-C_{S_2} x} \Pr(X'_i \geq x)\end{aligned}$$

where X'_i is an exponential random variable such that $\Pr(X'_i \geq x) \leq \Pr(X_i \geq x)$ for all x. It is easy to see that an exponential variable of rate C_B would satisfy the requirement, as it amounts to no packet from S_1 being relayed. We obtain the distribution of variable X'_i as follows. We model X'_i as the interarrival time of a thinned version of the Poisson process with rate C_B where the thinning factor is obtained using the modified matching strategy just described for source 2. It is easy to see that the interarrival time in the residual stream obtained using BGM would have to be at least as large as that obtained using this strategy. The rate of this exponential random variable can be shown to be ${\lambda_r}^* = C_B \frac{C_B + C_{S_1} e^{-C_B \Delta}}{C_B + C_{S_1}}$ (details are omitted due to space constraints). Therefore

$$\begin{aligned}1 - \epsilon(S_2, B) &\geq \int_0^{\Delta} C_{S_2} e^{-C_{S_2} x} \Pr(X'_i \geq x) \\ &= C_{S_2} \frac{(1 - e^{-(C_{S_2} + {\lambda_r}^*)\Delta}}{C_{S_2} + {\lambda_r}^*}.\end{aligned}$$

Subsituting the value for ${\lambda_r}^*$ and computing the corresponding values when S_2 has higher priority, we get the expressions in the theorem. □

2. The outer bound is obtained using the optimality of BGM algorithm. Let node S_i transmit at rates C_{S_i}. Then, the sum information relay rate obtained by using the BGM algorithm on the joint incoming process is given by

$$\sum_i \lambda_r(i) = (C_{S_1} + C_{S_2})(1 - f_e\left(\sum_i C_{S_i}, C_B\right). \quad (20)$$

Since BGM inserts the least fraction of dummy packets[22], this is the maximum sum-rate achievable for the given transmission rates. For each individual source S_i, the best rate possible is obtained if the other source is completely ignored. Therefore, by replacing $\sum_j C_{S_j}$ by C_{S_i} in (20), we can obtain the remaining conditions that specify the outer bound. □

Proof of Theorem 3:

Consider the modified point processes as defined in the proof of Theorem 1. X'_i denotes the ith step size of the random walk between two absorbing barriers. The average delay incurred by the BGM algorithm is equal to the expected mean size of the random walk without including the steps that hit either boundaries. Following the exposition in Example 2.16 in [31, p. 67],

the cumulative distribution of the step size (or delay Δ_i) in the interval $(0, \Delta)$ is given by

$$\Pr(X_i \le x) = \frac{1 - \frac{C_{S_1}}{C_B}\exp(\Delta^* + x)(C_{S_1} - C_B)}{1 - \frac{C_{S_1}^2}{C_B^2}\exp(\Delta^*(C_{S_1} - C_B))}. \quad (21)$$

Using the expression above, the average delay $\bar{\boldsymbol{\Delta}}$ for the BGM algorithm with strict delay Δ can be evaluated as

$$\begin{aligned}\bar{\boldsymbol{\Delta}} &= \mathbb{E}\{X_i' | X_i' \in (0, \Delta^*)\} \\ &= \frac{1 + \exp(\Delta^*(C_{S_1} - C_B))\,[\Delta^*(C_{S_1} - C_B) - 1]}{(C_B - C_{S_1})\,[1 - \exp(\Delta^*(C_{S_1} - C_B))]}.\end{aligned}$$

If $C_B > C_{S_1}$, then as $\Delta^* \to \infty$,

$$\begin{aligned}\bar{\boldsymbol{\Delta}} &= \frac{1 + \exp(\Delta^*(C_{S_1} - C_B))\Delta^*(C_{S_1} - C_B)}{(C_B - C_{S_1})\,[1 - \exp(\Delta^*(C_{S_1} - C_B))]} \\ &= \frac{1}{C_B - C_{S_1}}.\end{aligned}$$

This implies that if $\bar{\boldsymbol{\Delta}} > \frac{1}{C_B - C_{S_1}}$, then the BGM algorithm with $\Delta^* = \infty$ would be sufficient, and more importantly, optimal. It is easy to see that for small values of Δ, the average delay $\bar{\boldsymbol{\Delta}} \approx \frac{\Delta^*}{2}$. In other words, when the allowed delay is very small, relaxing the constraint does not provide significant improvement. □

Proof of Lemma 1:

Consider the modified point processes as defined in the proof of Theorem 1. X_i' denotes the ith step size of the random walk between two absorbing barriers. Consider a subsequence $\hat{X}_i$ of X_i', wherein Z' contains all points in X' that are strictly greater than 0. In other words, $\hat{X}_i$ does not represent any dummy packets. Accordingly, the erasure variable $E(i) = 1_{0 < \hat{X}_i < \Delta}$ because a packet is relayed whenever the random walk does not hit either barriers. Since the point processes are renewal processes, the resulting random walk is stationary and the distribution for X_i' given by (21). Therefore, the erasure $E(i)$ is a stationary and ergodic Markov chain and the capacity of the erasure channel is given by

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_i E(i) &= 1 - \Pr\{\hat{X}_i = \Delta\} \\ &= 1 - \frac{\Pr\{X_i' = \Delta\}}{(1 - \Pr\{X_i' = 0\})} \\ &= 1 - \frac{1 - \frac{\lambda_{S_1}}{\lambda_B}}{\frac{\lambda_B}{\lambda_{S_1}}e^{-\Delta(\lambda_{S_1} - \lambda_B)} - \frac{\lambda_{S_1}}{\lambda_B}} \\ &= 1 - \epsilon(S_1, B).\end{aligned}$$ □

Proof of Lemma 2:

From (17), we know that $\boldsymbol{\lambda_r}(\boldsymbol{S}, \bar{\boldsymbol{B}})$ is an achievable relay rate vector when nodes in $\boldsymbol{B}$ are covert. It remains to be seen that the condition $H(\boldsymbol{S}|\hat{\boldsymbol{S}}) \ge \alpha$ guarantees an anonymity α. For this purpose, it is sufficient to show that

$$H(\boldsymbol{S}|\boldsymbol{\tau}) \le H(\boldsymbol{S}|\hat{\boldsymbol{S}}).$$

Let $\hat{\mathcal{Y}}$ be the schedules generated assuming $\hat{\boldsymbol{S}}$ was a session and none of the nodes were covert. The transmission rates of nodes in $\hat{\mathcal{Y}}$ are assumed identical to $\boldsymbol{\tau}$. For the nodes that are the sources in $\boldsymbol{S}$, the schedules are independent in $\boldsymbol{\tau}$ and $\hat{\mathcal{Y}}$. Session $\hat{\boldsymbol{S}}$ 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 $\boldsymbol{S}$. Therefore, the schedules are independent in $\boldsymbol{\tau}$ as well. Since the remaining nodes relay all received packets within negligible processing delay, $p(\boldsymbol{\tau}|\boldsymbol{S}) = p(\hat{\mathcal{Y}}|\boldsymbol{S})$. Then, using the data processing inequality ($\boldsymbol{S} - \hat{\boldsymbol{S}} - \hat{\mathcal{Y}}$)

$$H(\boldsymbol{S}|\boldsymbol{\tau}) = H(\boldsymbol{S}|\hat{\mathcal{Y}}) \le H(\boldsymbol{S}|\hat{\boldsymbol{S}}).$$

Consider any realization of the variables $\boldsymbol{S}$, $\hat{\boldsymbol{S}}$. Suppose $\exists \boldsymbol{B}_1 \ne \boldsymbol{B}_2$ such that $\boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B}_1) = \boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B}_2) = \hat{\boldsymbol{S}}$. Then, we can write $\boldsymbol{B}_1 = (\boldsymbol{B}, \boldsymbol{B}_1'), \boldsymbol{B}_2 = (\boldsymbol{B}, \boldsymbol{B}_2')$ where $\boldsymbol{B}_1' = (B_{11}, \ldots, B_{1m})$, $\boldsymbol{B}_2' = (B_{21}, \ldots, B_{2n})$, and $\boldsymbol{B}_1' \bigcap \boldsymbol{B}_2' = \phi$. We know that

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

Suppose none of the paths in $\boldsymbol{T}(\boldsymbol{S}, \bar{\boldsymbol{B}})$ contain $\bar{\boldsymbol{B}}_1' \bigcup \bar{\boldsymbol{B}}_2'$, then it does not matter if those relays are covert or not, in which case the subset of covert relays would be $\boldsymbol{B}$.

If $\exists P \in \boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B})$ that contains B_{11}, then $\boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B}_1)$ would contain a path that ends in B_{11}, whereas $\boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B}_2)$ cannot contain such a path. Therefore, we have a contradiction. □

Proof of Theorem 4:

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

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

From the definition of $d(\boldsymbol{S}, \hat{\boldsymbol{S}})$, it is easy to see that if $\nexists \boldsymbol{B}$ s.t. $\hat{\boldsymbol{S}} = \boldsymbol{T}(\boldsymbol{S}, \boldsymbol{B})$, then $q^*(\hat{\boldsymbol{S}}|\boldsymbol{S}) = 0$. Given $\boldsymbol{S}, \hat{\boldsymbol{S}}$, Lemma 2 shows that the set of covert relays $\boldsymbol{B}$ are uniquely determined. Therefore, we can equivalently write $q^*(\hat{\boldsymbol{S}}|\boldsymbol{S}) = q^*(\boldsymbol{B}|\boldsymbol{S})$. Therefore, $q^*(\hat{\boldsymbol{S}}|\boldsymbol{S})$ specifies a valid selection strategy. Since $H(\boldsymbol{S})$ is fixed *a priori*, $I(\boldsymbol{S}; \hat{\boldsymbol{S}}) \le (1-\alpha)H(\boldsymbol{S})$ ensures that an anonymity α is guaranteed. Further, for every $\boldsymbol{B}$, the function d evaluates the difference in achievable rate vectors $\boldsymbol{\lambda_r}(\boldsymbol{S}, \phi)$ and $\boldsymbol{\lambda_r}(\boldsymbol{S}, B)$. Taking expectation over $q^*(\boldsymbol{B}|\boldsymbol{S})$, it is easy to see that the distortion D is achievable with α–anonymity. □

REFERENCES

[1] N. West, *The SIGINT Secrets: The Signal Intelligence War: 1900 to Today*. New York: William Morrow, 1988.

[2] V. L. Voydock and S. T. Kent, "Security mechanisms in high-level network protocols," *ACM Comp. Surv.*, vol. 15, pp. 135–171, 1983.

[3] J.-F. Raymond, "Traffic analysis: Protocols, attacks, design issues and open problems," in *Designing Privacy Enhancing Technologies: Proceedings of International Workshop on Design Issues in Anonymity and Unobservability (Lecture Notes in Computer Science)*, H. Federrath, Ed. Berlin, Germany: Springer-Verlag, 2001, vol. 2009, pp. 10–29.

[4] Q. Sun, D. R. Simon, Y. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in *Proc. 2002 IEEE Symp. Security and Privacy*, Berkeley, CA, May 2002, p. 19.

[5] N. Matthewson and R. Dingledine, "Practical traffic analysis: Extending and resisting statistical disclosure," presented at the Privacy Enhancing Technologies: 4th International Workshop, Toronto, ON, Canada, May 2004.

[6] E. W. Felten and M. A. Schneider, "Timing attacks on web privacy," in *Proc. ACM Conf. Computer and Communications Security*, Athens, Greece, Nov. 2000, pp. 25–32.

[7] D. X. Song, D. Wagner, and X. Tian, "Timing analysis of keystrokes and timing attacks on SSH," in *Proc. 10th USENIX Security Sym.*, Washington, DC, Aug. 2001, pp. 25–40.

[8] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, 1949.

[9] D. Chaum, "Untraceable electronic mail, return addresses and digital pseudonyms," *Commun. ACM*, vol. 24, pp. 84–88, Feb. 1981.

[10] C. Díaz and A. Serjantov, "Generalizing mixes," in *Proc. Privacy Enhancing Technologies Workshop (PET 2003) (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, Apr. 2003, vol. 2760.

[11] D. Kesdogan, J. Egner, and R. Buschkes, "Stop-and-go MIXes providing probabilistic security in an open system," in *Proc. 2nd Int. Workshop on Information Hiding (IH'98)*, Portland, OR, Apr. 1998, vol. 1525, Lecture Notes in Computer Science, pp. 83–98.

[12] C. Gulcu and G. Tsudik, "Mixing e-mail with Babel," in *Proc. Symp. Network and Distributed System Security*, San Diego, CA, Feb. 1996, pp. 2–19.

[13] G. Danezis, R. Dingledine, and N. Mathewson, "Mixminion: Design of a type III anonymous remailer protocol," in *Proc. 2003 Symp. Security and Privacy*, Oakland, CA, May 2003, pp. 2–15.

[14] Y. Zhu, X. Fu, B. Graham, R. Bettati, and W. Zhao, "On flow correlation attacks and countermeasures in mix networks," in *Proc. Privacy Enhancing Technologies Workshop*, Toronto, ON, Canada, May 2004, pp. 207–225.

[15] B. Radosavljevic and B. Hajek, "Hiding traffic flow in communication networks," in *Proc. Military Communications Conf.*, San Diego, CA, 1992, pp. 1096–1100.

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

[17] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 3, pp. 339–348, May 1978.

[18] S. Axelsson, "lIntrusion Detection Systems: A Taxonomy and Survey," Chalmers Univ. Technol., Sweden, Tech. Rep., 2000.

[19] T. He and L. Tong, "Detection of information flows," *IEEE Trans. Inf. Theory*, submitted for publication.

[20] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[21] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proc. Privacy Enhancing Technologies Workshop (PET 2002) (Lecture Notes in Computer Science)*, R. Dingledine and P. Syverson, Eds. Berlin, Germany: Springer-Verlag, Apr. 2002, vol. 2482.

[22] A. Blum, D. Song, and S. Venkataraman, "Detection of interactive stepping stones: Algorithms and confidence bounds," in *Conference on Recent Advance in Intrusion Detection (RAID)*, Sophia-Antipolis, France, Sep. 2004, pp. 258–277.

[23] S. Boucheron and M. R. Salamatian, "About priority encoding transmission," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 699–705, Mar. 2000.

[24] N. Shacham and P. McKenney, "Pakcet recovery in high-speed networks using coding and buffer management," in *Proc. IEEE INFOCOM*, San Francisco, CA, 1990, pp. 124–131.

[25] L. Rizzo, "Effective erasure codes for reliable computer communication protocols," *Proc. ACM SIGCOMM Computer Communication Rev.*, vol. 27, pp. 24–36, Jun. 1997.

[26] T. He, P. Venkitasubramaniam, and L. Tong, "Packet scheduling against stepping-stone attacks with chaff," in *Proc. IEEE Military Communications Conf.*, Washington, DC, Oct. 2006.

[27] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[28] D. Neuhoff and L. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 701–713, Sep. 1982.

[29] P. Venkitasubramaniam and L. Tong, "Throughput-anonymity tradeoff in wireless networks under latency constraints," in *Proc. 2008 IEEE INFOCOM*, Phoenix, AZ, Apr. 2008, pp. 807–815.

[30] P. Venkitasubramaniam and L. Tong, "Anonymity with minimum latency in multihop networks," in *Proc. 2008 IEEE Symp. Security and Privacy*, Oakland, CA, May 2008.

[31] D. Cox and H. Miller, *The Theory of Stochastic Processes*. New York: Wiley, 1965.