A Quantitative Analysis of Anonymous Communications

Yong Guan, Xinwen Fu, Riccardo Bettati, Wei Zhao

Abstract— **This paper aims to quantitatively analyze anonymous communication systems with regard to anonymity properties. Various anonymous communication systems have been designed and implemented. However, there are few formal and quantitative analyses on how these systems perform. System developers often informally argued the security goals which their systems can achieve. Such results were likely vague and not persuasive. In this paper, we use a probabilistic method to investigate the anonymity behavior of anonymous communication systems. In particular, we study the probability that the true identity of a sender can be discovered in an anonymous communication system given that some nodes have been compromised. It is through this analysis that we can identify a number of design guidelines for systems aimed at providing communication anonymity. For example, contrary to what one would intuitively expect, our analytic results show that the probability that the true identity of a sender can be discovered may not always decrease as the length of communication path increases. We also found that the complexity of path topology does not have significant impact in terms of anonymity behavior.**

Keywords—**Anonymous communication, network security, sender/receiver anonymity, rerouting.**

I. INTRODUCTION

THIS paper aims to quantitatively analyze anonymous communication systems (ACS) with regard to anonymity properties. With the rapid growth and public acceptance of the Internet as a means of communication and information dissemination, concerns about privacy and security on the Internet have grown. Protection of *privacy* and *anonymity* becomes a basic requirement for many on-line Internet applications such as E-Voting, E-Banking, E-Commerce, E-Mail, E-Auctions, and many others. Anonymity, in particular is important to protect, as this greatly simplifies protection of privacy. Many anonymous communication systems have been developed, which protect the identity of the participants in various forms: sender anonymity protects the identity of the sender, while receiver anonymity does this for the receiver. Mutual anonymity guarantees that both parties of a communication remain anonymous to each other at the same time. Finally, some systems provide unlinkability of sender and receiver, that is, they guarantee that nobody could learn the communication relation between the sender and receiver except the sender and receiver themselves. Among the various forms of anonymity, sender anonymity is most demanded in the current Internet applications. In this paper, we will therefore focus primarily on sender anonymity.

Sender anonymity on the Internet is most commonly achieved by transmitting the message to its destination through one or more intermediate nodes in order to hide the true identity of the sender. The message is effectively *rerouted* and the path that the message traverses is called *rerouting path*. In this paper, we study rerouting-based anonymous communication systems in terms of their ability to provide sender anonymity. Selection of rerouting paths is critical for rerouting-based anonymous communication systems. The key issues in path selection are (1) how to choose path length and (2) how to choose path topology. We study how different path selection strategies affect the ability to provide sender anonymity. We measure this ability as the probability that the true identity of a sender can be discovered. This measure can then be applied to quantitatively evaluate the various anonymous communication systems.

We investigate the anonymity behavior of these systems under a *passive* adversary model: An adversary can compromise *one or more* nodes, which are rerouting-capable hosts (e.g., onion router in Onion-Routing [13], [31], [33], [34] and jondo in Crowds [26]). An adversary agent at such a compromised node can gather information about messages that traverse the node. In particular, it may be able to discover and report the immediate predecessor and successor node for each message traversing the compromised node. We assume that the adversary collects all the information from those compromised nodes and attempts to derive the identity of the sender of a message.

Through our quantitative analysis, we identified several insightful results:

- Contrary to what people would intuitively expect, the system may not be safer as path length increases. That is, the probability that the true identity of a sender can be discovered may not always decrease as the path length increases.
- When the expected path length is sufficiently large, the difference of probabilities that the true sender can be discovered under different path length distributions is relatively small.
- The complexity of path topology does not have significant impact on this probability. While complicated paths perform better than simple paths, the difference is relatively small.
- As expected, the probability that the true identity of a sender can be discovered increases when the number of compromised nodes in the system increases. In addition, the deterioration of the system behavior goes sublinearly as the number of compromised nodes increases.

As a result of this study, we found several well-known anonymous communication systems are not using the best path selection strategies. We therefore believe that our results are greatly helpful for the current and future development of anonymous communication systems.

The remainder of this paper is organized as follows. Section II gives an overview of the anonymous communication systems. In Section III, we present the system model and discuss the key issues in path selection for anonymous communication systems. In Section IV, we describe the threat model and discuss how the adversary determines the identity of the sender of a message under this model. In Section V, we report our analytical and numerical results. Finally, in Section VI, we present our conclu-

Last modified by Yong Guan on August 1, 2001.

The authors are with the Department of Computer Science, Texas A&M University, College Station, TX 77843-3112. E-mail:{yguan, xinwenfu, bettati, zhao}@cs.tamu.edu.

sions.

II. Overview of Anonymous Communication Systems

In this section, we survey the past work related to anonymity, including DC-Net [5], [39], Mixes [4], [18], [20], Anonymizer [1], Anonymous Remailer [2], LPWA [10], Onion Routing [13], [31], [33], [34], Crowds [26], Hordes [28], Freedom [12], and Pipenet [8].

Many anonymous communication systems have been designed and implemented that provide various types of anonymity, such as sender anonymity, receiver anonymity, and mutual anonymity, unlinkability of sender and receiver, or combinations thereof. As mentioned in Section I, sender anonymity is typical and most demanded in current Internet applications. Systems providing sender anonymity can be categorized into two classes: *rerouting-based* systems and *non-rerouting-based* systems.

To the best of our knowledge, DC-Net [5] is the only non-rerouting-based anonymous communication systems. In DC-Net, each participant shares secret coin flips with other pairs and announces the parity of the flips the participant has seen to all other participants and the receiver. The total parity should be even since each flip is announced twice. By incorrectly stating the parity the sender has seen, she can send a message to the receiver. The receiver gets the message if she finds that the total parity is odd. Nobody except the sender herself knows who sends it. The advantage of DC-Net is that it does not introduce extra overhead like the longer rerouting delay and extra amount of rerouting traffic in rerouting-based system. However, due to its lack of scalability in practice, none of the current on-line applications employ this method.

Most anonymous communication systems used widely on the Internet employ rerouting of a message through a number of intermediate nodes from the source to the destination in order to hide the true identity of the sender. Even though rerouting introduces extra delay and typically increases the amount of traffic due to longer routes, it is a scalable and practical approach when such overheads are within tolerable limits for most online Internet applications. In this paper, we will focus on the analysis of rerouting-based anonymous communication systems. In the following, we will briefly overview a number of such communication systems, categorizing them according to their path selection strategies.

Anonymizer [1] provides fast, anonymous, interactive communication services. It is essentially a web proxy (called *Anonymizer Server*) that filters out the identifying headers and source addresses from web browsers' requests. Instead of user's true identity (e.g., IP address), a web server can only learn the identity of the Anonymizer Server. In this approach, all rerouting paths have a single intermediate node, which is the Anonymizer server.

Anonymous Remailer [2] is mainly used for email anonymity. It employs rerouting of an email through a sequence of multiple mail remailers and then to its recipient such that the true origin of the email can be hidden. These mail remailers are scattered around the world.

Onion-routing [13], [31], [34], [33] provides anonymous Internet connection services. It builds a rerouting path within a network of onion routers, which in turn are similar to real-time Chaum Mixes [4]. The basic idea of mixes is that each message is sent over a series of independent stations. A mix is a store-and-forward device that accepts a number of fixed-length messages from different sources, discards repeats, performs cryptographic transformation on the messages, and then outputs the message to the next destination in an order not predictable from the order of inputs.

In Onion Routing, the data packet is broken into fixed-size cells (128bytes long, at present) and each cell is encrypted multiple times (once for each onion router on the path). Thus, a recursively layered data structure called *onion* is constructed. Onion is the packet transmitted along the rerouting path. The fixed size of an onion limits a route to a maximum of 11 nodes in the current implementation.

Onion Routing I [31], [34], [33] uses a network of five Onion Routing nodes operating at the Naval Research Laboratory. It forces a fixed length (five hops, i.e., 4 intermediate onion routers) for all routes.

Onion Routing II [33] can support a network of up to 50 core Onion Routers. For each rerouting path through an onion routing network, each hop is chosen at random. The rerouting path may contain cycles, although only cycles with one or more intermediate nodes are allowed. The number of intermediate nodes in a route, L, is chosen based on repeated flips of a weighted coin. This idea is borrowed from Crowds [26]. The expected route length is completely determined by the weight of flipping the coin (called *forwarding probability*).

Crowds [26] aims at protecting the users' web-browsing anonymity. Like Onion Routing, the Crowds protocol uses a series of cooperating proxies (called jondo) to maintain anonymity within the group. Unlike Onion Routing, the sender does not determine the whole path. Instead, the path is chosen randomly on a hop-by-hop basis. Each hop decides whether to submit the request directly to the end server or to forward it to another randomly-chosen member according to forwarding probability. The expected path length is controlled by the forwarding probability. Cycles are allowed on the path. The receiver is known to any intermediate node on the route. Once a path out of a crowd is chosen, it is used for all the anonymous communication from the sender to the receiver within a 24-hour period. At some specific time instant, new members can join the crowd and new paths can be formed.

Freedom Network [12] also aims at providing anonymity for web browsing. From the user's point of view, Freedom is very similar to Onion Routing. Freedom consists of a set of nodes (called *Anonymous Internet Proxy*) which run on top of the existing Internet infrastructure. To communicate with a web server, the user first selects a series of nodes to form a rerouting path, and then uses this path to forward the requests to its destination.

The Freedom Route Creation Protocol allows the sender to randomly choose the path, but the path length is fixed to be 3 [35]. The Freedom client user interface does not allow the user to specify a path containing cycle. The Freedom client is either to have all the intermediate nodes in the path chosen, or to choose a preferred first node and last node, and the intermediary nodes are picked at random.

Hordes [28] employs multiple jondos similar to those used in the Crowds protocol to anonymously route a packet towards the receiver. For the forwarding path, Hordes uses the same strategy as Crowds. However, the major difference is that it uses multicast services to anonymously route the reply to the sender instead of using the reverse path of the request. Same as Crowds, Hordes also allows the cycle on the forwarding path and the expected forwarding path length is determined by the forwarding probability.

Similar to Anonymizer, *Lucent Personalized Web Assistant* (LPWA) [10] uses a single proxy server that accepts connections from the sender of an anonymous connection and forwards them on to the host that the sender wishes to contact anonymously. All the receiver can learn is the proxy's address, thus the sender address is hidden from the receiver. The proxy itself can determine the sender's identity. It is obvious that LPWA uses the rerouting path with only one intermediate node which is same as Anonymizer.

PipeNet [8] is a simple anonymous protocol. It is based on the idea of virtual link encryption. Before the sender starts to send the data, it establishes a rerouting path (called anonymous connection in PipeNet). The rerouting path is a path in the network, along which the first node is the sender, the last node is the receiver, and the rest nodes are the intermediate nodes. Each node in the path shares a key with the sender and knows the nodes immediately in front of and behind it. Each node also knows how many hops away it is from the sender. PipeNet always generate a rerouting path with 3 or 4 intermediate nodes.

III. System Model and Key Issues in Path Selection

Our system model used in the following discussion is an abstraction of the systems mentioned above. It will therefore lend itself well to discuss the key issues in path selection for rerouting-based anonymous communication systems.

A. System Model

A rerouting-based anonymous communication system consists of a set of N nodes ($\{0, 1, \cdots, N-1\}$) which collaborate with each other to achieve anonymity. We assume that the receiver is always compromised and is not part of the N nodes. As we consider an IP-based network, we assume that an underlying routing subsystem determines a unique path between any pairs of hosts. For our purposes, we model the network at transport layer and assume that every host can communicate with every other host. The network therefore can be modeled as a fully-connected graph. All the edges in this graph represent the paths from the source host to destination host through some routers in the network. To hide the true identity of the sender, the message is transmitted from its source to its destination through one or more intermediate nodes.

Figure 1 shows a system of 16 nodes, where Node 0 is the sender of a message. The message is transmitted along a rerouting path $< 0, 5, 2, 7, 11, 8, R >$ determined by anonymous communication system and finally arrives at Node R. The path length[1] is 5.

[1] In this paper, the path length is defined as the number of intermediate nodes on the rerouting path.

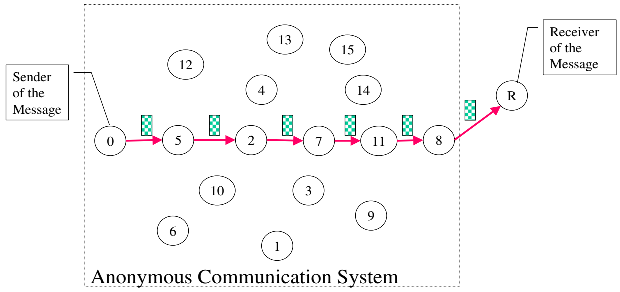

Fig. 1. System Model

B. Key Issues in Path Selection

The key issues in path selection for rerouting-based anonymous communication systems are (1) how to choose the length of rerouting path (path length) and (2) how to choose the sequence of intermediate nodes on the path (path topology).

B.1 Choosing Path Length

Choosing path length is one critical decision in the design and implementation of anonymous communication systems. There are two kinds of strategies that can be used: fixed length and variable length. In the case of variable length, the path length is a random variable conforming to a specific probability distribution. The following shows several strategies already used by existing systems such as Crowds, Onion-routing, PipeNet, Freedom, and others.

Algorithm L1 : Fixed Length. Whenever a sender wants to send a message to some receiver, the anonymous communication system always chooses a fixed number of intermediate nodes to form the rerouting path to transfer the message. Onion-Routing I and Freedom use this method.

Algorithm L2 : Variable Length with geometric path length distribution. The path length is randomly chosen as a non-negative number according to the following distribution, for $x \geq 0$,

$$P(L = x) = (1 - p_f)p_f^x, \tag{1}$$

where p_f is the forwarding probability that controls the expected length of the rerouting path. The expected path length is $\frac{p_f}{1-p_f}$. Crowds and Onion-Routing II use this method. In practice, the path length cannot be unbounded, since the user's on-line activity always has performance requirement in term of response time. If the path length is to be bounded, i.e., the above probability distribution will have to be modified as follows:

$$P(L = x) = \begin{cases} (1 - p_f)p_f^x, & 0 \leq x < MAXLEN; \\ p_f^{MAXLEN}, & \text{otherwise}, \end{cases} \tag{2}$$

then the expected path length is $\frac{p_f(1-p_f^{MAXLEN})}{1-p_f}$.

Algorithm L3 : Variable Length with uniform path length distribution. The path length is randomly chosen as a non-negative number between a and b according to the uniform distribution, i.e., for $a \leq x < b$,

$$P(L = x) = \frac{1}{b - a}, \tag{3}$$

where $a \geq 0$ and $b > a$. The expected path length for this uniform distribution is $\frac{b-a}{2}$.

It is up to the system developer to determine the type of path length (fixed or variable) and its parameters. In this process, the following issues must be addressed:
1. If paths of fixed length are used, what path length should be used?
2. If paths of variable length are used, what probability distribution for the path length should be used?
3. It is usually expected that the variable length should be much better than fixed length. Is this always true?

This paper evaluates path length selection methods and have these questions answered.

B.2 Choosing Path Topology

Once the path length has been chosen, the rerouting path must be chosen by selecting intermediate nodes. Depending on the constraints posed on the choice for intermediate nodes, the paths can be classified as *simple paths* and *complicated paths*. In a simple path, cycles are not allowed. That is, a node can be chosen on a rerouting path at most once. In a complicated path, a node can be chosen on a rerouting path more than once. In this paper, we consider the following methods:

Algorithm T1 : the rerouting path doesn't allow cycles. No intermediate node can show up in more than one different places on the rerouting path.

Algorithm T2 : the rerouting path allows *disjoint* cycles, that is, cycles that do not share common nodes. Cycles must be non-reflective: only the intermediate node as the starting and ending point of a cycle can show up exactly two times.

Algorithm T3 : the rerouting path allows arbitrary non-reflective cycles. Any intermediate node can show up arbitrary times at any places. We do not consider reflective cycles, because they have no effect on the adversary's ability to infer the sender identity.

Clearly, algorithm T1 can be used to construct the simple path. Algorithms T2 and T3 are for the complicated path.

When it comes to selecting good paths, the following issues must be addressed:
1. Do complicated paths always perform better than simple paths?
2. Does the requirement in T2 that cycles be disjoint have an effect on the capability to provide anonymity?

Table I shows a combination of strategies for path selection algorithm. We will study how these different strategies impact on the performance behavior of anonymous communication systems in the following sections.

TABLE I
COMBINATION OF PATH SELECTION STRATEGIES

Path Length and Topology	Simple Path	Complicated Path
Fixed Length	Case A	Case C
Variable Length	Case B	Case D

IV. THREAT MODEL AND ADVERSARY ALGORITHM

In this section, we first present the threat model and then discuss how the adversary determines the identity of the sender of a message under this model.

A. Threat Model

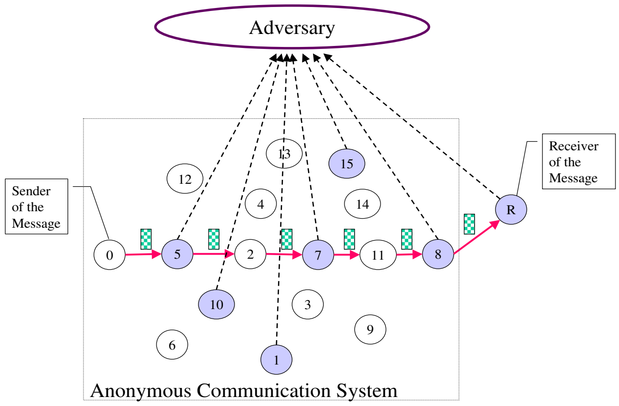

Fig. 2. Threat Model

In this paper, we consider a *passive* adversary model: There is an adversary who intends to detect the identity of senders passively by monitoring packets. In order to have access to packets, the adversary has previously compromised one or more nodes, and now is able to monitor the activity of the compromised nodes in the way we describe below. In our model, we assume that an anonymous communication system consists of N nodes, in which there are M compromised nodes. In addition to the M compromised nodes, the receiver is compromised as well. An agent of the adversary at a compromised node can discover and report the immediate predecessor and successor node for each message traversing the compromised node. We assume that the adversary collects all the information from those compromised nodes and uses it to derive the identity of the sender of a message. In Figure 2, the dark nodes $1, 5, 7, 8, 10, 15$, and R are compromised nodes. A message is transmitted along the rerouting path $< 0, 5, 2, 7, 11, 8, R >$ determined by a path selection algorithm of the anonymous communication system from its source 0 to its destination R. In this case, the true sender 0 can definitely be determined by the adversary.

Our analysis is based on the worst-case one.

- When the sender constructs the rerouting path, it does not know which nodes have been compromised. Thus, some compromised nodes may be on the rerouting path.
- On the other hand, the adversary has all the knowledge about path selection algorithm.
- The adversary collects all the information from the agents on the compromised nodes and tries to derive the true identity of the sender.
- To make the case even worse, we assume that there is one and only one message transmitted through the rerouting path from its source to its destination. Therefore, the adversary does not have to worry about the inconsistent information collected by its agents.

In previous evaluations of anonymous communication systems, various attack models were assumed [26], [28], [33]. As it turns out, Many of these models are special cases of our model described above. For example:

- Attack by observing respondent [28] and end server [26]: This corresponds to the case in our model that the receiver is compromised.
- Attack by active and passive path traceback [28] and collaborating jondos [26]: This corresponds to the case in our model that the rerouting path can be reconstructed by the attacker using the routing information or other monitoring information and at least $\frac{L}{2}$ nodes on the rerouting path are compromised (L is the path length).
- Attack by local eavesdropper [26], [28]: This corresponds to the case in our model that the sender of the message is compromised.

In the following, we discuss how the adversary derives the identity of the sender in the case of the fixed and variable path lengths.

B. Overview of Adversary Algorithm for Fixed Path Length

The Adversary Algorithm aims at deriving the true identity of the sender of a message. Figure 3 shows the framework for adversary algorithm for fixed path lengths. In the first two steps, the adversary collects the information about the path selection algorithm and its parameters. Following this, the adversary tries to eliminate the nodes that could not be the sender. Finally, the adversary algorithm returns a node S from the remaining nodes as the source of a message.

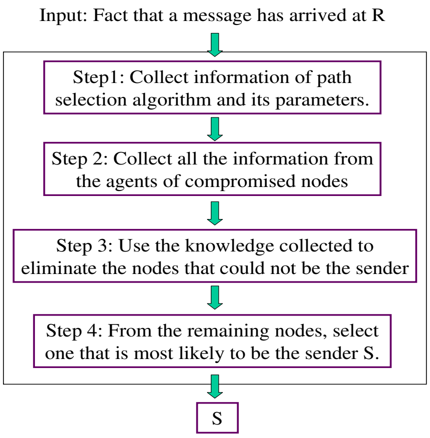

Fig. 3. Framework for Adversary Algorithm (Fixed Path Length)

We use the probability that the true sender can be discovered (i.e., $P(S = s)$) as the measure to evaluate the anonymity behavior of anonymous communication systems.

C. Details of the Adversary Algorithm for Fixed Path Lengths

In this section, we discuss the details of each step in the adversary algorithm shown in Figure 3.

Step 1: Collecting information of path selection algorithm and its parameters. In reality, the adversary may have all or part of the knowledge about the path selection algorithm and its parameters for the anonymous communication system under attack. In the following, we assume that the adversary has full knowledge about the path selection algorithm. That is, the path selection algorithm and its parameters are known to the adversary [2].

Step 2: Collecting information from the agents residing on compromised nodes. As discussed earlier, an adversary agent at a compromised node can discover and report the immediate predecessor and successor node for each message traversing the compromised node. We assume that the adversary collects all the information from those compromised nodes.

Step 3: Using the knowledge collected to eliminate the nodes that could not be the sender. Based on the information collected in the first two steps, the adversary can eliminate nodes from the set of possible senders of the message by applying the following rules. We use the *non-sender set* (NS) to represent the set of all those nodes that could not be the true sender. In this step, the non-sender set is constructed by adding all the nodes that could not be the sender. The remaining nodes form the set of potential senders RS. The nodes in RS cannot be added into non-sender set for certain. At any point in time, the true sender must be in the set RS. It is obvious that

$$|NS| + |RS| = N. \tag{4}$$

In the following, we discuss how to construct the non-sender set NS under different path selection strategies given that a fixed-path-length approach is used.

1. Elimination rules for the case of simple path

R1: The compromised node itself can be added in non-sender set if it is not the true sender.

R2: The node immediately after the compromised node on the rerouting path of a message cannot be the sender.

R3: The node immediately before the receiver cannot be the sender if the path length is larger than 0 and the receiver is compromised and collaborates with the adversary.

R4: All the nodes that the adversary can find after the first compromised node cannot be the sender.

2. Elimination rules for the case of complicated path (with one or more disjoint cycles)

R5: The compromised node itself can be added in non-sender set if it is not the true sender.

R6: If the cycle can be definitely determined by the adversary, that is, the cycle length and cycle start and end point are known to the adversary, then this becomes the shorter path case in Elimination Rules for the Case of Simple Path.

R7: If the cycle length is unknown but cycle start and end point are known, then

R7.1: The node immediately after the compromised node on the rerouting path of a message cannot be the sender.

R7.2: The node immediately before the receiver cannot be the sender if the path length is larger than 0 and the receiver is compromised and collaborates with the adversary.

R8: If neither cycle length nor cycle start and end point are known, then this case becomes identical to the case of Elimination Rules for the Case of Complicated Path (arbitrary path without disjoint constraints).

3. Elimination rules for the case of complicated path (arbitrary path without disjoint constraints)

R9: If the rerouting path can be reconstructed completely based on the adversary's observation and knowledge, the identity of

[2] This assumption follows common practice: it makes little sense to assume that encryption algorithms are not known to the adversary. The only thing the adversary does not know is the seed of the random number generation used by the path selection algorithm.

the sender can always be correctly identified.
R10: Otherwise, only the compromised nodes are sure not to be the sender.

Using these rules, we can distinguish which nodes can not be the sender, and so construct the non-sender set NS. This set is complete, i.e., no other nodes can be added into the non-sender set.

Step 4: Selecting one node S that is most likely to be the sender from the remaining node set RS.

In Step 3, the set of potential senders RS was constructed. In Step 4, the adversary tries to derive which node is most likely to be the true sender and returns this node as the sender of the message.

In some cases, due to the physical location of the sender and of the compromised nodes or some other information, the adversary may have the knowledge in term of *a priori* probability that each node in the system can be the true sender. That is, the adversary may believe that each node may have a different *a priori* probability p_i to be the sender. Thus we can calculate the probability that the true sender can be discovered as follows,

$$P(S = s|F) = \frac{p_s}{\sum_{i \in RS} p_i}. \quad (5)$$

In most cases, the adversary does not have such knowledge. So in the remainder of this paper, without loss of generality, we assume that any node in RS can be the sender of a message with equal probability. Thus, the probability that the identity of the sender can be discovered is.

$$P(S = s|F) = \frac{1}{|RS|} = \frac{1}{N - |NS|}. \quad (6)$$

D. Extension to Variable Path Length

As we discussed in Section III, the path length can be a random number conforming to some specific probability distribution. We now extend the adversary algorithm for fixed path lengths in Figure 3 to this case.

The following theorem is central in describing the adversary's ability to identify the sender for variable length paths.

Theorem 1: If the path selection algorithm in an anonymous communication system chooses the variable path length between the two values a and b according to some specific distribution, i.e., $P(L = i)$ for $a \leq i \leq b$, and the adversary can learn that the actual path length can be any value between c and d by analyzing the information F collected from the agents on the compromised nodes, where $a \leq c \leq d \leq b$, the probability that the identity of the sender can be discovered is given as follows:

$$\begin{aligned} P(S = s|F) &= \sum_{i=c}^{d} (P(S = s|L = i, F) \\ &\quad \times \frac{P(L = i)}{\sum_{k=c}^{d} P(L = k)}). \end{aligned} \quad (7)$$

The proof of Theorem 1 can be found in the Appendix.

Based on Theorem 1, the major steps in the framework for the Adversary Algorithm for variable length paths is given in Figure 4. The difference between fixed-length Adversary algorithm and variable-length Adversary Algorithm is that the adversary only knows the path length distribution and maximum possible and minimum possible path length c and d in variable-length Adversary algorithm instead of the exact path length in fixed-length adversary algorithm.

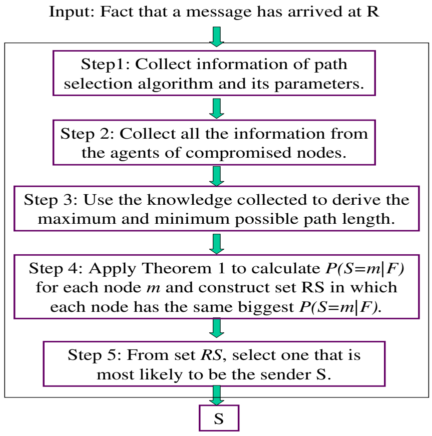

Fig. 4. Framework for Adversary Algorithm (Variable Path Length)

V. Security Analysis

In this section, we analyze the impact of path selection strategies on the probability that the true identity of the sender can be discovered. We first present analytical results. We will then discuss observations on the numerical results.

A. Analytical Results

We will start our analysis with a conditional probability. We will then present the general results followed by closed-form formulas for some special cases.

A.1 A Conditional Probability

Here, we derive the conditional probability that the identity of the true sender is discovered given that there are exactly i compromised nodes on a randomly selected simple path.

First, we note that for simple paths, there are $\binom{L+1}{i}$ instances of simple paths that have length of L and contain exactly i compromised nodes [3]. For example, if L = 4 and i = 1, we have the following $\binom{4+1}{1} = 5$ instances:

$$\begin{array}{ll} \text{Instance 0 :} & S^* \; i_1 \; i_2 \; i_3 \; i_4 \; R^* \\ \text{Instance 1 :} & S \; i_1^* \; i_2 \; i_3 \; i_4 \; R^* \\ \text{Instance 2 :} & S \; i_1 \; i_2^* \; i_3 \; i_4 \; R^* \\ \text{Instance 3 :} & S \; i_1 \; i_2 \; i_3^* \; i_4 \; R^* \\ \text{Instance 4 :} & S \; i_1 \; i_2 \; i_3 \; i_4^* \; R^* \end{array}$$

In the above, i_k represents the k^{th} intermediate node on the path. We use the notation $*$ to indicate that a node is compromised.

Let $f(i, j, L)$ be the probability that the identity of the true sender can be discovered for instance j. That is,

$$P(S = s|i, j, L) = f(i, j, L). \quad (8)$$

[3] Recall that we assume that the receiver is always compromised. So the i compromised nodes here do not include the receiver.

$f(i,j,L)$ can be calculated by applying the elimination rules to the instance. For example, for Instance 1 listed above, we know that i_1, i_2 and i_4 cannot be the sender per rules R1, R2, and R3, respectively. Thus, for this instance, the non-sender set is given by

$$NS = \{i_1, i_2, i_4\}. \tag{9}$$

Then, by (6),

$$f(1,1,4) = \frac{1}{N - |NS|} = \frac{1}{N-3}. \tag{10}$$

Formally, a procedure of computing $f(i,j,L)$ can easily be established by using the elimination rules as operators. (The details of the procedure is not given here due to the space limitation.) With given $f(i,j,L)$, we can obtain the probability that the sender's identity can be discovered for a given number of compromised nodes, as illustrated in the following lemma:

Lemma 1: For a simple path of length L and i compromised nodes, the conditional probability that the identity of the true sender s is discovered is given by

$$P(S = s|i) = \frac{1}{\binom{L+1}{i}} \sum_{j=0}^{\binom{L+1}{i}-1} f(i,j,L). \tag{11}$$

The lemma is easily obtained with the law of the total probability and the definition of $f(i,j,L)$.

A.2 General Results

Based on Lemma 1, we can now obtain the general results on the probability that the sender can be identified. These results are presented for fixed-length paths in Theorems 2 and for variable-length paths in Theorem 3, respectively.

First, we deal with the case of simple paths with fixed-length.

Theorem 2: For a system with M compromised nodes and a simple path of length L, the probability that the true sender s can be discovered is given by

$$P(S=s) = \sum_{i=c}^{d} \left(\left(\frac{1}{\binom{L+1}{i}} \sum_{j=0}^{\binom{L+1}{i}-1} f(i,j,L)\right) \times \frac{\binom{M}{i}\binom{N-M}{L+1-i}}{\binom{N}{L+1}} \right). \tag{12}$$

where

$$c = \max(0, L+1-(N-M)), \tag{13}$$

and

$$d = \min(M, L). \tag{14}$$

Proof: Recall that the path selection algorithm constructs the rerouting path by randomly choosing the intermediate nodes. All the nodes are chosen as intermediate nodes on the path with equal probability. Also note that path selection algorithm does not know which nodes are compromised.

Then, K, the number of compromised nodes on the path, can be any integral value between $\max(0, L+1-(N-M))$ and $\min(M,L)$. By the law of total probability, we have

$$P(S=s) = \sum_{i=\max(0,L+1-(N-M))}^{\min(M,L)} P(S=s|K=i)P(K=i), \tag{15}$$

where $P(K=i)$ represents the probability that a simple path of length L contains exactly i compromised nodes. Since both the nodes on the rerouting path and the compromised nodes are randomly chosen, we have

$$P(K=i) = \frac{\binom{M}{i}\binom{N-M}{L+1-i}}{\binom{N}{L+1}}. \tag{16}$$

Substituting (11) and (16) into (15), we get (12).

■

Now, we consider the case simple paths with variable-length. Formally, let random variable L be the path length chosen by the sender, random variable C be the number of compromised nodes on a path, L' be the path length considered by the adversary, J be an index marking instances of paths that have length L and i compromised nodes. Using these notations, we have the following theorem,

Theorem 3: Assume that, in a system with M compromised nodes, simple paths with variable-length are used and that the path length L conforms to the probability distribution $P(L=l)$, where $a \le l \le b$. Then, the probability that the true sender s can be discovered is given by

$$P(S=s) = \sum_{l=a}^{b}\sum_{i=c}^{d} \frac{\sum_{j=0}^{z}\sum_{l'=u}^{v} g(l',i,j,l)q(l',i,j,l)}{\binom{l+1}{i}} P(L=l) \tag{17}$$

where

$$c = max(0, l+1-(N-M)) \tag{18}$$

$$d = min(M,l) \tag{19}$$

$$z = \binom{l+1}{i} - 1 \tag{20}$$

$$u = h(i,j,l) \tag{21}$$

$$v = b-(M-k) \tag{22}$$

$$g(l',i,j,l) = P(S=s|L'=l', J=j, C=i, L=l) \tag{23}$$

$$q(l',i,j,l) = P(L'=l'|J=j, C=i, L=l) \tag{24}$$

Like $f()$, functions $g()$ and $q()$ can be computed by applying operators related to the rules used in the Adversary Algorithm shown in Figure 4.

Sketch of proof By the law of total probability, we have

$$P(S=s) = \sum_{l=a}^{b} P(S=s|L=l)P(L=l). \tag{25}$$

$P(S=s|L=l)$ can be further computed by applying the law of total probability. The (17) follows.

So far, we have obtained the general analytical results for the case of simple paths. Similar approach can be taken to deal with the case of complicated paths. The only difference is that for the latter, the construction of the instance list for a path of length L with i compromised nodes is much more complicated. The related results are in [17] and will not be presented here due to the space limitations.

A.3 Special Cases

In Section V-A.2, we have derived the general results for computing the probability that the true sender can be discovered.

Here we analyze two special cases to obtain closed-form formulas. While these two special cases are simple, their closed-form formulas will help us to analytically verify certain properties we observed in the numerical analysis described in Section V-B.

In the first special case, we consider a system using a fixed-length simple path and having *exactly one* compromised node. The probability that the true sender can be discovered can be easily determined as follows:

Theorem 4: For a system having exactly one compromised node and using a fixed length path, when $1 \leq N \leq 3$, we have

$$P(S = s) = 1. \tag{26}$$

When $N = 4$, we have

$$P(S = s) = \begin{cases} \frac{3}{4}, & L = 1; \\ \frac{3}{4}, & L = 2; \\ \frac{7}{8}, & L = 3. \end{cases} \tag{27}$$

When $N \geq 5$, we have

$$P(S = s) = \begin{cases} \frac{3}{N}, & L = 1; \\ \frac{3}{N}, & L = 2; \\ \frac{2}{N} + \frac{N^2-5N+7}{N(N-2)(N-3)}, & L = 3; \\ \frac{1}{N} + \frac{N^2-4N+6}{N(N-2)(N-3)} + \frac{L-4}{N(N-2)(N-3)}, & L \geq 4. \end{cases} \tag{28}$$

In the second special case, a system that uses variable-length paths zero or one length. The distribution of the length is given as follow:

$$P(L = x) = \begin{cases} p, & x = 0; \\ 1 - p, & x = 1. \end{cases} \tag{29}$$

In this case, the following theorem can be obtained.

Theorem 5: For a system that uses variable-length paths and has a length distribution conforming to (29), we have

$$\begin{aligned} P(S = s) &= \frac{M}{N} + (1-p)\frac{M}{(N-1)}(1 - \frac{M}{N}) \\ &+ \left(p + \frac{N-1-M}{N-1} \times (1-p)\right) \\ &\times (1 - \frac{M}{N}) \\ &\times \left(p^2 + \frac{1}{N-1-M}(1-p)^2\right). \end{aligned} \tag{30}$$

The proofs of the above two theorems are given in the Appendix.

B. Observations

In this section, we numerically compute the values of P(S = s) of a number of anonymous communication systems and analyze how path length, path topology, and number of compromised nodes impact the values of P(S = s).

B.1 Impact of Path Length

We will start with the case of fixed length. We then consider the systems in which the path length can vary.

B.1.a Case of Fixed Length Paths.

Figure 5 compares the detection probability of the sender for systems with fixed path length selection and with varying path topology selection strategies and varying numbers of compromised nodes.[4] We have the following observations from Figure 5:

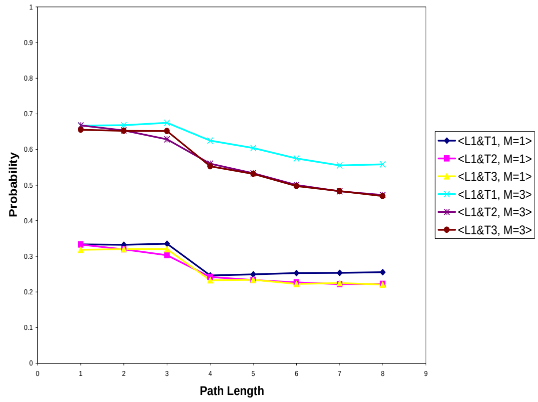

Fig. 5. $P(S = s)$ vs. Path Length (N=9 and Fixed Length)

- In many cases, the value of $P(S = s)$ decreases as the path length increases. This coincides with general intuition: the more a message gets rerouted, the more difficult it is for the adversary to infer the sender.
- However, $P(S = s)$ is not always monotonically decreasing as the path length increases. For example, for the systems where Algorithms L1 and T1 are used and a single compromised node, $P(S = s)$ reaches its minimum at $L = 4$. After that, $P(S = s)$ becomes an increasing function of L.

The second observation above is against intuition. People tend to think that the longer the path length is, the better the anonymity should be. Our results show that this is NOT always true. This phenomenon can be actually be explained: As the path length increases, the possibility increases that compromised nodes are chosen as the intermediate nodes. In this way, the adversary would get better knowledge of the path, improving its chance to identify the sender.

This observation can be confirmed analytically. Consider (28) with $N \geq 5$. By applying classical calculus techniques, it is easy to find that $P(S = s)$ reaches its minimum at $L = 4$ and its maximum at $L = 3$. That is, in this kind of systems, increasing path length may not necessarily improve the performance. It is the best to chose $L = 4$ while it is the worst to use $L = 3$!

B.1.b Case of Variable Length Paths.

Here, the path length is a random variable. We consider the following typical distributions of path length. Some of them are used in existing anonymous communication systems:

- Fixed path length

$$P(L = x) = \begin{cases} 1, & \text{x = c;} \\ 0, & \text{Otherwise.} \end{cases} \tag{31}$$

[4]The label $< L1\&T2, M = 3 >$ in Figure 5 represents that path length selection algorithm $L1$ and path topology selection algorithm $T2$ in Section III are used and there are 3 compromised nodes in the system. This style of label in this section can be explained in the same way.

- Variable path length conforming to geometrical distribution

$$P(L = x) = (1 - P_f)P_f^x, \text{ for } x \geq 0. \quad (32)$$

- Variable path length conforming to uniform distribution

$$P(L = x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b; \\ 0, & \text{Otherwise.} \end{cases} \quad (33)$$

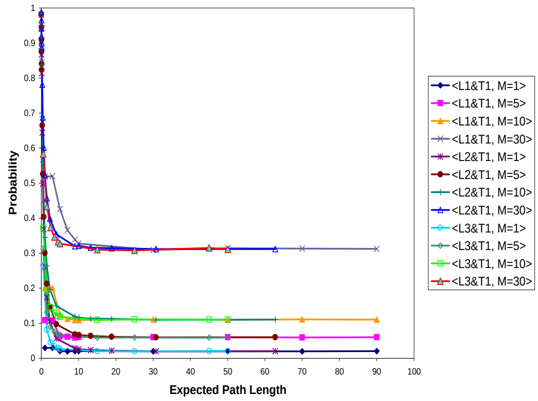

Fig. 6. $P(S = s)$ vs. Expected Path Length (N=100, Simple Path and Variable Length)

Figure 6 shows the numerical results about detection probability under different path length distributions. From Figure 6, we can make the following observations:

- When the expected path length is sufficiently large (say, larger than 15), the probability that the true sender can be discovered under different path length distributions become very close. The difference due to using different length distributions is relatively small.
- When the expected path length is small, $P(S = s)$ is sensitive to the expected path length. Generally speaking, when the expected path length increases, $P(S = s)$ decreases in most cases. Again, in some special cases, this may not be always true.
- When the expected length is small (say, less than 10), the system that uses uniform distribution performs the best.

Let us again use a closed-form formula to verify the non-monotonicity of $P(S = s)$. Consider the simple case where the variable path length conforms to the distribution defined in (29). Note that $1 - p$ is the expected path length.

From (30), we have

$$\frac{\partial P(S = s)}{\partial p} = -Ap^2 - Bp - C, \quad (34)$$

where

$$A = \frac{3M(M-N)^2}{(N-1)(M-N+1)N}, \quad (35)$$

$$B = \frac{2(M-N)(-N^2+N+2MN-M^2+M)}{(N-1)(M-N+1)N}, \quad (36)$$

and

$$C = \frac{(M-N)(-M^2+MN-2-4M+2N)}{(N-1)(M-N+1)N}. \quad (37)$$

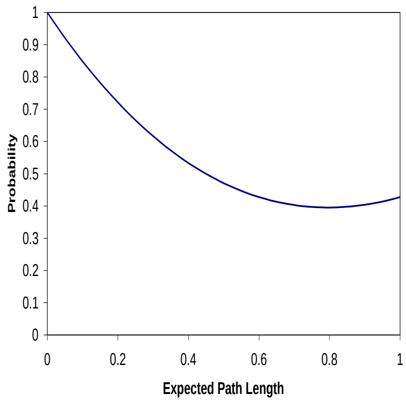

Fig. 7. $P(S = s)$ vs. Expected Path Length ($N = 11, M = 2$, Variable Length Conforming to the Distribution in (29))

When $\frac{\partial P(S=s)}{\partial p} = 0$, i.e., $p = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$, $P(S = s)$ achieves its minimum value. There may be two values for p here. We should choose the one in $[0, 1]$. Figure 7 illustrates the impact of p on $P(S = s)$. In this case, we can observe that $P(S = s)$ achieves its minimum value 0.395 for $p = 0.204$.

B.2 Impact of Path Topology

Due to space limitation, we will only consider the impact of path topology under fixed path length strategy here.

Path topology is determined by the way the intermediate nodes on the path are chosen. If one node can be chosen more than once, the path is a complicated one. We compare the three path topology selection algorithms of T1, T2, and T3, respectively. We assume that the path length are fixed. Figure 8 shows

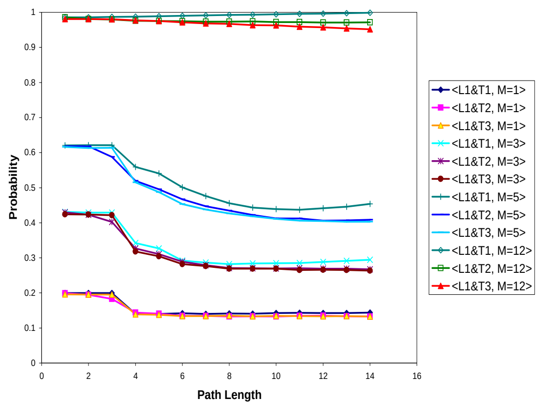

Fig. 8. $P(S = s)$ vs. Path Length (N=15 and Fixed Length)

the impact of complexity of path topology on $P(S = s)$. We can see that complicated path performs better than simple path. However, the difference are relatively small.

B.3 Impact of Number of Compromised Nodes

Figures 9 and 10[5] show how, M, the number of compromised nodes, affects on $P(S = s)$. As expected, the probability that

[5] The label $< L1\&T_2, L = 3 >$ in Figures 9 and 10 represents that path length selection algorithm L_1 and path topology selection algorithm T_2 in Section III are used and the path length is 3.

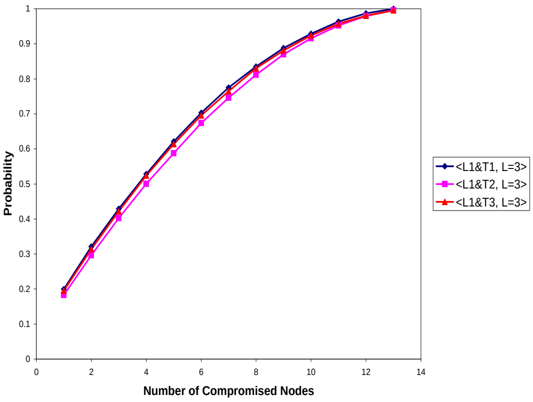

Fig. 9. $P(S = s)$ vs. Number of Compromised Nodes (N=15, and Fixed Length $L = 3$)

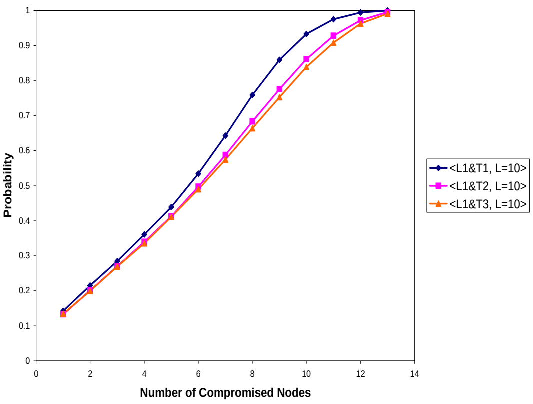

Fig. 10. $P(S = s)$ vs. Number of Compromised Nodes (N=15, and Fixed Length $L = 10$)

the true sender can be discovered increases as the number of compromised nodes increases. This conforms to general intuition. In addition, the deterioration of the system behavior goes sublinearly as the number of compromised nodes increases.

VI. Conclusions

In this paper, we quantitatively analyze the anonymity behavior of anonymous communication systems under different rerouting path selection strategies. We considered several path selection methods used by applications and modeled the behavior of the adversary. We measure the system anonymity in terms of the probability that the identity of a message sender can be discovered by the adversary. Our main results from this study are:

1. A general intuition has been that the longer the rerouting path, the better the system's anonymity. While this is true in many cases, our analytical result shows that the anonymity of the system may NOT always be improved as path length increases. A longer path of rerouting may not result in a better anonymity. We provide formulas that allow designers identify optimal rerouting path lengths under varying conditions.

2. While complicated paths perform better than simple ones, the difference on system anonymity due to different complexity of path topology is usually relatively small.

3. As expected, system anonymity becomes more vulnerable as the number of compromised nodes increases. In addition, the deterioration of the system behavior goes sublinearly as the number of compromised nodes increases.

Following these analytical results, we can see that several existing anonymous communication systems are not using the best path selection strategies. For example, Freedom [12], [35] should use simple with fixed length 4, instead of 3 as in the current system.

Thus, the results reported in this paper can help system developers to properly design path selection algorithm and consequently improve their anonymous communication systems.

Acknowledgments

This work was partially sponsored by NSF under contract number EIA-0081761, by DARPA under contract number F30602-99-1-0531, and by Texas Higher Education Coordinating Board under its Advanced Technology Program.

References

[1] The Anonymizer, http:/www.anonymizer.com/.
[2] Anonymous Remailer, http://www.lcs.mit.edu/research/anonymous.html.
[3] O. Berthold, H. Federrath, and M. Köhntopp, *Project Anonymity and Unobservability in the Internet*, Proceedings of the tenth conference on Computers, freedom privacy: challenging the assumptions April 4 - 7, 2000, Toronto, ON Canada.
[4] D. Chaum, *Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms*, Communications of the ACM, v. 24, n. 2, pp. 84-88, 1981.
[5] D. Chaum, *The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability*, Journal of Cryptology 1/1 (1988), pp. 65-75, (draft received May 13, 1985).
[6] R. C. H. Cheng, *Generating Beta Variatew with Nonintegral Shape Parameters*, Communications of the ACM, v. 21, n. 4, pp. 317-322, 1978.
[7] J. Claessens, B. Preneel, J. Vandewalle, *Solutions for anonymous communication on the Internet*, Proceedings of IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology, pp. 298 -303, 1999.
[8] W. Dai, *PipeNet 1.1*, http:/www.eskimo.com/ weidai/pipenet.txt, 1998.
[9] J. D. Faires and R. L. Burden, *Numerical Methods*, PWS-KENT Publishing Company, Boston, MA, 1993
[10] E. Gabber, P. B. Gibbons, Y. Matias, and A. Mayer, *How to make personalized web browing simple, secure, and anonymous*, Proceedings of Financial Cryptography'97-LNCS 1318. Springer-Verlag, 1997.
[11] J. E. Gentle, *Random number generation and more monte carlo methods*, Springer-Verlag, New York, 1998.
[12] I. Goldberg and A. Shostack, *Freedom network 1.0 architecture and protocols*, http://www.freedom.net/info/freedompapers/index.html, 1999.
[13] D. Goldschlag, M. Reed, and P. Syverson, *Onion Routing for Anonymous and Private Internet Connections*, Communications of the ACM, v. 42, n. 2, pp. 39-41, 1999.
[14] Y. Guan, C. Li, D. Xuan, R. Bettati, and W. Zhao, *Preventing traffic analysis for real-time communication networks*, Proceedings of IEEE MILCOM1999, vol. 1, pp. 744-750, Nov. 1999.
[15] Y. Guan, X. Fu, R. Bettati, and W. Zhao, *Efficient Traffic Camouflaging in Mission-critical QoS-guaranteed Networks*, Proceedings of IEEE Information Assurance and Security Workshop, pp. 143-149, West Point, June 2000.
[16] Y. Guan, X. Fu, D. Xuan, P. Shenoy, R. Bettati, and W. Zhao, *NetCamo: Camouflaging Network Traffic for QoS-Guaranteed Mission Critical Applications*, IEEE Transactions on System, Man, and Cybernetics, Special Issue on Information Assurance, Vol. 31, No. 4, pp. 253-266, July 2001.
[17] Y. Guan, X. Fu, R. Bettati, and W. Zhao, *A Quantitative Analysis of Anonymous Communications*, Technical Report TR01-016, Dept. of Computer Science, Texas A&M University, July 2001.
[18] C. Gülcü and G. Tsudik, *Mixing Email with Babel*, Proceedings of the 1996 Symposium on Network and Distributed System Security, 1996.
[19] C. A. R. Hoare, *Communicating Sequential Processes*, Prentice-Hall, 1985.
[20] A. Jerichow, J. Müller, A. Pfitamann, B. Pfitzmann, and M. Waidner, *Real-Time Mixes: A Bandwidth-Efficient Anonymity Protocol*, IEEE Journal on Selected Areas in Communications, v. 16, n. 4, pp. 495-509, 1998.
[21] Lucent Personalized Web Assistant, http://www.bell-labs.com/projects/lpwa.

[22] S. Kent and R. Atkinson, *Security Architecture for the Internet Protocol*, RFC2401, 1998.
[23] R. E. Newman-Wolfe, B. R. Venkatraman, *High Level Prevention of Traffic Analysis*, Proceedings of the Seventh Annual Computer Security and Applications Conference, San Antonio, Texas, Dec. 2-6, pp. 102-109, 1991.
[24] R. E. Newman-Wolfe, B. R. Venkatraman, *Performance Analysis of a Method for High Level Prevention of Traffic Analysis*, Proceedings of the 8th Annual Computer Security and Applications Conference, San Antonio, Texas, pp. 123-130, Nov. 30-Dec. 4, 1992.
[25] M. Reed, P. Syverson, and D. Goldschlag, *Anonymous Connections and Onion Routing*, IEEE Journal on Selected Areas in Communications, v. 16, n. 4, pp. 482-494, 1998.
[26] M. K. Reiter and A. D. Rubin, *Crowds: Anonymity for Web Transactions*, ACM Transactions on Information and System Security, v. 1, n. 1, pp. 66-92, 1998.
[27] S. Schneider and A. Sidiropoulos, *CSP and Anonymity*, ESORICS'96, E. Kurth, G. Martella, and E. Montolivio, eds., LNCS vol. 1146, Springer-Verlag, pp. 198-218, 1996.
[28] C. Shields and B. N. Levine, *A Protocol for Anonymous Communication Over the Internet*, Proceedings of the 7th ACM Conference on Computer and Communication Security, Athens, Greece, Nov. 1-4, 2000.
[29] V. Scarlata, B.N. Levine, and C. Shields, *Responder Anonymity and Anonymous Peer-to-Peer File Sharing*, UMass Technical report 01-xx, Submitted for publication, May 2001.
[30] R. C. Summers, *Secure Computing*, McGraw-Hill Publishing Company, 1997.
[31] P. Syverson, D. Goldschlag, and M. Reed, *Anonymous Connections and Onion Routing*, Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, IEEE CS Press, pp. 44-54, May 1997.
[32] P. Syverson and S. Stubblebine, *Group Principles and the Formalization of Anonymity*, World Congress on Formal Methods'99, Toulouse, France, LNCS 1708 Springer-Verlag, pp. 814-833, Sept. 1999.
[33] P. Syverson, G. Tsudik, M. Reed, and C. Landwehr, *Towards an Analysis of Onion Routing Security*, Workshop on Design Issues in Anonymity and Unobservability, Berkeley, CA, July 2000.
[34] P. Syverson, M. Reed, and D. Goldschlag, *Onion Routing Access Configuration*, DISCEX 2000: Proceedings of the DARPA Information Survivability Conference and Exposition, Hilton Head, SC, IEEE CS Press, pp. 34-40, January 2000.
[35] Anton Stiglic, *Personal Communication*, anton@zks.net, Zero-Knowledge Systems Inc., May 2001.
[36] A. Teich, M. Frankel, R. Kling, and Y. -C. Lee, *Anonymous Communication Policies for the Internet: Results and Recommendations of the AAAS Conference*, The Information Society, 15(2), 1999.
[37] B. R. Venkatraman, R. E. Newman-Wolfe, *Performance Analysis of a Method for High Level Prevention of Traffic Analysis Using Measurements from a Campus Network*, Proceedings of the Tenth Annual Computer Security and Applications Conference, Orlando, Florida, pp. 288-297, Dec. 1994.
[38] Internet Policy Institute (IPI), *Report from the National Workshop on Internet Voting released*, National Workshop on Internet Voting http://www.netvoting.org/Resources/InternetVotingReport.pdf, March 6, 2001.
[39] M. Waidner, *Unconditional Sender and Recipient Untraceability in spite of active attacks*, Eurocrypt'89, April 1989.
[40] M. Wright, M. Adler, B.N. Levine, and C. Shields, *An Analysis of the Degradation of Anonymous Protocols*, UMass Technical report 01-xx, Submitted for publication, May 2001.
[41] Zero-knowledge Systems, http:/www.zeroknowledge.com/.

We understand that INFOCOM has strict limitation on number of pages a paper can have. The material provided in Appendix are for the convenience of the reviewers. We do not intend to publish this material in the proceedings of INFOCOM if this paper is accepted.

APPENDIX

A. Proof of Theorem 1

Theorem 1: If the path selection algorithm in an anonymous communication system chooses the variable path length between the two values a and b according to some specific distribution, i.e., $P(L = i)$ for $a \leq i \leq b$, and the adversary can learn that the actual path length can be any value between c and d by analyzing the information F collected from the agents on the compromised nodes, where $a \leq c \leq d \leq b$, the probability that the identity of the sender can be discovered is given as follows:

$$P(S = s|F) = \sum_{i=c}^{d} (P(S = s|L = i, F) \times \frac{P(L = i)}{\sum_{k=c}^{d} P(L = k)}). \quad (38)$$

Proof:
By the conditional probability formula,

$$P(S = s|F) = \frac{P(S = s, F)}{P(F)} \quad (39)$$

$$= \frac{\sum_{j=a}^{b} P(S = s, F, L = j)}{P(F)} \quad (40)$$

$$= \frac{\sum_{j=a}^{b} P(S = s|F, L = j)P(F, L = j)}{P(F)} \quad (41)$$

$$= \sum_{j=a}^{b} P(S = s|F, L = j)\frac{P(F, L = j)}{P(F)} \quad (42)$$

$$= \sum_{j=a}^{b} P(S = s|F, L = j)P(L = j|F) \quad (43)$$

Because of F, (43) becomes

$$P(S = s|F) = \sum_{j=c}^{d} P(S = s|F, L = j)P(L = j|F). \quad (44)$$

Also by the conditional probability formula,

$$P(L = j|F) = \frac{P(L = j, F)}{P(F)} \quad (45)$$

$$= \frac{P(F|L = j)P(L = j)}{\sum_{k=a}^{b} P(F|L = k)P(L = k)}. \quad (46)$$

Here, $P(F|L=j)$ represents the probability that the path length can be in $[c,d]$ when the true path length is j. By combinatorial rules, we have

$$P(F|L=j) = \begin{cases} 0, & j < c \text{ or } j > d; \\ \frac{P_{N-c}^{j-c}}{P_N^j}, & c \le j \le d. \end{cases} \quad (47)$$

When $j \ge c$,

$$P_N^j = P_N^c P_{N-c}^{j-c} \quad (48)$$

Substituting (48) into (47), then

$$P(F|L=j) = \begin{cases} 0 & j < c \text{ or } j > d \\ \frac{1}{P_N^c} & c \le j \le d \end{cases} \quad (49)$$

Substituting (49) into (46), we have,

$$P(L=j|F) = \begin{cases} 0 & j < c \text{ or } j > d \\ \frac{P(L=j)}{\sum_{k=c}^{d} P(L=k)} & c \le j \le d \end{cases} \quad (50)$$

Substituting (50) into (44), we get (38).

■

B. *Proof of Theorem 4*

Theorem 4: For a system having exactly one compromised node and using a fixed length path, when $1 \le N \le 3$, we have

$$P(S=s) = 1. \quad (51)$$

When $N=4$, we have

$$P(S=s) = F(N,1,L) = \begin{cases} \frac{3}{4}, & L=1; \\ \frac{3}{4}, & L=2; \\ \frac{7}{8}, & L=3. \end{cases} \quad (52)$$

When $N \ge 5$, we have

$$P(S=s) = F(N,1,L) = \begin{cases} \frac{3}{N}, & L=1; \\ \frac{3}{N}, & L=2; \\ \frac{2}{N} + \frac{N^2-5N+7}{N(N-2)(N-3)}, & L=3; \\ \frac{1}{N} + \frac{N^2-4N+6}{N(N-2)(N-3)} + \frac{L-4}{N(N-2)(N-3)}, & L \ge 4. \end{cases} \quad (53)$$

Proof:

When N=1, it is obvious that $P(S=s)=1$ since the message can only be sent from its sender to its receiver directly (i.e., no rerouting).

When N=2, there are two cases for path length:

- L=0: This is the same case as $N=1$, the true sender can be determined definitely, i.e., $P(S=s)=1$.
- L=1: Since there is one compromised node in the system, either the true sender or the only intermediate node is compromised. In either case, $P(S=s)=1$.

So we have $P(S=s)=1$ when N=2.

When N=3, there are three cases for path length:

- L=0: Same as the case when $N=1$, the true sender can be determined definitely, i.e., $P(S=s)=1$.
- L=1: Since there is one compromised node in the system, we have the following three instances:

Instance 0:	$S^* \; i_1 R^*$
Instance 1:	$S \; i_1^* R^*$
Instance 2:	$S \; i_1 R^*$

Following the elimination rules R_1, R_2 and R_3 it is obvious that $P(S=s)=1$ for Instance 0 and Instance 1. For Instance 2, we know that the cpmpromised node is not on the path and the intermediate node on the path could not be the sender. Thus, there is the only remaining node which could be the sender, i.e., the sender itself. So $P(S=s)=1$ for Instance 2. So we have $P(S=s)=1$ when N=3 and L=1.

- L=2: Here, the compromised node must be on the path. We also have three instances:

Instance 0 :	$S^* \; i_1 \; i_2 R^*$
Instance 1 :	$S \; i_1^* \; i_2 R^*$
Instance 2 :	$S \; i_1 \; i_2^* R^*$

It is obvious that $P(S=s)=1$ for Instance 0. For Instance 1 and Instance 2, we can always eliminate two intermediate nodes which could not be the sender. Now the sender itself is the only remaining node in the set. Thus, we have $P(S=s)=1$ when N=3 and L=2.

So $P(S=s)=1$ when N=3.

We skip the proof for case of $N=4$, since it is similar to the case of $N \ge 5$.

We prove the case of $N \ge 5$. Since the path selection algorithm uses simple path with fixed length, we know $0 \le L \le N-1$.

Case 1: $L=0$**.** It is obvious that $P(S=s)=1$.

Case 2: $L=1$**.** We have the following three instances:

Instance 0 :	$S^* \; i_1 R^*$
Instance 1 :	$S \; i_1^* R^*$
Instance 2 :	$S \; i_1 R^*$

Following the elimination rules, we have

- Instance 0: The true sender can be determined definitely.
- Instance 1: The true sender can also be determined definitely.
- Instance 2: The 1^{st} intermediate node and the compromised node j (not on the path) could not be the true sender, i.e., $NS = \{i_1, j\}$.

So we have

$$P(S=s|Instance\ 0) = 1, \quad (54)$$

$$P(S=s|Instance\ 1) = 1, \quad (55)$$

$$P(S=s|Instance\ 2) = \frac{1}{N-2}. \quad (56)$$

Thus,

$$P(S=s) = \sum_{k=0}^{2} P(S=s|Instance\ k)P(Instance\ k) \quad (57)$$

$$= 1 \times \frac{1}{\binom{2}{1}} \times \frac{\binom{N-1}{1}}{\binom{N}{2}} + 1 \times \frac{1}{\binom{2}{1}} \times \frac{\binom{N-1}{1}}{\binom{N}{2}} \quad (58)$$

$$+\frac{1}{N-2} \times \frac{\binom{N-1}{2}}{\binom{N}{2}} \quad (59)$$

$$= \frac{3}{N}. \quad (60)$$

Case 3: $L = 2$**.** Using a similar argument to case 2, we have

$$P(S = s) = \frac{3}{N}. \quad (61)$$

Case 4: $L = 3$**.** We have the following five instances:

$$\begin{array}{ll} \text{Instance 0}: & S^* \, i_1 \, i_2 \, i_3 R^* \\ \text{Instance 1}: & S \, i_1^* \, i_2 \, i_3 R^* \\ \text{Instance 2}: & S \, i_1 \, i_2^* \, i_3 R^* \\ \text{Instance 3}: & S \, i_1 \, i_2 \, i_3^* R^* \\ \text{Instance 4}: & S \, i_1 \, i_2 \, i_3 R^* \end{array}$$

Following the elimination rules R_1,R_2 and R_3, we have

- Instance 0: The true sender can be determined definitely.
- Instance 1: The true sender can also be determined definitely.
- Instance 2: 1^{st}, 2^{nd} and 3^{rd} intermediate nodes could not be the true sender.
- Instance 3: 2^{nd} and 3^{rd} intermediate nodes could not be the true sender.
- Instance 4: 3^{rd} intermediate node and the compromised node (not on the path) could not be the true sender.

So we have

$$\begin{aligned} P(S = s|Instance\, 0) &= 1, & (62) \\ P(S = s|Instance\, 1) &= 1, & (63) \\ P(S = s|Instance\, 2) &= \frac{1}{N-3}, & (64) \\ P(S = s|Instance\, 3) &= \frac{1}{N-2} & (65) \end{aligned}$$

and

$$P(S = s|Instance\, 4) = \frac{1}{N-2}. \quad (66)$$

Thus,

$$\begin{aligned} &P(S = s) \\ &= \sum_{k=0}^{4} P(S = s|Instance\ k)P(Instance\ k) & (67) \\ &= (1 + 1 + \frac{1}{N-3} + \frac{1}{N-2}) \times \frac{1}{\binom{4}{1}} \times \frac{\binom{N-1}{3}}{\binom{N}{4}} & (68) \\ &\quad + \frac{1}{N-2} \times \frac{\binom{N-1}{4}}{\binom{N}{4}} & (69) \\ &= \frac{2}{N} + \frac{N^2 - 5N + 7}{N(N-2)(N-3)}. & (70) \end{aligned}$$

Case 5: $L \geq 4$**.** We have the following $(L + 2)$ instances:

$$\begin{array}{rl} \text{(Instance 0)} & S^* \, i_1 \, i_2 \cdots i_{L-1} \, i_L R^* \\ \text{(Instance 1)} & S \, i_1^* \, i_2 \cdots i_{L-1} \, i_L R^* \\ \text{(Instance 2)} & S \, i_1 \, i_2^* \cdots i_{L-1} \, i_L R^* \\ \cdots & \\ \text{(Instance L-1)} & S \, i_1 \, i_2 \cdots i_{L-1}^* \, i_L R^* \\ \text{(Instance L)} & S \, i_1 \, i_2 \cdots i_{L-1} \, i_L^* R^* \\ \text{(Instance L+1)} & S \, i_1 \, i_2 \cdots i_{L-1} \, i_L R^* \end{array} \quad (71)$$

Following elimination rules, we have

- Instance 0: The true sender can be determined definitely.
- For Instance k, where $1 \leq k < L - 1$: Exactly three of the intermediate nodes on the path could not be the true sender, i.e., $NS = \{i_k, i_{k+1}, i_L\}$.
- Instance L-1: the $(L - 2)^{th}$, $(L - 1)^{th}$ and L^{th} intermediate nodes could not be the true sender, i.e., $NS = \{i_{L-1}, i_L\}$.
- Instance L: the $(L - 1)^{th}$ and L^{th} intermediate nodes could not be the true sender, i.e., $NS = \{i_{L-2}, i_{L-1}, i_L\}$.
- Instance L+1: the L^{th} intermediate node and the compromised node j (not on the path) could not be the true sender, i.e., $NS = \{i_L, j\}$.

So we have

$$\begin{aligned} P(S = s|Instance\, 0) &= 1, & (72) \\ P(S = s|Instance\, k) &= \frac{1}{N-3}, & (73) \\ P(S = s|Instance\, L) &= \frac{1}{N-2}, & (74) \\ P(S = s|Instance\, L + 1) &= \frac{1}{N-2}, & (75) \end{aligned}$$

where $1 \leq k \leq L - 1$.

Thus,

$$\begin{aligned} &P(S = s) \\ &= \sum_{k=0}^{L+1} P(S = s|Instance\ k)P(Instance\ k) & (76) \\ &= (1 + \frac{L-1}{N-3} + \frac{1}{N-2}) \times \frac{1}{\binom{L+1}{1}} \times \frac{\binom{N-1}{L}}{\binom{N}{L+1}} & (77) \\ &\quad + \frac{1}{N-2} \times \frac{\binom{N-1}{L+1}}{\binom{N}{L+1}} & (78) \\ &= \frac{1}{N} + \frac{N^2 - 4N + 6}{N(N-2)(N-3)} + \frac{L-4}{N(N-2)(N-3)} & (79) \end{aligned}$$

where $L \geq 4$.

Theorem 4 now follows.

■

C. Proof of Theorem 5

Theorem 5: For a system that uses variable length paths. If the distribution of the length conforms with (29), we have

$$P(S = s) = \frac{M}{N} + (1 - p)\frac{M}{(N-1)}(1 - \frac{M}{N})$$

$$
\begin{aligned}
&+(p+\frac{N-1-M}{N-1}\times(1-p)) \\
&\times(1-\frac{M}{N}) \\
&\times(p^2+\frac{1}{N-1-M}(1-p)^2). \qquad (80)
\end{aligned}
$$

Proof: We partition the entire sample space into three parts,
$\mathcal{A}$: The sender is compromised;
$\mathcal{B}$: The sender is not compromised ($\overline{\mathcal{A}}$), one of the compromised nodes receives a packet (denoted as $\mathcal{G}$);
$\mathcal{C}$: The sender is not compromised ($\overline{\mathcal{A}}$), no compromised nodes receive packet ($\overline{\mathcal{G}}$)
By the law of total probability,

$$
\begin{aligned}
P(S=s) &= P(S=s,\mathcal{A})+P(S=s,\mathcal{B}) \\
&\quad +P(S=s,C) \qquad (81)\\
&= P(S=s|\mathcal{A})P(\mathcal{A}) \\
&\quad +P(S=s|\mathcal{B})P(\mathcal{B}) \\
&\quad +P(S=s|C)P(C) \qquad (82)
\end{aligned}
$$

Now we calculate the three terms in (81):
(1) $P(S=s|\mathcal{A})P(\mathcal{A})$. Clearly,

$$
\begin{aligned}
P(\mathcal{A}) &= \frac{M}{N} \qquad (83)\\
P(S=s|\mathcal{A}) &= 1 \qquad (84)
\end{aligned}
$$

So

$$P(S=s|\mathcal{A})P(\mathcal{A})=\frac{M}{N} \qquad (85)$$

(2) $P(S=s|\mathcal{B})P(\mathcal{B})$. By the definition of event $\mathcal{B}$, we can see

$$
\begin{aligned}
P(\mathcal{B}) &= P(\overline{\mathcal{A}},\mathcal{G}) \qquad (86)\\
&= P(\mathcal{G}|\overline{\mathcal{A}})P(\overline{\mathcal{A}}) \qquad (87)
\end{aligned}
$$

Morevoer,

$$
\begin{aligned}
P(\overline{\mathcal{A}}) &= 1-\frac{M}{N} \qquad (88)\\
P(\mathcal{G}|\overline{\mathcal{A}}) &= P(\mathcal{G},L=0|\overline{\mathcal{A}}) \\
&\quad +P(\mathcal{G},L=1|\overline{\mathcal{A}}) \qquad (89)\\
&= P(\mathcal{G}|\overline{\mathcal{A}},L=0)P(L=0|\overline{\mathcal{A}}) \\
&\quad +P(\mathcal{G}|\overline{\mathcal{A}},L=1)P(L=1|\overline{\mathcal{A}}) \qquad (90)
\end{aligned}
$$

It's easy to see that when path length $L=0$ and the sender is not copromised (event $\overline{\mathcal{A}}$),

$$P(\mathcal{G}|\overline{\mathcal{A}},L=0)=0 \qquad (91)$$

Moreover, we can see the fact

$$
\begin{aligned}
P(L=0|\overline{\mathcal{A}}) &= p \qquad (92)\\
P(L=1|\overline{\mathcal{A}}) &= 1-p \qquad (93)
\end{aligned}
$$

Substituting (91) , (92) and (93) into (90), we have

$$P(\mathcal{G}|\overline{\mathcal{A}})=\frac{M}{2(N-1)} \qquad (94)$$

Then

$$P(\mathcal{B})=(1-\frac{M}{N})(1-p)\frac{M}{(N-1)} \qquad (95)$$

Moreover,

$$P(S=s|\mathcal{B})=1 \qquad (96)$$

So

$$P(\mathcal{B})P(S=s|\mathcal{B})=(1-\frac{M}{N})(1-p)\frac{M}{(N-1)} \qquad (97)$$

(3) $P(S=s|C)P(C)$. By the definition of event C, we can see

$$
\begin{aligned}
P(C) &= P(\overline{\mathcal{A}},\overline{\mathcal{G}}) \qquad (98)\\
&= P(\overline{\mathcal{G}}|\overline{\mathcal{A}})P(\overline{\mathcal{A}}) \qquad (99)
\end{aligned}
$$

Then

$$
\begin{aligned}
P(\overline{\mathcal{G}}|\overline{\mathcal{A}}) &= P(\overline{\mathcal{G}},L=0|\overline{\mathcal{A}})+P(\overline{\mathcal{G}},L=1|\overline{\mathcal{A}}) \qquad (100)\\
&= P(\overline{\mathcal{G}}|L=0,\overline{\mathcal{A}})P(L=0|\overline{\mathcal{A}}) \\
&\quad +P(\overline{\mathcal{G}}|L=1,\overline{\mathcal{A}})P(L=1|\overline{\mathcal{A}}) \qquad (101)
\end{aligned}
$$

Clearly,when path length $L=0$, there will be no internediate nodes to receive packets

$$P(\overline{\mathcal{G}}|L=0,\overline{\mathcal{A}})=1 \qquad (102)$$

When path length $L=1$, and the sender is not compromised (event $\overline{\mathcal{A}}$), the probability that no intermediate nodes receive packets is

$$P(\overline{\mathcal{G}}|L=1,\overline{\mathcal{A}})=\frac{N-1-M}{N-1} \qquad (103)$$

Substituting (102) and (103) into (101), we get

$$P(\overline{\mathcal{G}}|\overline{\mathcal{A}})=p+\frac{N-1-M}{N-1}\times(1-p) \qquad (104)$$

Then, we have

$$
\begin{aligned}
P(C) &= (1\times p+\frac{N-1-M}{N-1}\times(1-p)) \\
&\quad \times(1-\frac{M}{N}) \qquad (105)
\end{aligned}
$$

By the law of total probability,

$$
\begin{aligned}
P(S=s|C) &= P(S=s,L=0|C) \\
&\quad +P(S=s,L=1|C) \qquad (106)\\
&= P(S=s|L=0,C)P(L=0|C) \\
&\quad +P(S=s|L=1,C)P(L=1|C) \qquad (107)
\end{aligned}
$$

Moreover, we know

$$
\begin{aligned}
P(L=0|C) &= p \qquad (108)\\
P(L=1|C) &= 1-p \qquad (109)
\end{aligned}
$$

Now in this case the adversary must guess the length of the path, $L^{'} = 0$ represents the event that the observer guesses that the path length is 0, while $L^{'} = 1$ represents the event that the observer guesses that the path length is 1. So we have

$$\begin{aligned} &P(S=s|L=0,C) \\ &\quad= P(S=s,L^{'}=0|L=0,C) \\ &\qquad+P(S=s,L^{'}=1|L=0,C) \\ &\quad= P(S=s|L=0,C,L^{'}=0) \\ &\qquad\times P(L^{'}=0|L=0,C) \\ &\qquad+P(S=s|L=0,C,L^{'}=1) \\ &\qquad\times P(L^{'}=1|L=0,C) \qquad (110) \end{aligned}$$

Clearly when the actual path length $L = 0$ and the observer guesses that the path length is 0 (event $L^{'} = 0$), if the sender is not compromised, and no compromised nodes receive packet (event C), the adversary will be certain to identify the sender. That is,

$$P(S=s|L=0,C,L^{'}=0) = 1 \qquad (111)$$

Consequently,

$$\begin{aligned} P(L^{'}=0|L=0,C) &= p & (112) \\ P(L^{'}=1|L=0,C) &= 1-p & (113) \\ P(S=s|L=0,C,L^{'}=1) &= 0 & (114) \end{aligned}$$

Substituting (111), (112), (113) and (114) into (110), we have

$$\begin{aligned} P(S=s|L=0,C) &= 1\times p + 0\times(1-p) \\ &= p \qquad (115) \end{aligned}$$

Using the same method, we have

$$\begin{aligned} &P(S=s|L=1,C) \\ &\quad= P(S=s,L^{'}=0|L=1,C) \\ &\qquad+P(S=s,L^{'}=1|L=1,C) \\ &\quad= (P(S=s|L=1,C,L^{'}=0) \\ &\qquad\times P(L^{'}=0|L=1,C)) \\ &\qquad+(P(S=s|L=1,C,L^{'}=1) \\ &\qquad\times P(L^{'}=1|L=1,C)) \\ &\quad= 0\times p \\ &\qquad+\frac{1}{N-1-M}\times(1-p) \\ &\quad= \frac{1}{N-1-M}\times(1-p). \qquad (116) \end{aligned}$$

So

$$P(S=s|C) = p^2 + \frac{1}{N-1-M}(1-p)^2 \qquad (117)$$

and

$$\begin{aligned} P(S=s|C)P(C) \qquad (118) \\ = (p+\frac{N-1-M}{N-1}\times(1-p)) \\ \times(1-\frac{M}{N}) \\ \times(p^2+\frac{1}{N-1-M}(1-p)^2) \qquad (119) \end{aligned}$$

Finally, substituting (85), (97) and (119) into (81), we have (80) established.

■