HIDING TRAFFIC FLOW IN COMMUNICATION NETWORKS

BRANKO RADOSAVLJEVIC and BRUCE HAJEK

Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering
University of Illinois, Urbana, Illinois 61801

ABSTRACT This paper considers hiding traffic flow on a communication network from an outside eavesdropper. It is assumed that the eavesdropper can detect transmissions, but cannot understand the encrypted messages being sent. The solution we propose for the network is to choose a transmitter activation schedule independent of the traffic demand, and to use this schedule for all data transfers.

I. INTRODUCTION

Consider a communication network with an outside eavesdropper. Suppose that the eavesdropper can detect the presence of transmissions, but due to encryption cannot understand the messages being sent. Then the eavesdropper can still deduce information about the network. For example, he may be able to:

1) Analyze traffic flow and detect changes in traffic patterns.
2) Determine the identity of a command node.
3) Anticipate and jam acknowledgements.
4) Follow messages across the network.
5) Deduce the original sender of a message.
6) Correlate traffic patterns with other observable activities.

To secure the network, one could use dummy packets and randomized acknowledgement times. More systematically, we propose the following technique for hiding traffic flow from an eavesdropper. (See Figure 1. This is a brief overview; the concepts are defined more precisely in Section II.) Using only the network topology as input, the transmitter schedule generator produces a *transmitter* activation schedule before any network traffic demands are known. From then on, the transmitter activation schedule repeats indefinitely, and is used for all data transfers. Furthermore, if a transmitter is scheduled to transmit, and has no data to send, it sends encrypted dummy messages instead. Later, when the network needs to transmit data, an algorithm called the link decision generator produces a *link* activation schedule (and, if the traffic demand is end-to-end, a set of routes, as well). The link decision generator uses as input the transmitter activation schedule specified in advance, the network topology, and the traffic demand. The resulting link activation schedule must be consistent with the transmitter activation schedule, and it should satisfy the traffic demands efficiently.

The transmitter schedule is chosen in advance because, by assumption, this is what the eavesdropper can detect. However, to actually transfer data, the network must specify a link activation schedule (and, for end-to-end demands, a set of routes). The network is free to choose any link activation schedule that "sounds like" the transmitter activation schedule to the eavesdropper. Then the eavesdropper gets no information about the traffic flow (except possibly upper bounds on capacity), because he always hears the same repeated pattern of transmissions.

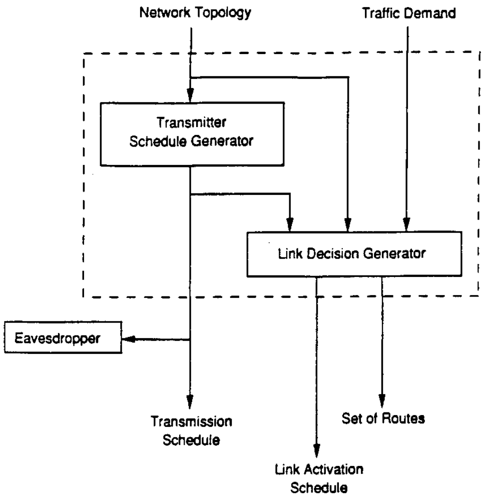

Figure 1: Hiding traffic flow from an eavesdropper.

However, choosing a schedule in advance can lead to a reduction in throughput. For example, ccnsider a radio network in which nodes cannot both transmit and receive at the same time. If every node transmits all the time, then no data will flow, but it is not clear which nodes to turn off, and when. This leads to the following questions:

1) For a given network, what is the best schedule to choose in advance?
2) Using this schedule, what is the worst-case reduction in throughput?
3) Given a prearranged schedule, how can one use it to efficiently satisfy a set of traffic demands?

Our goal is to answer these questions.

Previous techniques to counter traffic flow analysis include the following. These works assume, as we do, that an eavesdrop-

0-7803-0585-X/92 $3.00 © 1992 IEEE

per is able to detect transmissions, and thus do not consider such issues as LPD (low probability of detection) signaling.

- The most basic is to use link-level encryption (i.e., a separate cryptosystem for each communication link), together with bit stuffing, to obtain continuous ciphertext traffic on every link [1,2,3,4,5,6]. However, this approach is not appropriate in a network in which nodes cannot transmit and receive simultaneously, because no information can be transferred if every node is always transmitting.
- Broadcast every message to every node in the network but use a destination-specific deciphering key [6]. Voydock and Kent feel this would require too much bandwidth for a general network.
- Assuming there are many users per node, conceal user-to-user traffic, but not node-to-node [1,6]. This is done by securely encrypting the user tags, but leaving the node addresses available to the intervening nodes (for routing).
- Introduce dummy messages between node-to-node pairs to obtain fixed or random traffic flow [7,6]. However, it seems that these works assume a point-to-point network, i.e., a network for which links can be activated independently of other links. Zavgren [8], in presenting a channel access algorithm for packet radio networks, suggests using randomized acknowledgement times. Randomized acknowledgement times have been proposed for other reasons, as well, such as to avoid message collisions.

Another problem related to traffic flow security is that of covert channels, where an authorized user or process inside the network illegally communicates sensitive information to an unauthorized entity [2,4]. In one scenario, the authorized user communicates by modifying traffic flow through the network, which is then observed by an unauthorized eavesdropper [7]. This problem can be countered using methods presented in this paper.

II. DEFINITIONS

A. Interference models

A communication network is represented by a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of links. There is a link from node a to node b, denoted (a, b), if b can hear a. Given a set of active transmitters, the interference model specifies which sets of links can be simultaneously active. In other words, given a set of active transmitters, the interference model defines a collection of *sets* of links, and all the links contained in any one set in the collection can be simultaneously activated. In addition, we will sometimes consider whether a set of links can be simultaneously active, without specifying which transmitters are active. In this case, we assume that the only active transmitters are the ones at the tail of each active link. (The tail of link (a, b) is node a.) We consider in this paper the following interference models.

1. **Spread spectrum broadcast with perfect capture.** In this model, there is one transmitter per node, and a node is said to transmit when its transmitter is used. A transmitting node always broadcasts, i.e., transmits the same message on all outgoing links. However, we assume that a message can have only one intended recipient. (Every neighbor wants a different message.) Consider an arbitrary node a. If one or more of the nodes that a can hear are transmitting, a can choose exactly one of them from which to receive data. Finally, every node can transmit or receive, but not both simultaneously. This interference model can possibly represent an idealized version of a packet radio network with one transmitter/receiver per node, using spread spectrum signaling with transmitter- or receiver-directed spreading sequences. As another example, this model is consistent with ARNS [9], a communications protocol designed for a multiple satellite network with directional antennas.

   Two terms that have been used to describe packet radio networks are primary and secondary interference. For our purposes, primary interference is said to occur when two active links share a common node. Thus, primary interference occurs if a node attempts any one of the following: (1) to transmit to two nodes simultaneously, (2) to receive from two nodes simultaneously, or (3) to transmit and receive simultaneously. In contrast, secondary interference occurs when an intended recipient of a message can hear an active transmitter other than the desired transmitter. Then stations under the spread spectrum broadcast model are sensitive to primary interference, but not secondary interference.

2. **Narrowband broadcast.** This model is the same as the spread spectrum broadcast model, with the following difference. An arbitrary node a can receive data only if exactly one of the nodes it can hear is transmitting. If two or more of the nodes that a can hear are transmitting, then a receives nothing, due to destructive interfence. This holds whether or not a is the intended recipient. Thus, this model recognizes both primary interference and secondary interference. It can be used to represent an idealized version of a narrowband packet radio network.

3. **Point-to-point networks.** In this model, each communication link has a separate transmitter. A node can transmit and receive different messages on all outgoing and incoming links simultaneously. For example, this could be used to model a hardwired store-and-forward packet network, such as the Internet. For our problem, this model is not very interesting. One can simply activate all the links all the time to hide the actual traffic flow, as has been recommended by many authors.

B. Schedules

We consider two types of schedules — *transmitter* activation schedules and *link* activation schedules. A transmitter activation schedule specifies when each transmitter in a communication network is active. Similarly, a link activation schedule specifies when each link is active. One can obtain a link activation schedule from a transmitter activation schedule by specifying an intended receiver for each active transmitter at each instant of time. A link schedule must satisfy the conditions of the interference model being used.

More precisely, a transmitter activation schedule S^T is an indexed family $S^T = (\lambda_i, T_i : 1 \leq i \leq A)$, where $\lambda_i \geq 0$ for each i, each T_i is a set of transmitters, and A is a finite integer. The λ_i are called activation times and the T_i activation sets. When considering throughput (as opposed to delay), the order in which a schedule's activation sets are specified is not important.

One interpretation of this schedule is to activate the transmitter sets one at a time, activating set T_i for λ_i time units. Other interpretations are possible [10]. Time is assumed to be continuous.[1] The schedule length Λ is defined by

$$\Lambda = \sum_{i=1}^{A} \lambda_i .$$

A link activation schedule $S^L = (\lambda'_i, L_i : 1 \leq i \leq A')$ is defined similarly, but now each set of simultaneously active links L_i must be noninterfering, as defined by the interference model. A link activation schedule S^L is *consistent* with a transmitter activation schedule S^T (denoted $S^L \prec S^T$) if (1) both schedules have the same length, and (2) at each instant of time, the set of active links specified by S^L and the set of active transmitters specified by S^T are compatible, as defined by the interference model.

The difference between transmitter activation schedules and link activation schedules is the crux of this work. In order to hide traffic flow from an eavesdropper, a transmitter activation schedule is specified in advance, because this is what the eavesdropper can detect. However, to actually transfer data, the network uses a link activation schedule. The network can choose any link activation schedule consistent with the transmitter activation schedule chosen in advance.

C. Efficiency

Using a fixed transmitter activation schedule puts constraints on the link activation schedule, and this can lead to a reduction in throughput. The purpose of this section is to define a quantity, called the efficiency, that characterizes the minimum achievable throughput. However, the value of this quantity will depend on the way traffic demands for the network are specified.

There are at least two types of traffic demands: link demands and end-to-end demands. Link demands specify a desired amount of traffic for each link in a network. The actual origins and destinations of the message traffic are unspecified. In contrast, end-to-end demands specify a desired amount of traffic for each origin-destination pair. End-to-end demands, together with a set of routes for each origin-destination pair, generate link demands. Efficiency will be defined for both ways of specifying traffic demands.

Traffic specified by link demands

A set of link demands is represented by a nonnegative vector $r \in R^E$. A link activation schedule $S^L = (\lambda_i, L_i; 1 \leq i \leq A)$ *satisfies* r if

$$\sum_{\{i: e \in L_i\}} \lambda_i \geq r_e \qquad \text{for all } e \in E. \qquad (1)$$

As discussed in [10], there are at least three possible interpretations of S^L. For clarity in this paper, we adopt the following interpretation. The goal is to empty the network when each link e initially has r_e units of data to transfer, measured in units of time. If S^L is a link schedule with length Λ (where $\Lambda = \sum_{i=1}^{A} \lambda_i$), and S^L satisfies (1), then S^L can be used to transfer the data as follows. Activate the link sets one at a time, activating set L_i for λ_i time units. This transfers all the data in Λ time units.

Following are a list of definitions. By assumption, we start with a fixed interference model and a fixed network G. For a link activation schedule S^L and a link demand vector r, define $f_{S^L}(r)$, the rate of S^L for r, to be

$$f_{S^L}(r) = \min_{e \in E} \left(\frac{\Lambda^{-1} \sum_{\{i: e \in L_i\}} \lambda_i}{r_e} \right) .$$

As before, $\Lambda = \sum_{i=1}^{A} \lambda_i$ is the length of S^L. The quantity $f_{S^L}(r)$ is the fraction of r satisfied in one unit of time, averaged over the length of the schedule. For example, if $f_{S^L}(r) = 1/6$, then S^L satisfies 1/6 of r in one unit of time, on the average.

A related quantity is defined for *transmitter* activation schedules. Let S^T be a transmitter activation schedule and r a link demand vector. Then $f_{S^T}(r)$, the max rate of S^T for r, is defined to be

$$f_{S^T}(r) = \max_{S^L \prec S^T} f_{S^L}(r) ,$$

where the maximum is over all link schedules S^L consistent with S^T. Recall that a link schedule S^L is consistent with a transmitter schedule S^T if (1) S^L and S^T have the same length Λ, and (2) at each instant of time, the active links specified by S^L and the active transmitters specified by S^T are compatible, as defined by the interference model. (However, consistent schedules can have different numbers of activation sets.) The definition of $f_{S^T}(r)$ corresponds to fixing S^T and choosing the best S^L consistent with S^T to transmit r.

For comparison, we consider having no constraints on the link activation schedule (other than the constraints dictated by the interference model). For a link demand vector r, define $f(r)$, the max rate for r, to be

$$f(r) = \max_{\text{all } S^L} f_{S^L}(r) .$$

An optimal link activation schedule for r is any shortest-length link activation schedule that satisfies r. If $\Lambda(r)$ is the associated schedule length, then $f(r) = 1/\Lambda(r)$. An optimal *transmitter* activation schedule for r is one that is consistent with an optimal link activation schedule.

Given a transmitter activation schedule S^T and a link demand vector r, define $\eta_G(S^T, r)$, the efficiency of S^T for r, to be

$$\eta_G(S^T, r) = \frac{f_{S^T}(r)}{f(r)} .$$

The quantity $\eta_G(S^T, r)$ is used to compare the maximum throughput with S^T fixed to the maximum throughput without this constraint.

Next, we consider the worst possible performance for a fixed transmitter activation schedule S^T. Define $\eta_G(S^T)$, the min efficiency of S^T, to be

$$\eta_G(S^T) = \min_{r} \eta_G(S^T, r) .$$

Finally, define η_G, the max-min efficiency for G, to be

$$\eta_G = \max_{S^T} \eta_G(S^T) .$$

[1] Data is assumed to be infinitely divisible. This is not a strictly valid assumption to make, but it may be a good approximation if the schedule activation times are large compared to the minimum data size.

The quantity η_G is a function of only the interference model and the network G. It represents the fundamental cost of using a fixed transmitter activation schedule. By choosing S^T to be the most efficient transmitter activation schedule, the achievable efficiency $\eta_G(S^T, r)$ will always be greater than or equal to η_G.

Traffic specified by end-to-end demands

A set of end-to-end demands is represented by a vector $q \in R^{V \times V}$. In this case, to talk about satisfying a set of end-to-end demands, we must define a set of routes, as well. For each origin-destination pair, a set of routes R specifies which paths are used to send data from the origin to the destination, and what fraction of the data traffic is sent along each path. Then a set of routes R coupled with a set of end-to-end demands q generates a set of link demands r. We denote the resulting link demand vector r as Rq, to indicate the linear dependence of r on q, for R fixed. The value of r_e for a link e is the sum of the traffic associated with all the paths passing through e.

It is assumed we start with a fixed interference model and a fixed network G. In addition, it is assumed that all links have equal capacity. For a link activation schedule S^L and an end-to-end demand vector q, define $g_{S^L}(q)$, the rate of S^L for q, to be

$$g_{S^L}(q) = \max_R f_{S^L}(Rq) ,$$

where the maximum is over all sets of routes R, and f_{S^L} is the corresponding quantity for link demands, defined previously. The remaining quantities ($g_{S^T}(q)$, $g(q)$, $\eta_G(S^T, q)$, $\eta_G(S^T)$, and η_G) are defined analogously to those defined for link demands, substituting $g_{S^L}(q)$ for $f_{S^L}(r)$.

It can be shown that $[\eta_G]_{e-e} \geq [\eta_G]_{link}$, where "e-e" stands for end-to-end demands and "link" for link demands. An intuitive explanation of this result is that end-to-end demands have the added flexibility of choosing routes for the data. This leads to greater efficiency than with fixed-route link demands.

III. EXAMPLE 1 — THE SPREAD SPECTRUM BROADCAST MODEL

A. Link demands

This subsection contains results concerning the spread spectrum broadcast model used with link demands. This is indicated by the letters "SL". Our results for narrowband broadcast models are not included in this paper due to space limitations.

SL-1. If any node has both an incoming and an outgoing link, then $\eta_G \leq 1/2$. Otherwise, $\eta_G = 1$. Thus, if there are any bidirectional links, then $\eta_G \leq 1/2$.

SL-2. For any transmitter activation schedule S^T,

$$\eta_G(S^T) = \min_{r \in \mathcal{I}} \eta_G(S^T, r) , \qquad (2)$$

where $\mathcal{I}$ is the set of all link demand vectors of the form X_I and X_I is the indicator vector of an independent (non-interfering) set of links. (For the current interference model, I is independent if all the links in I are node disjoint.) This narrows the set of link demands we must consider to compute $\eta_G(S^T)$.

SL-3.

$$\eta_G(S^T) = \min_{(a,b) \in E} \beta(a, b) ,$$

where $\beta(a, b)$ is the fraction of time node a is on and node b is off using S^T. This formula for $\eta_G(S^T)$ is easy to compute.

SL-4. For every network G with n nodes,

$$\eta_G \geq \begin{cases} (1/4)\,[1 + 1/(n-1)] & n \text{ even} \\ (1/4)(1 + 1/n) & n \text{ odd} \end{cases} \qquad (3)$$

SL-5. For complete (bidirectional) graphs, (3) holds with equality. Thus, for a fixed number of nodes, the complete graph has the least efficiency.

SL-6. For any bipartite graph, and thus for any tree, $\eta_G \geq 1/2$. In general, η_G is lower bounded by $1/\chi_f(G)$, where $\chi_f(G)$ is the fractional chromatic number of G (and thus a similar result holds for the usual chromatic number as well), though this is a weak bound.

SL-7. Using hash functions [11], we can construct a transmitter schedule S^T, with less than or equal to $2n$ activation sets, that achieves

$$\eta_G(S^T) \geq (1/4)(1 + 1/2n) .$$

SL-8. Let r_{eq} denote the link demand with one unit of traffic on each link. Then it is possible for a schedule to be r_{eq}-optimal but not max-min optimal, and vice versa.

B. End-to-end demands

This subsection contains results concerning the spread spectrum broadcast model used with end-to-end demands. This is indicated by the letters "SE".

SE-1. Some results carry over from the link demand case. First, $\eta_G \leq 1/2$ if any link in G has both an incoming and an outgoing link, and $\eta_G = 1$ otherwise. Second, for every transmitter schedule S^T,

$$\eta_G(S^T) = \min_{q \in \mathcal{I}} \eta_G(S^T, q) ,$$

where, in this case, $\mathcal{I}$ is the set of all end-to-end demands that "look like" the indicator vector of a set of independent links. That is, each q in $\mathcal{I}$ is the sum of one-unit demands from nodes a_i to b_i, where each $(a_i, b_i) \in E$, and the set of links $\{(a_i, b_i)\}$ are independent. Third, as discussed on p. 4, the link value of η_G lower bounds the end-to-end value, so SL-3, SL-4, and SL-6 give lower bounds to η_G.

SE-2. For a complete graph with n nodes,

$$\eta_G = \begin{cases} \frac{5}{12}\left[1 + \frac{1}{5n(n-1)}\right] & \text{even } n \\ \frac{5}{12}\left[1 + \frac{5n-1-4nc_n}{5n(n-1)}\right] & \text{odd } n \geq 13 \end{cases} \qquad (4)$$

where $c_n = 2^m \Big/ \binom{n}{m}$ and $m = \lfloor n/2 \rfloor$. In addition, for all odd n,

$$\frac{5}{12}\left(1 + \frac{1}{5n}\right) \leq \eta_G \leq \frac{5}{12}\left[1 + \frac{5n-1}{5n(n-1)}\right] .$$

For both even and odd n, a transmitter schedule that achieves (4) is the *50-50 schedule*, which activates each set of $\lfloor n/2 \rfloor$ nodes and each set of $\lceil n/2 \rceil$ nodes for an equal amount of time.

SE-3. For a fixed set of nodes A, the *volley schedule* activates first A, then $V \setminus A$, both for an equal amount of time. For a complete graph with n nodes, the volley schedule, using any A with $|A| = \lfloor n/2 \rfloor$, achieves $\eta_G(S^T) = 1/3$.

SE-4. For a bidirectional ring with n nodes,

$$\eta_G = \begin{cases} 1/2 & n \text{ even or } n = 3 \\ (n-1)/2n & \text{odd } n \geq 5 \end{cases}$$

In particular, $\eta_G = 0.4$ for a pentagon. Thus, in this case, complete graphs do not have the minimum efficiency.

SE-5. There exist graphs for which the set of transmitter activation schedules that are max-min optimal (achieve $\eta_G(S^T) = \eta_G$) for end-to-end demands is disjoint from the set of transmitter activation schedules that are max-min optimal for link demands.

IV. OTHER CONSIDERATIONS

This section contains some miscellaneous items about choosing a schedule in advance to maintain traffic flow security.

An encryption scheme used for traffic hiding should have certain properties. First, if a fixed dummy packet is used, it should not always be mapped to the same ciphertext or its function will become obvious. This behavior can be accomplished by changing keys, using block chaining (Cipher Block Chaining mode of the Data Encryption Standard), or using a stream cipher with feedback key generation [4,6]. Second, as a packet moves from one node to the next, if its ciphertext representation remains the same, the eavesdropper may be able to follow this pattern of bits, even if he does not understand them. Thus, there should be some link-level encryption to change the ciphertext from hop to hop. However, this does not eliminate the need for end-to-end encryption. As discussed in [6], relying solely on link-level encryption can be dangerous, because this requires that every node on a message's path be secure. Thus, it may be appropriate to use end-to-end encryption, together with a simple link-level encryption layered on top to obscure the data flow.

Another point to consider is that some people may feel uneasy choosing a schedule in advance and using it repeatedly. If the eavesdropper discovers the schedule, he can use this knowledge to decide when to jam. However, there are two replies to this concern. First, the eavesdropper will not know whether he is jamming data or dummy messages, so his jamming may not be efficient after all. Second, even a simple schedule can be manipulated to obtain an equivalent schedule with a very long period that effectively appears random. Recall that for purposes of throughput, a schedule is just a scale-invariant weighting on the set of possible activation sets. The period of time allotted to an activation set can be broken up and shuffled around with that of the other activation sets, and the entire mixture can be expanded to have a very long period. However, these changes are not equivalent with respect to delay. Nevertheless, the resulting mixed schedule can be created in such a way that each small part has good delay characteristics.

A final point to consider is that it may be desirable to choose a schedule that is independent of the network topology. (The schedule would depend only on the number of nodes in the network.) Then knowledge of the topology would be denied to the eavesdropper. For the spread spectrum broadcast model, the hash function schedule described in SL-7 is independent of the topology, and has worst-case efficiency greater than 1/4 (for both link and end-to-end demands).

V. SUMMARY

In this paper, we presented a method to hide traffic flow on a communication network from an outside eavesdropper. The network uses a fixed transmitter activation schedule that gives no information about the traffic demand. For the spread spectrum broadcast model, we analyzed the cost of using this scheme, as reflected in a possible reduction of throughput.

In work not presented here, we have studied the performance of this scheme on the narrowband broadcast model. In addition, we have considered the problem of choosing an efficient link activation schedule consistent with a given transmitter activation schedule, as a function of the network traffic demand.

ACKNOWLEDGEMENT

This research was supported in part by the Joint Services Electronics Program under Grant No. N00014-90-J-1270.

References

[1] R. Benjamin, "Security considerations in communications systems and networks," *IEE Proceedings, Pt. I*, vol. 137, pp. 61–72, April 1990.

[2] J. A. Cooper, *Computer and Communications Security.* McGraw-Hill, 1989.

[3] D. W. Davies and W. L. Price, *Security for Computer Networks.* John Wiley and Sons, 1984.

[4] D. E. Denning, *Cryptography and Data Security.* Addison-Wesley, 1982.

[5] E. H. Lipper, B. Melamed, R. J. T. Morris, and P. Zave, "A multi-level secure message switch with minimal TCB: Architectural outline and security analysis," in *Fourth Aerospace Computer Security Conference*, pp. 242–249, 1988.

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

[7] A. B. Jeng and M. D. Abrams, "On network covert channel analysis," in *AIAA/ASIS/IEEE Third Aerospace Computer Security Conference*, pp. 95–103, 1987.

[8] J. R. Zavgren, "The moment-of-silence channel-access algorithm," in *IEEE Military Communications Conference (MILCOM)*, 1989.

[9] R. P. Kosowsky, I. M. Jacobs, and K. S. Gilhousen, "ARNS: A new link layer protocol," in *IEEE Military Communications Conference (MILCOM)*, 1988.

[10] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Transactions on Information Theory*, vol. 34, pp. 910–917, September 1988.

[11] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *Journal of Computer and System Sciences*, vol. 18, pp. 143–154, 1979.