Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Maximum Throughput of Clandestine Relay

Ting He, Lang Tong, and Ananthram Swami

Abstract— **The maximum throughput of relaying information flows while concealing their presence is studied. The concealment is achieved by embedding transmissions of information flows into truly independent transmission schedules that resemble the normal transmission behaviors without any flow. Such embedding may reduce the throughput for delay-sensitive flows, and the paper provides a quantitative characterization of the level of reduction. Under a strict or average delay constraint, the maximum normalized throughput is measured by the efficiency of the optimal relay algorithms that embed the most flow into given transmission schedules. Exact analytical solutions and closed-form approximations are derived for renewal schedules, verified by simulations on both synthetic traffic and traces. The results reveal general relationships between the clandestine throughput and system parameters including delay constraints, traffic load, and traffic distributions. In particular, the throughput is found to be negatively related to the burstiness of the cover traffic. Moreover, simulations show that the throughputs of renewal traffic with certain power-law interarrival distributions can closely approximate those of actual traces.**

Keywords: **Information flow/Relayed traffic flow, Clandestine relay, Anonymous networking, Intrusion detection.**

I. Introduction

We consider the problem of relaying information flows as a clandestine operation. We call a relay node a *clandestine relay* if it hides the presence of information flows across it from monitoring agents[1]. Besides its natural applications in intelligence operations, clandestine relay can be part of the anonymous networking paradigm in which the presence of information flows is hidden from traffic analyzers [1]. Understanding clandestine relay also has implications in network security. In the so-called wormhole attack [2], for example, the intruder may channel a flow of information packets through a tunnel unknown to the source and the destination. To what degree the intruder can relay information flows without being detected by networked traffic monitors is at the heart of the problem addressed in this paper.

We assume that traffic monitors are omnipresent; all nodes are subject to monitoring, and their timing traces are sent to a fusion center and analyzed with unbounded computation power. We will, however, restrict ourselves to timing information only. Other flow information, e.g., addresses, flow types, and packet content, will certainly make the monitors more powerful, but the availability of such information makes the analysis specific to certain network setup. For example, such information may not be available if an anonymous routing protocol is used [3].

It is apparent that, if timing is the only information available, it would not be possible to track a specific packet. We assume that packets in an information flow are subject to delay constraints. Such constraints may be strict in the form of a maximum tolerable delay, or flexible in the form of an upper bound on the average delay. If packets in an information flow must be forwarded within a deadline (strict or average), then the transmission timings on the relay route will exhibit certain statistical correlations, and it is such correlations that make it possible to detect the presence of information flows through traffic analysis.

Given that nodes performing clandestine relaying cannot hide the act of transmission, they have to embed transmissions of the information flow into their normal transmission schedules, which provide “cover traffic” for the desired flow. For example, a particular type of cover traffic may be generated from statistically independent transmission schedules. If relay nodes use a fraction of such transmission epochs to relay an information flow, then the traffic analyzer, no matter how powerful it is, will be unable to infer the presence of this flow.

A. Summary of Results, Contributions, and Limitations

The main contributions of this paper include a characterization of the maximum throughput of a clandestine relay relative to that of a normal relay and a study of various factors affecting the throughput. Our specific contributions are:

Optimal flow-embedding algorithms: We develop simple yet provably optimal algorithms to embed information flows into given transmission schedules of cover traffic under a strict or average delay constraint. In particular, the *First In, First Out (FIFO)* algorithm is shown to be optimal under the strict delay constraint and near optimal under the average delay constraint. A clandestine relay can use these algorithms to achieve the maximum throughput while hiding the presence of flow.

Research was sponsored in part by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-06-3-0001 and the National Science Foundation under award CCF-0635070. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence, or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

T. He is with the IBM T.J. Watson Research Center, Hawthorne, NY 10532, USA the@us.ibm.com

L. Tong is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA lt35@cornell.edu

A. Swami is with the Army Research Laboratory, Adelphia, MD 20783, USA aswami@arl.army.mil

[1] In contrast, a *covert relay* means a relay node that hides its identity although the presence of flow may be detectable.

978-1-4244-5871-4/09/$26.00 ©2009 IEEE

Throughput analysis: We characterize the maximum (normalized) clandestine throughput by analyzing the efficiency of the proposed algorithms. Assuming that the cover traffic follows renewal processes, we compute the clandestine throughput based on the limiting distribution of a special Markov process constructed from the embedding algorithm. In the strict delay case, we show that the throughput grows linearly with the maximum delay under tight delay constraints, but then slows down and converges to the total traffic rate polynomially as the delay constraint is relaxed. Moreover, the burstier the cover traffic, the lower the clandestine throughput. The average delay case is shown to be approximately equivalent to the strict delay case with a delay bound twice as large, and the same observation holds.

Simulation studies: We complement the analysis with simulations on both renewal traffic and network traces. Besides confirming the analytical results, and in particular that traffic with exponential *inter-packet delays (IPD's)* has higher throughputs than those with power-law IPD's, the simulations also show that the throughputs of renewal traffic with power-law IPD's closely approximate those of actual traces.

This paper aims to obtain insights on the fundamental limits of clandestine relay operations. Our results are limited by the models and assumptions made for analytical tractability. For example, the renewal traffic model may not be accurate for some networking operations. However, a case study shows that with proper interarrival distributions, renewal traffic can resemble network traces reasonably well (Section V-B). Our network model is also simplified in that we have presented the two-hop relay scenario. Space limitation prevents us from treating the general case of multiple flows over arbitrary hops, but approaches similar to those in [4], [5] can be used.

B. Related Work

The problem of characterizing the maximum throughput of clandestine relay has not been formally studied in the past, but problems sharing common concepts have been investigated. The problem of avoiding traffic analysis using special relays was first considered in [6], where relays called *Mixes* collect packets from multiple users and relay them after encryption and mixing to remove the correlation between incoming and outgoing traffic. While Mix effectively hides the routes of individual packets, a study in [7] showed that long streams of packets under delay constraints can still be correlated. To prevent such flow correlation, the method of *cover traffic* is used to pad the actual traffic with dummy packets such that the overall transmission activities stay fixed [1]. Although fixed scheduling hides the correlation, it is inefficient and requires synchronization across the network, and the fixed patterns themselves might as well reveal the flow. In this paper, we overcome these issues by considering stochastic transmission schedules that resemble the nodes' normal transmission behavior when there is no flow.

Another line of related work is from the traffic analyzer's perspective. Motivated by the detection of *stepping-stone attacks* [8], the problem is to detect relayed traffic flows based on transmission patterns. Although various detectors have been developed (see references in [4], [8]), [4] showed that it is always possible to evade detection by embedding flows in normal transmissions, and the efficiency of such embedding gives a fundamental limit on flow detectability, which is rigorously analyzed under Poisson traffic model in [4] and extended to general renewal processes in [9].

This paper extends our earlier work in [9] in several directions: earlier work only considered flows under strict delay constraint, whereas here we also consider average delay constraint and focus more on closed-form solutions that provide tractable insights; earlier work ignores packet sizes by modeling each schedule as a point process, whereas here we consider variable packet sizes, which allows us to model packet splits and merges.

The rest of the paper is organized as follows. Section II defines the problem, and Section III presents the embedding algorithms, which are analyzed in Section IV. Section V presents simulation results. Then Section VI concludes the paper.

II. Problem Statement

For clarity of presentation, we use uppercase letters to denote random variables, lowercase letters for realizations, boldface letters for vectors, and plain letters for scalars.

A. Flow Models

Denote the incoming and outgoing transmission schedules of a relay node by ON-OFF processes $\mathbf{S}_i$ $(i = 1, 2)$

$$\mathbf{S}_i \triangleq ([S_i^s(k),\ S_i^t(k)])_{k=1}^{\infty}, \tag{1}$$

where $S_i^s(k)$ is the starting time and $S_i^t(k)$ the terminating time of the kth packet[2], with a *packet length*[3] $L_i(k) \triangleq S_i^t(k) - S_i^s(k)$. Schedules $(\mathbf{S}_1,\ \mathbf{S}_2)$ specify the generation of cover traffic, which is what the traffic analyzer can observe. In particular, if the schedules represent truly independent transmission activities, then the act of relay will be invisible to the traffic analyzer, and thus the relay operation is "clandestine".

Under predetermined schedules, the act of relay can be considered a process of embedding an information flow into these schedules. Specifically, as illustrated in Fig. 1, we model such embedding by a decomposition

$$L_i(k) = L_i^{(I)}(k) + L_i^{(C)}(k), \tag{2}$$

where $L_i^{(I)}(k)$ denotes the *effective packet length*, defined as the length of the part of the packet that belongs to an information flow (defined below), and $L_i^{(C)}(k)$ the length of the remaining part, called *chaff noise*. Chaff noise models transmissions that are not part of the flow, including dummy packets, dropped packets, superfluous data padded in packets, and multiplexed packets from other flows. Information bits

[2] Assume that packets in the same schedule do not overlap.

[3] Since the problem is defined in time domain, we measure a packet length by the duration of transmitting that packet.

and chaff bits can be mixed in any order within a packet, and $L_i^{(I)}(k)$, $L_i^{(C)}(k)$ denote their total lengths, respectively (either $L_i^{(I)}(k)$ or $L_i^{(C)}(k)$ can be zero).

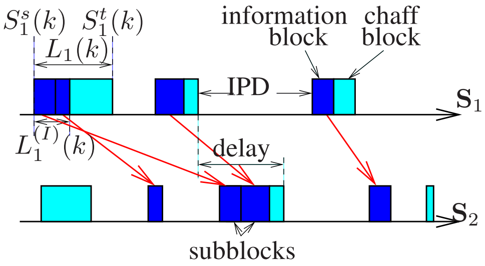

Fig. 1. Decompose schedules $\mathbf{S}_i$ ($i = 1, 2$) into information-carrying subschedules and chaff noise, where the information-carrying subschedules have to match with each other (denoted by arrows) under certain constraints.

We say that transmission schedules $(\mathbf{S}_1, \mathbf{S}_2)$ contain an *embedded information flow* if they can be decomposed as in (2) such that the following definition holds.

Definition 2.1: A pair of transmission schedules $(\mathbf{S}_1, \mathbf{S}_2)$ with effective packet lengths $(\mathbf{L}_1^{(I)}, \mathbf{L}_2^{(I)})$ is a (two-hop) *information flow* if the following conditions hold a.s. (*almost surely*):

1) *Flow-conservation:* $\sum_{m=1}^{\infty} L_1^{(I)}(m) = \sum_{n=1}^{\infty} L_2^{(I)}(n)$;
2) *Causality:* $\sum_{m:\, S_1^t(m)\le t} L_1^{(I)}(m) \ge \sum_{n:\, S_2^s(n)\le t} L_2^{(I)}(n)$ for all t;
3) *Bounded delay:* under strict delay bound Δ,
$$\sum_{m:\, S_1^t(m)\le t} L_1^{(I)}(m) \le \sum_{n:\, S_2^t(n)\le t+\Delta} L_2^{(I)}(n), \quad \forall t; \quad (3)$$
under average delay bound $\bar{\Delta}$,
$$\lim_{t\to\infty} \frac{\sum_{S_2^t(n)\le t} L_2^{(I)}(n) S_2^t(n) - \sum_{S_1^t(m)\le t} L_1^{(I)}(m) S_1^t(m)}{\sum_{S_2^t(n)\le t} L_2^{(I)}(n)} \le \bar{\Delta}. \quad (4)$$

The flow-conservation constraint defines a relay operation by requiring the effective traffic volume to be conserved during relay. The causality constraint ensures that information in a packet can be relayed only after the whole packet arrives, allowing packet-level transformation such as decryption and re-encryption. The delay constraint imposes requirement on the timeliness of relayed information[4], where a strict delay bound enforces every information bit to leave the relay within Δ time of arrival, and an average delay bound only requires the time-averaged delay per information bit to be bounded[5]. The above definition allows packets to be combined, split, delayed, and permuted during relay. Intuitively, the conditions guarantee that there exists a decomposition of the information blocks into subblocks such that the subblocks in the two schedules are in 1-1 correspondence and of equal length (see Fig. 1). We assume that Δ and $\bar{\Delta}$ are known.

B. Clandestine Relay Throughput

The constraints in Definition 2.1 imply that not every transmission in given schedules can be used to relay information. We measure the efficiency of relaying information flow under given schedules by the asymptotic fraction of embedded information flow, stated as follows.

Definition 2.2: Given transmission schedules $(\mathbf{S}_1, \mathbf{S}_2)$, the *maximum normalized throughput of a clandestine relay* (*clandestine relay throughput*) under these schedules is defined as the maximum asymptotic fraction of embedded information flows, i.e.,

$$C(\mathbf{S}_1, \mathbf{S}_2) \overset{\Delta}{=} \sup\{r \in [0, 1] : \exists (\mathbf{L}_i^{(I)})_{i=1}^2 \text{ such that:}$$
$$\text{1) } (\mathbf{S}_i)_{i=1}^2 \text{ with effective packet lengths } (\mathbf{L}_i^{(I)})_{i=1}^2 \text{ is an information flow ;}$$
$$\text{2) } \liminf_{N\to\infty} \frac{\sum_{k=1}^{N} L_1^{(I)}(k) + L_2^{(I)}(k)}{\sum_{k=1}^{N} L_1(k) + L_2(k)} \ge r \text{ a.s.}\}. \quad (5)$$

Under this definition, the clandestine relay throughput is the long-term fraction of information blocks (in length), maximized over all possible ways of embedding them into the given schedules. Intuitively, the clandestine relay throughput tells us for a relay node with incoming schedule $\mathbf{S}_1$ and outgoing schedule $\mathbf{S}_2$, what fraction of the transmission time can be used to relay information.

III. Optimal Embedding Algorithms

It is difficult to compute the clandestine relay throughput directly by Definition 2.2 because it involves an optimization over numerous possible ways of embedding information flows. In this section, we will present algorithms that can compute the clandestine relay throughput efficiently. The idea is to embed the most information bits into given realizations of transmission schedules under flow constraints.

A. Optimal Embedding under the Strict Delay Constraint

For information flows with strict delay, the optimal embedding algorithm turns out to be a simple FIFO matching. The algorithm, called "Strict Greedy Match" (SGM), sequentially scans given transmission schedules and matches each incoming packet with the first relay packet that satisfies the causality and the delay constraints, as shown in Algorithm 1. Specifically, variables m, n denote the indices of the current packets, l_1, l_2 their remaining (unmatched) packet lengths, and C the total length of matched information blocks. The algorithm skips the packets that do not satisfy the causality or the strict delay constraint (lines 3–6) and computes the total length of packets that can be matched (line 8). The fraction of information flow is thus C divided by the total packet length in $\mathbf{s}_1$ and $\mathbf{s}_2$.

[4] We distinguish "delay", denoting the time between the complete arrival of an incoming packet and the complete departure of a relay packet, from "inter-packet delay" (IPD), standing for the OFF period between two consecutive packets transmitted by the same node (see Fig. 1).

[5] The strict and the average delay constraints are extended from [5], where the original models ignore packet sizes.

Algorithm 1 Strict Greedy Match (SGM)

```
Require: Schedules $(\mathbf{s}_1, \mathbf{s}_2)$, maximum delay $\Delta$.
Ensure: Return the maximum length of information blocks
    in $(\mathbf{s}_1, \mathbf{s}_2)$ under strict delay $\Delta$.
 1: initialize: $m, n \leftarrow 1$, $l_i \leftarrow s_i^t(1) - s_i^s(1)$ $(i = 1, 2)$,
    $C \leftarrow 0$
 2: while $m, n$ are valid indices in $\mathbf{s}_1, \mathbf{s}_2$ do
 3:   if $s_2^s(n) < s_1^t(m)$ then {noncausal}
 4:     $l_2 \leftarrow 0$ {skip the packet in $\mathbf{s}_2$}
 5:   else if $s_1^t(m) < s_2^t(n) - \Delta$ then {delay $> \Delta$}
 6:     $l_1 \leftarrow 0$ {skip the packet in $\mathbf{s}_1$}
 7:   else {find a valid match}
 8:     $C \leftarrow C + 2\min(l_1, l_2)$
 9:     $l_1 \leftarrow l_1 - \min(l_1, l_2)$, $l_2 \leftarrow l_2 - \min(l_1, l_2)$
10:   if $l_1 = 0$ then {update indices}
11:     $m \leftarrow m + 1$, $l_1 \leftarrow s_1^t(m) - s_1^s(m)$
12:   if $l_2 = 0$ then
13:     $n \leftarrow n + 1$, $l_2 \leftarrow s_2^t(n) - s_2^s(n)$
14: return $C$
```

This algorithm is an extension of the algorithm "Bounded Greedy Match" proposed in [8], which ignores packet lengths. Since SGM is sequential, it is suitable for online embedding of information flows even if the incoming schedule is not known at the relay beforehand[6]. Simple as it is, SGM is actually optimal as stated in the following proposition.

Proposition 3.1: For any given schedules $(\mathbf{s}_1, \mathbf{s}_2)$ and strict delay constraint Δ, SGM maximizes the total length of embedded information blocks.

Proof: The proof borrows the idea in [8], where the key is to build a 1-1 correspondence between unmatched packets in our algorithm and the unmatched packets in an optimal algorithm M^*. Suppose a data unit dt_1 in $\mathbf{s}_1$ is matched by M^* but not SGM. Then since SGM always tries to match earlier packets first, the relay of dt_1 according to M^* must be matched to some data unit before dt_1 by SGM. The process repeats until reaching a data unit dt_m that is matched in SGM but unmatched in M^*. Thus, the length of matched blocks in SGM is no smaller than that in M^*. ■

B. Optimal Embedding of Flows under the Average Delay Constraint

As we relax the delay constraint to a bounded average delay, the role of the constraint changes dramatically from a hard bound to a "budget" which can be spread over many packets. To perform optimal embedding, we resort to the duality between maximizing the amount of matched data and minimizing the average delay, which leads to an algorithm called "Average Greedy Match" (AGM), shown in Algorithm 2. Algorithm AGM iteratively finds a maximum matching such that the average delay is bounded by $\bar{\Delta}$. In each iteration (lines 2–12), AGM finds the pair of packets that has the minimum delay among all the causal pairs of nonempty packets (line 3; a packet is called "empty" when its unmatched length is zero) and computes the new average delay assuming that this pair is matched (line 4). If the new average delay is within $\bar{\Delta}$, then the matching is finalized (lines 6–9) and the iteration continues; otherwise, we only match a portion of these packets to meet the delay budget $\bar{\Delta}$ and the iteration stops (lines 11–12). The following proposition shows the optimality of AGM.

Algorithm 2 Average Greedy Match (AGM)

```
Require: Schedules $(\mathbf{s}_1, \mathbf{s}_2)$, maximum average delay $\bar{\Delta}$.
Ensure: Return the maximum total length of information
    blocks in $(\mathbf{s}_1, \mathbf{s}_2)$ under average delay constraint $\bar{\Delta}$.
 1: initialize: $C \leftarrow 0$, $\bar{d} \leftarrow 0$, $\mathbf{l}_1, \mathbf{l}_2 \leftarrow$ initial packet lengths
 2: while $\exists$ a causal pair of nonempty packets do
 3:   $(m, n) \leftarrow$ indices of the packet pair with the minimum
      delay among all causal, nonempty pairs
 4:   compute the new average delay:
        $\bar{d}_0 \leftarrow \dfrac{C\bar{d} + 2(s_2^t(n) - s_1^t(m))\min(l_1(m), l_2(n))}{C + 2\min(l_1(m), l_2(n))}$
 5:   if $\bar{d}_0 \le \bar{\Delta}$ then {the delay constraint is satisfied}
 6:     $C \leftarrow C + 2\min(l_1(m), l_2(n))$
 7:     $\bar{d} \leftarrow \bar{d}_0$
 8:     $l_1(m) \leftarrow l_1(m) - \min(l_1(m), l_2(n))$
 9:     $l_2(n) \leftarrow l_2(n) - \min(l_1(m), l_2(n))$
10:   else {the delay constraint is violated}
11:     $C \leftarrow C + C(\bar{\Delta} - \bar{d})/(s_2^t(n) - s_1^t(m) - \bar{\Delta})$
12:     go to line 13
13: return $C$
```

Proposition 3.2: For any given $(\mathbf{s}_1, \mathbf{s}_2)$ and average delay constraint $\bar{\Delta}$, AGM maximizes the total length of embedded information blocks.

Proof: The proof is by contradiction. Assume another algorithm M^* can embed more data under the same constraint. Then if embedding the same amount of data, M^* must achieve a smaller average delay than AGM. This contradicts the fact that AGM minimizes the average delay for any given amount of matched data. ■

Unlike SGM, AGM cannot be used for online scheduling of relay because it is not sequential[7], unless the relay knows the incoming schedule, e.g., by exchanging seeds for the random schedule generators. It is, however, representative of what can be achieved by sequential algorithms (see Fig. 3).

IV. Computing the Clandestine Relay Throughput

This section is dedicated to analytical characterization of the clandestine relay throughput. Throughout the section, we consider fixed-length packets for simplicity, i.e., every packet

[6] The source can overcome dropped packets by using forward error correction codes.

[7] Algorithm AGM may scan the schedules multiple times with a total complexity of $O(N^2)$, where N is the number of packets, whereas SGM has complexity $O(N)$.

has equal length l. Variable-length packets can be handled by partitioning them into smaller "packets" of fixed length[8]. For fixed-length packets, it suffices to specify the starting times of a transmission schedule:

$$\mathbf{S}_i = (S_i^s(k))_{k=1}^{\infty}, \quad i = 1,\, 2, \tag{6}$$

or $\mathbf{S}_i = (S_i(k))_{k=1}^{\infty}$ for short, where each *interarrival time* $S_i(k+1) - S_i(k)$ is the summation of packet length l and the IPD. In this section, we consider a special family of schedules with i.i.d. interarrival times, i.e., *renewal processes*. While network traffic is not really renewal, we observe in simulations that renewal traffic with certain interarrival distribution has similar clandestine relay throughputs as traces (see Section V-B) and thus assume renewal traffic for rigorous analysis.

A. Clandestine Relay Throughput under the Strict Delay Constraint

For flows with strict delay, the clandestine relay throughput is computed by the optimal embedding algorithm SGM. For fixed-length packets, we see from Algorithm 1 that in the jth iteration ($j \geq 1$), one of the following cases will happen:

1) if $s_2(n) - s_1(m) < l$, then the matching is noncausal, and $n \leftarrow n + 1$;
2) if $s_2(n) - s_1(m) > \Delta$, then the delay is too large, and $m \leftarrow m + 1$;
3) if $s_2(n) - s_1(m) \in [l,\, \Delta]$, then this is a valid pair of information packets, and $n \leftarrow n + 1$, $m \leftarrow m + 1$.

Combining the above, we obtain an evolution of the pairwise packet delay $Y_j \overset{\Delta}{=} S_2(n) - S_1(m)$ (capital letters denote random variables):

$$Y_j = \begin{cases} Y_{j-1} + V_j & \text{if } Y_{j-1} < l \\ Y_{j-1} - U_j & \text{if } Y_{j-1} > \Delta \\ Y_{j-1} + V_j - U_j & \text{o.w.,} \end{cases} \tag{7}$$

where $U_j \overset{\Delta}{=} S_1(m) - S_1(m-1)$, $V_j \overset{\Delta}{=} S_2(n) - S_2(n-1)$ denote the next interarrivals in $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively.

The merit of introducing Y_j is that it bridges the algorithm and the analysis. On the one hand, each Y_j within the interval $[l,\, \Delta]$ corresponds to a pair of information packets, whereas each Y_j outside this interval corresponds to a chaff packet; on the other hand, the behavior of Y_j's is easily analyzable because of the following property.

Proposition 4.1: If $\mathbf{S}_i$ ($i = 1,\, 2$) are renewal processes, then $\mathbf{Y} \overset{\Delta}{=} (Y_j)_{j=1}^{\infty}$ ($Y_1 \overset{\Delta}{=} V_1 = U_1$) is a Markov process with transition (7).

Proof: Since the interarrivals U_j, V_j are independent for difference j, Y_j depends on $(Y_k)_{k \leq j-1}$ only through Y_{j-1}. Therefore, $\mathbf{Y}$ is Markovian. ■

The Markovianess of $\mathbf{Y}$ leads to a unique property of SGM that given the current Y_j, whether or not the next packet can relay information is independent of the history. The result is that in the long run, the behavior of SGM looks like repetitions of short durations, and thus the fraction of matched packets will converge as time increases. This is the idea of stochastic stability. Based on this idea, we derive an analytical solution to the clandestine relay throughput.

Theorem 4.2: If $\mathbf{S}_1$ and $\mathbf{S}_2$ are i.i.d. renewal processes with interarrival *probability density function (pdf)* $f(x)$ ($f(x) \equiv 0$ for $x < l$), then the fraction of packets matched by SGM converges a.s., and the limit (i.e., the clandestine relay throughput under the strict delay constraint) is

$$C_d(\Delta) \overset{\Delta}{=} C(\mathbf{S}_1,\, \mathbf{S}_2;\, \Delta) = \frac{2 - 2q}{2 - q}, \tag{8}$$

where $q \overset{\Delta}{=} \lim_{j \to \infty} \Pr\{Y_j \notin [l,\, \Delta]\}$ can be computed by $1 + H(l) - H(\Delta)$, where $H(x)$ ($x \in \mathbb{R}$) is the invariant *cumulative distribution function (cdf)* of $\mathbf{Y}$. Furthermore, $H(x)$ is the solution to

$$\begin{aligned} H(x) = L(x) &+ \int_{-\infty}^{l} H(y) f(x - y) dy \\ &+ \int_{l}^{\Delta} H(y) g(x - y) dy + \int_{\Delta}^{\infty} H(y) f(y - x) dy, \end{aligned} \tag{9}$$

where $g(x)$ is the convolution of $f(x)$ and $f(-x)$, defined as $g(x) \overset{\Delta}{=} \int_0^{\infty} f(y) f(y - x) dy$, and

$$\begin{aligned} L(x) \overset{\Delta}{=} [F(x - l) - G(x - l)]\, H(l) + \big[G(x - \Delta) \\ + F(\Delta - x) - 1\big] H(\Delta) \end{aligned} \tag{10}$$

with $F(x)$, $G(x)$ being the cdf's of $f(x)$, $g(x)$, respectively.

Proof: Assume $\mathbf{Y}$ has the property that the frequency for Y_j to fall outside the interval $[l,\, \Delta]$ converges a.s. to a constant, defined as q. Then since each Y_j outside $[l,\, \Delta]$ represents a chaff packet whereas each Y_j inside the interval represents a pair of information packets, we see that SGM converges a.s., and the clandestine relay throughput, which is the limiting fraction of information packets, is given by $2(1 - q)/(2 - q)$. By definition, $q = H(l) + 1 - H(\Delta)$ for the invariant cdf $H(x)$ of $\mathbf{Y}$, and it can be shown that this invariant cdf must satisfy (9). For details, see [10]. ■

Theorem 4.2 provides both numerical and analytical methods of computing the clandestine relay throughput. Numerically, since SGM converges, we can run SGM on sufficiently long realizations of the schedules; analytically, we can directly solve (9) to compute (8). Equation (9) is a *Fredholm integral equation of the second kind* and can be solved by *Neumann Series* through iterations [11].

In general, there is no closed-form expression for the clandestine relay throughput[9]. Instead, we focus on regimes where clean results can be obtained to provide insight into the relationship between the clandestine relay throughput and

[8] The optimality of the proposed algorithms holds for arbitrary packet lengths; further analysis is left for future work.

[9] More explicit upper and lower bounds for special families of distributions have been obtained and will be reported elsewhere.

the traffic distribution. For the special case of exponential interarrival distribution, i.e., $f(x) = \lambda e^{-\lambda x}$ and $l = 0$, the schedules are Poisson processes of rate λ, and we have a closed-form solution

$$C_d^{\text{Exp}}(\Delta) = \frac{\lambda\Delta}{1+\lambda\Delta}. \tag{11}$$

For large delays ($\lambda(\Delta - l) \gg 1$), it has been shown in [12] that for interarrival times with finite variance[10],

$$C_d(\Delta) \approx \frac{\lambda(\Delta - l)}{\gamma + \lambda(\Delta - l)}, \tag{12}$$

where λ is the traffic rate ($\lambda \overset{\Delta}{=} 1/\mathbb{E}[U]$ for interarrival U), and $\gamma \overset{\Delta}{=} \text{Var}[U]/(\mathbb{E}[U])^2$ is the *dispersion coefficient* of the interarrival distribution. For small delays, we provide the following approximation.

Corollary 4.3: Under the conditions in Theorem 4.2, if $\Delta < 2l$, then

$$C_d(\Delta) \approx \lambda(\Delta - l), \tag{13}$$

where λ is the traffic rate.

Proof: If $\Delta < 2l$, then different incoming packets will have different candidate relay packets, and thus, we can approximate $C_d(\Delta)$ by the probability that a particular packet is matched[11]. As illustrated in Fig. 2, for a packet $S_1(k)$ to be matched to the first relay packet $S_2(j)$ after it, it has to completely fall into an interval of length Δ right before $S_2(j)$, i.e., its starting time needs to fall into an interval of length $\Delta - l$. Since the processes are independent, $S_1(k)$ is uniformly distributed between $S_2(j-1)$ and $S_2(j)$, and thus the probability of matching is $(\Delta - l)\mathbb{E}[1/U]$ for an interarrival U. By Jensen's inequality, it is lower bounded by $(\Delta - l)/\mathbb{E}[U] = \lambda(\Delta - l)$. ■

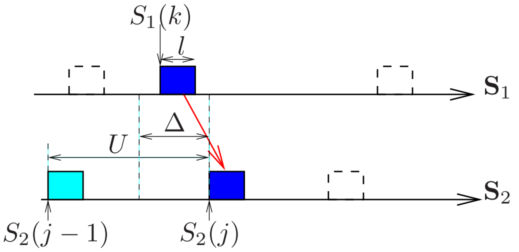

Fig. 2. At a small delay ($\Delta < 2l$), the clandestine relay throughput can be approximated by the probability that a random incoming packet falls within the Δ-length window before its candidate relay packet.

From (12) and (13), we see that the clandestine relay throughput is only a function of the *effective delay* $\lambda(\Delta - l)$ (i.e., the flexible portion of the delay $(\Delta - l)$ normalized by the mean interarrival $1/\lambda$), and is invariant to the scaling of delay, packet length, and traffic rate as long as this effective delay remains the same. The clandestine relay throughput converges to one as the effective delay increases, and the convergence rate is linear at small delays but $O(1/(\lambda(\Delta - l)))$ at large delays. Moreover, Corollary 4.3 reveals a surprising fact that although at larger delays, the interarrival distribution matters (e.g., through the dispersion coefficient), at small delays, the distribution becomes immaterial. The borderline $\Delta^* = 2l$ of Corollary 4.3 actually acts as a threshold separating the two regimes (see Fig. 3).

B. Clandestine Relay Throughput under the Average Delay Constraint

Under the average delay constraint, the constraint is flexible because large delays are allowed as long as there are sufficiently many small delays to average them out. This nature of the constraint leads to the following property of the clandestine relay throughput.

Proposition 4.4: The clandestine relay throughput under the average delay constraint $C_a(\bar{\Delta}) \overset{\Delta}{=} C(\mathbf{S}_1,\ \mathbf{S}_2;\ \bar{\Delta})$ is a concave function with respect to $\bar{\Delta}$ for any $\mathbf{S}_i$ $(i = 1,\ 2)$.

Proof: For any $p \in [0,\ 1]$ and $\Delta_1,\ \Delta_2 > 0$ s.t. $\bar{\Delta} = p\Delta_1 + (1-p)\Delta_2$, one way to achieve an average delay $\bar{\Delta}$ is to use AGM with constraint Δ_1 for p fraction of the time and Δ_2 for the rest[12], which will yield a throughput of $pC_a(\Delta_1) + (1-p)C_a(\Delta_2)$. By definition, $C_a(\bar{\Delta})$ must be no smaller than this linear combination, and thus $C_a(\cdot)$ is concave. ■

In AGM, whether a packet can be matched depends on the delays of all the other matched packets. Such global correlation renders the previous Markovian approach inapplicable. Instead, we seek to bound $C_a(\bar{\Delta})$. It is easy to see that using SGM with a strict delay constraint Δ such that the average delay is $\bar{\Delta}$ provides a lower bound on $C_a(\bar{\Delta})$.

Theorem 4.5: Given i.i.d. renewal processes $\mathbf{S}_1,\ \mathbf{S}_2$ with interarrival pdf $f(x)$, then the clandestine relay throughput under the average delay constraint $\bar{\Delta}$ satisfies

$$C_a(\bar{\Delta}) \geq C_d(2\bar{\Delta} - l). \tag{14}$$

Proof: Note that by symmetry of the transition (7), the limiting distribution $\mathbf{Y}$ must be symmetric around $(l+\Delta)/2$. It implies that the average delay achieved by SGM using a strict delay constraint Δ is equal to $(l + \Delta)/2$. Therefore, using SGM with $\Delta = 2\bar{\Delta} - l$ satisfies the average delay constraint and yields a lower bound $C_d(2\bar{\Delta} - l)$. ■

Theorem 4.5 provides an explicit relationship between the two types of clandestine relay throughput. Actually, a stronger claim holds that any time-sharing of SGM under various delay constraints such that the overall average delay constraint is satisfied will provide a lower bound on $C_a(\bar{\Delta})$. We have, however, observed in simulations that $C_d(\cdot)$ is also a concave function of Δ. Therefore, time-sharing will not improve the throughput, and the lower bound in (14) is in fact the maximum clandestine relay throughput that can be achieved by SGM under an average delay $\bar{\Delta}$.

[10] The original result in [12] is for point processes, but it can be easily generalized to ON-OFF processes with fixed packet lengths.

[11] This is an approximation because the matchings of incoming packets are correlated.

[12] The time-sharing can be implemented by dividing the time window into two parts with the ratio $p/(1-p)$ and then increasing the window size.

As shown in Section V-A, the lower bound provides a good approximation of the clandestine relay throughput computed by AGM, indicating that SGM is near optimal under the average delay constraint. Combining this with the previous results in (12) and (13) shows that relaxing a strict delay constraint to an average delay constraint doubles the effective delay, and: (i) at small delays ($\Delta = \bar{\Delta} < 3l/2$), the clandestine relay throughput under the average delay constraint is twice as large as that under the strict delay constraint, and (ii) at large delays, the fraction of chaff noise (i.e., $1-$ clandestine relay throughput) under the average delay constraint is about half of that under the strict delay constraint if the dispersion coefficient is finite; both results have been verified numerically (see Fig. 3).

V. Simulations

A. Simulations on Renewal Processes

Fixing the mean interarrival time $1/\lambda$ and the packet length l ($\lambda l \leq 1$), we simulate the embedding algorithms on renewal processes of the shifted exponential and the Pareto distributions; see Table I for their properties. These distributions represent transmission schedules with exponential and power-law IPD's, respectively. Since the exponential IPD is typically assumed in analysis, whereas the power-law IPD has been shown to fit network traces [13], we use them to investigate the influence of the Poisson assumption with respect to the clandestine relay throughput.

TABLE I

Interarrival distributions in the simulations

Distribution	PDF	Dispersion coefficient
Shifted exponential	$\lambda' e^{-\lambda'(x-l)}$ ($\lambda' = \frac{\lambda}{1-\lambda l}$)	$\frac{1}{(1+\lambda' l)^2}$
Pareto	$\beta l^{\beta} x^{-\beta-1}$ ($\beta = \frac{1}{1-\lambda l}$)	$\frac{1}{\beta(\beta-2)}$ if $\lambda l > \frac{1}{2}$ ∞ o.w.

We first simulate the clandestine relay throughput for varying delay constraints, as shown in Fig. 3. The approximation in Corollary 4.3 matches the simulation exactly when it applies, whereas the approximation[13] (12) has some error which diminishes at large Δ (not shown). There exist thresholds $\Delta_d^* \approx 2l$ and $\Delta_a^* \approx 3l/2$ for C_d and C_a, respectively, below which the clandestine relay throughputs grow linearly with the delay, and their values are largely independent of the interarrival distribution (see (13), (14)). Above the thresholds, the growth slows down, and the shifted exponential distribution (bold curves) has higher throughputs than the Pareto distribution (plain curves). Since traffic with power-law IPD's is more likely to have large gaps and bursts compared with traffic with exponential IPD's, the result indicates a negative correlation between the clandestine relay throughput and the traffic burstiness, which confirms the asymptotic result in (12) because the Pareto distribution has a larger dispersion coefficient. Moreover, the lower bound on C_a achieved by SGM (dashed curves) is uniformly tight, suggesting that SGM is near optimal under the average delay constraint. Comparing the two types of clandestine relay throughputs shows that: (i) for $\Delta < \Delta_a^*$, $C_a(\Delta) \approx 2C_d(\Delta)$, and (ii) for $\Delta \gg \Delta_d^*$, $1 - C_a(\Delta) \approx (1 - C_d(\Delta))/2$ for the shifted exponential distribution, which are consistent with the analysis in Section IV-B[14].

We then study the effect of the packet length l under fixed delay constraints; see Fig. 4. The plot shows that there exists a packet length $l_d^* \approx \Delta/2$ such that the clandestine relay throughputs are maximized at l_d^*. The phenomenon can be explained by two contradictory effects of increasing l: on the one hand, it compresses the IPD (since the total mean interarrival is fixed) and thus reduces burstiness of the traffic; on the other hand, it tightens the causality constraint and makes it harder to embed flows. At large l ($l > l_d^* \approx \Delta/2$ for C_d and $l > l_a^* \approx 2\Delta/3$ for C_a), the approximation in Corollary 4.3 applies, and both throughputs decay linearly with l regardless of the distribution. In designing transmission schedules, these effects should be taken into account to maximize (or minimize) the clandestine relay throughput as the application requires.

Finally, we study the clandestine relay throughput for variable-length packets as compared with fixed-length packets[15]; see Fig. 5. Under both exponential and power-law IPD's, we compare the clandestine relay throughputs for fixed, uniformly distributed, and exponentially distributed packet lengths. In contrast to the fixed-packet-length case (Fig. 4), the clandestine relay throughput in the variable-packet-length case may simply decreases with the average packet length. This is mainly because for variable-length packets, the traffic burstiness is not reduced as much (and may even increase) as the length grows, and thus the effect of a tighter causality constraint dominates. How much the packet lengths vary negatively affects the clandestine relay throughput: the constant, uniform, and exponential distributions have increasing dispersion coefficients[16] and decreasing clandestine relay throughputs correspondingly.

B. Simulations on Traces

We simulate the proposed algorithms on network traces to study the clandestine relay throughput in practice. We use the traces LBL-PKT-4, which contains an hour's wide-area traffic between the Lawrence Berkeley Laboratory and the Internet[17]. The simulated clandestine relay throughput is then compared with the clandestine relay throughput of renewal processes, as shown in Fig. 6. We fit the traces with both the shifted exponential and the Pareto interarrival distributions, and find that the Pareto distribution with $\beta = 0.6$ gives a good approximation to the clandestine relay throughput

[13] Note that (12) is not applicable to the Pareto distribution here because its dispersion coefficient is infinite.

[14] The result does not apply to the Pareto distribution here because it has an infinite dispersion coefficient.

[15] Only C_d is plotted for clarity. Similar observation holds for C_a.

[16] Their dispersion coefficients are 0, 1/3, and 1, respectively.

[17] The traces were collected by Paxson and first used in his paper [13], from which we extract 134 TCP traces of 1000 packets each. The traces can be obtained from `http://ita.ee.lbl.gov/html/contrib/LBL-PKT.html`.

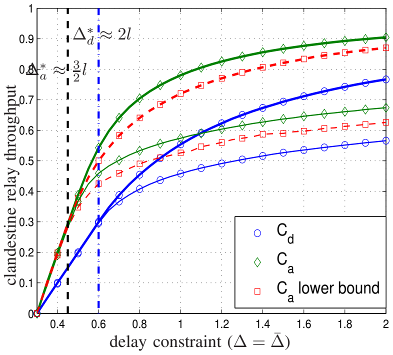

Fig. 3. Clandestine relay throughput vs. delay constraint ($\lambda = 1$, $l = 0.3$, 10^4 packets per process). Bold line: shifted exponential; plain line: Pareto.

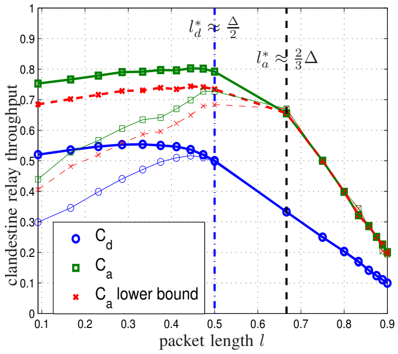

Fig. 4. Clandestine relay throughput vs. packet length ($\lambda = 1$, $\Delta = \bar{\Delta} = 1$, β varies between 1.1 and 10, 10^4 packets per process). Bold line: shifted exponential; plain line: Pareto.

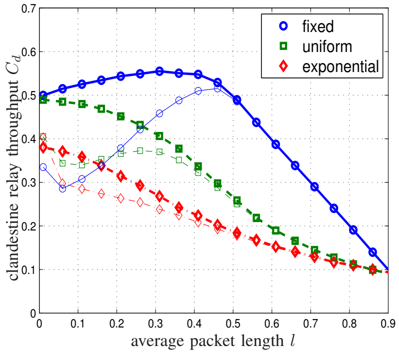

Fig. 5. Packet length distributions ($\lambda = 1$, $\Delta = 1$, mean interarrival $= 1/\lambda$, 10^4 packets per process, 10^3 Monte Carlo runs). Bold line: exponential IPD; plain line: power-law IPD.

of the traces, which is consistent with the previous studies in [13] that have claimed these traces to have Pareto-like interarrival distributions[18]. Since $\beta < 1$ implies infinite mean interarrival and zero traffic rate, the result suggests that traces have much higher bustiness and lower clandestine relay throughputs than renewal processes of the same rates.

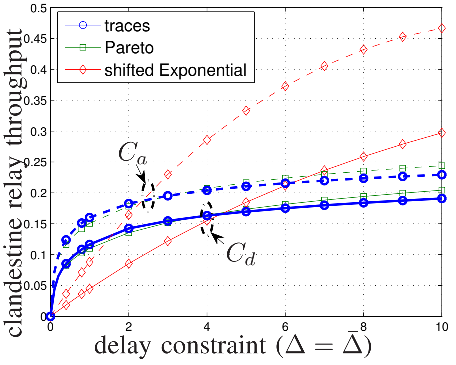

Fig. 6. Clandestine relay throughputs of traces and renewal processes (the packet length of each trace is estimated by its minimum interarrival, $\beta = 0.6$, $\lambda' = 0.1$, 10^3 packets per process).

VI. Conclusion

We have studied in detail the maximum throughput of a clandestine relay under stochastic transmission schedules and strict or average delay constraints. Efficient algorithms are developed to schedule the relay of flows under arbitrary transmission schedules with variable-length packets, and their efficiency is characterized analytically for renewal schedules with constant packet lengths. The result establishes a fundamental limit of clandestine communications and provides insights on how to constrain/improve it based on application needs by tuning transmission schedules.

References

[1] B.Radosavljevic and B. Hajek, "Hiding traffic flow in communication networks," in *Military Communications Conference*, 1992.

[2] Y. Hu, A. Perrig, and D. B. Johnson, "Packet Leashes: A Defense against Wormhole Attacks in Wireless Ad Hoc Networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, March 2003.

[3] J. Kong and X. Hong, "Anodr: Anonymous on demand routing with untraceable routes for mobile ad-hoc networks," in *ACM lnternational Symposium on Mobile Ad Hoc Nefworking and Computing*, Annapolis, MD, June 1–3 2003.

[4] T. He and L. Tong, "Detection of Information Flows," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4925–4945, November 2008.

[5] P. Venkitasubramaniam, T. He, and L. Tong, "Anonymous Networking amidst Eavesdroppers," *IEEE Transactions on Information Theory, Special Issue on Information-Thoeretic Security*, vol. 54, no. 6, pp. 2770–2784, June 2008.

[6] D. Chaum, "Untraceable electronic mail, return addresses and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–88, February 1981.

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

[8] A. Blum, D. Song, and S. Venkataraman, "Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds," in *the 7th International Symposium on Recent Advances in Intrusion Detection (RAID)*, Sophia Antipolis, France, Sept 2004.

[9] T. He, A. Agaskar, and L. Tong, "On Security-Aware Transmission Scheduling," in *Proc. 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, NV, March 2008.

[10] T. He, "Maximum Throughput of Clandestine Relay: Proof of Selected Theorems," IBM, Tech. Rep., September 2008, http://domino.research.ibm.com/comm/research_people.nsf/pages/ting.pubs.html/$FILE/rc24634.pdf.

[11] G. Arfken, *Mathematical Methods for Physicists*. Orlando, FL: Academic Press, 1985.

[12] S. Marano, V. Matta, and L. Tong, "Detectability of Information Flows under General Renewal Traffic," Sept, 2009, draft.

[13] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[18] There is a subtle difference in our results: [13] found $\beta \approx 0.9$ to fit the interarrival distribution of the traces, whereas we find that to fit the clandestine relay throughput, β should be even smaller