Too Many Cooks: Analyzing the Impact of Multiple Adversaries on Mix Networks

J David Smith
University of Florida
Email: emallson@ufl.edu

***Abstract*—Attacks on mix networks have been extensively studied in the literature. However, most work in this area only considers scenarios with one global active adversary. We examine this problem in the context of blending attacks by modeling more than one independent global active adversary using stochastic processes. We prove that the expected effectiveness of blending attacks changes, and that in some cases greater adversity actually *improves* the effectiveness of mix networks. We further present a model for analyzing the behavior of mix networks under attack by multiple adversaries, and demonstrate its efficacy. Our work shows that single-adversary models do not completely characterize the behavior of security systems, and that further research in this area is needed.**

I. Introduction

A wide range of adversarial models have been studied in security literature. The capabilities of an adversary range from almost none in the case of a local passive adversary to nearly all in the case of a global active adversary (GAA). While it is not uncommon for models to include multiple actors, these are generally acting in unison as a single unit. Far less common is the concept of multiple *oblivious* actors (MOA), who are each conducting their own attack without knowledge of the other actors.

We examine this scenario in the context of mix networks [1]. This choice restricts us to a single well-defined problem: how well does a mix network preserve anonymity under attack by MOAs? Serjantov, Dingledine, and Syverson [2] detailed the effectiveness of various attacks by GAAs and introduced the taxonomy that we will make use of in this work. In particular, they study the class of blending attacks, dividing them into certain / uncertain and exact / inexact based on their worst-case effectiveness. *Certain* attacks will always succeed, while *exact* attacks are capable of identifying a sender with no ambiguity.

They show that several basic kinds of mixes are vulnerable to exact, certain attacks. We show that these same mixes cease to be exact as long as multiple adversaries are attacking at the same time. The question of how realistic this scenario is naturally follows. We note that any attacker seeking to break sender/receiver anonymity with certainty must successfully attack or compromise every mix along the route a message takes. If multiple oblivious attackers are interested in breaking the anonymity of a message, then every attacker must do this. Thus the chance of $m > 1$ attackers acting at the same time on the same mix is the chance of m attackers seeking to trace a message. It is not hard to imagine a scenario in which this is the case.

For example: suppose an activist A is targeted by two different states, X and Y. Their goal is to identify with whom A is communicating in order to stifle dissent. However, these states may be indifferent to each other and thus not willing to collaborate unless necessary. In this case, they'd each conduct their attack independently – and in the process introduce inexactness and uncertainty into the result.

This paper is structured as follows. In section II we define the network and adversarial models and lay out the problem we aim to solve. In section III we present our solution to this problem. Section IV covers work related to this problem and our solution. We conclude with a discussion in section V.

II. Model and Problem Definition

We consider a single mix network with homogeneous mix nodes (e.g. all timed, all threshold, etc.), a single user of interest A, and m adversaries seeking to trace A's messages through the network. Note that each different kind of mix demands independent analysis, just as was done by Serjantov, Dingledine, and Syverson [2]. Due to the time constraints imposed by this being a course project, we only consider simple threshold mixes, which are vulnerable to exact, certain attacks at a cost of $n-1$ messages inserted in the $m = 1$ case.

The attack used is the flooding attack: the adversary sends $n - 1$ messages to the n-threshold mix in addition to the message of interest. During the flood, the adversary must delay all other incoming messages. In the single adversary case, the adversary may perform these actions at any location along the link. However, things become more complicated when additional adversaries enter the picture.

Suppose two adversaries $\mathcal{A}$ and $\mathcal{B}$ are independently executing a flooding attack on a single mix. They must observe at some point on the outbound links, and delay / insert at some point along the inbound links. We term these locations the *observation points* and *control points*, respectively. If $\mathcal{A}$ has a control point preceding $\mathcal{B}$'s control point, then clearly $\mathcal{A}$ will not be able to successfully conduct the attack because $\mathcal{B}$ will delay all of their messages while injecting additional messages. In this case, $m = 1$ for all intents and purposes. Thus, for this attack we consider only the number of attackers that are co-resident on last occupied control point.

We measure the effectiveness of this attack in terms of the effective anonymity set size [3]. The EAS is defined as the entropy of the probability distribution over possible senders. This provides a concrete measure of an attacker's certainty that it has identified the sender of the message of interest. However, note that each attacker has a distinct distribution over possible senders. Therefore, we measure EAS with respect to an individual adversary. With these definitions laid out, we define the problem we seek to solve:

Problem 1. Given a mix M, a message of interest e, a set of adversaries $\mathbb{A}$ with $|\mathbb{A}| = m$, and an adversary $\mathcal{A} \in \mathbb{A}$, determine the EAS of the sender of e with respect to $\mathcal{A}$.

III. Solution

We begin by noting that the messages sent by $\mathcal{B}$ are opaque to $\mathcal{A}$ and indistinguishable from e. Thus, if $\mathcal{A}$ manages to complete flooding M prior to $\mathcal{B}$ sending any messages, then $\mathcal{A}$ is successful in conducting an exact attack. However, if $\mathcal{B}$ sends any messages before M fires, then $\mathcal{A}$ has no way to tell which of the outgoing messages is the target. We can represent the success of $\mathcal{A}$ in terms of the number of messages $\mathcal{A}$ sent before the mix fires.

A. Special Case: m = 2

Theorem 1. When $m = 2$, the probability of A sending k messages prior to $n-1$ messages being sent is given by the kth term of the binomial formula:

$$\binom{n-1}{k} p_{\mathcal{A}}^{k} p_{\mathcal{B}}^{n-1-k}$$

where p_i is the probability of adversary $\mathcal{I}$ sending a message at any given time step.

Proof. First we define some notation. We write a sequence of messages being sent as $(x, \ldots, y)$ where each glyph represents the sender of a message. So (a, b, b) would be the sequence of $\mathcal{A}$ sending a message, followed by $\mathcal{B}$ sending two messages. We denote the probability of adversary $\mathcal{I}$ sending a message at any point in time as p_i.

Note that we are interested only in the *number* of messages sent by $\mathcal{A}$, so the sequences (a, b, b), (b, a, b) and (b, b, a) are all identical. Further, all three have equal probability: $p_a p_b^2$. The number of ways to write a sequence with k a's and $n - 1 - k$ b's is given by the binomial coefficient: $\binom{n-1}{k}$. Thus, the probability of $\mathcal{A}$ sending k messages before $n-1$ have been sent is given by

$$\binom{n-1}{k} p_a^{k} p_b^{n-1-k}$$

which is the kth term of the binomial formula.

□

This gives us sufficient information to compute the EAS with respect to $\mathcal{A}$.

$$\text{EAS}_{\mathcal{A}}(n-1) = H(\sigma)$$

where H is Shannon entropy and σ is the distribution over anonymity set size given by application of Theorem 1.

TABLE I
Expected Anonymity Set Size for $m = 2$ and various values of n, $p_{\mathcal{A}}$, and $p_{\mathcal{B}}$

$n-1$	$p_{\mathcal{A}}$	$p_{\mathcal{B}}$	EAS
10	0.5	0.5	1.89
10	0.2	0.8	2.21
10	0.05	0.95	2.29
50	0.5	0.5	3.34
100	0.5	0.5	4.00
100	0.2	0.8	4.43

B. General Case

We generalize this special case to a Markov Chain representation. Markov chains are stochastic processes which operate on a finite state space and discrete time, with the additional property of being *memoryless* [4]. A memoryless process is one for which the transition to the next state depends only on the current state:

$$\forall X, Y, Z : \Pr[Y \to Z] = \Pr[X \to Y \to Z]$$

We model the possible states of this system as m-tuples of natural numbers starting at 0. The initial state is $q_0 = (0, \ldots, 0)$, which indicates that all m adversaries have sent 0 messages. This state representation captures the possible combinations of message counts during an attack. Finally, we define the following transition function to model the possible ways the process can progress:

$$\Pr[X \to Y] = \begin{cases} \sum k < n-1 & p_i \\ \sum k \geq n-1 \land X = Y & 1 \\ \text{else} & 0 \end{cases}$$

where $\sum k$ is the sum of the number of messages sent by each adversary so far, and p_i is the probability of adversary i sending a message. Then, the final distribution of over all possible states is the *stationary distribution* of the Markov Chain. If we write the transition matrix P, then the stationary distribution π satisfies

$$\pi = \pi P$$

when written as a vector. This distribution gives the probability of remaining on each state as $t \to \infty$. We note that for the case $m = 2$, the probabilities given by π are the same as those predicted by Theorem 1. The stationary distribution is mapped to a distribution over anonymity set size by summing the probabilities of each sequence with $k_a = n - 1 - |\text{AS}|$. There is not a notationally straightforward way to write this, so the formula is omitted. The EAS is then computed as the entropy of the AS size distribution.

This representation can be generalized to other behaviors and attacks by changing the transition function. For example: if $\mathcal{A}$ always sends $n-1$ messages before checking the output of the mix, then the transition function becomes:

TABLE II
EXPECTED ANONYMITY SET SIZE FOR $m = 3$ AND VARIOUS VALUES OF n, $p_{\mathcal{A}}$, $p_{\mathcal{B}}$, AND $p_{\mathcal{C}}$

$n-1$	$p_{\mathcal{A}}$	$p_{\mathcal{B}}$	$p_{\mathcal{C}}$	EAS
10	0.33	0.33	0.33	2.09
10	0.2	0.5	0.3	2.21
10	0.05	0.8	0.15	2.29
50	0.33	0.33	0.33	3.59
100	0.33	0.33	0.33	4.26
100	0.2	0.5	0.3	4.43

$$\Pr\left[X \to Y\right] = \begin{cases} k_a < n-1 & p_i \\ k_a \geq n-1 \land X = Y & 1 \\ \text{else} & 0 \end{cases}$$

where k_a is the number of messages $\mathcal{A}$ has sent at state X.

IV. Related Work

The most closely related prior work is that of Serjantov, Dingledine, and Syverson [2]. They developed the taxonomy of attacks on mix networks that is further developed in this work. However, as was noted previously their work is limited to the single-adversary case. To the best of our knowledge, there is no prior work on this problem.

There is prior work on related problems. Game theoretic models have been devised to optimize the behavior of both the adversary and the principals [5, 6]. These models typically allow collaboration, which is something explicitly disallowed in our model. A logical next step for this work would be to develop a Stackelberg-game model as in [5] for this scenario to optimize adversary behavior. However, these models are known to be NP-hard [7] even though they are well-behaved [8].

Another related problem, detecting attackers, has been studied under a wide variety of scenarios. Rennhard and Plattner [9] introduced a new form of mix along with a method for detecting collusion attacks. Their work is not directly applicable to this problem because the adversaries are colluding instead of oblivious. Work in other fields has included detecting Sybil attacks on social networks [10] and mobile ad-hoc networks [11], detecting spoofing attacks in wireless and other networks [12], among many others.

Among all of these works collusion is either assumed or allowed, or a single adversary is assumed. Practically, there are scenarios where collusion will not occur. The example of rival states given in section I is one such scenario. By revisiting the assumption that adversaries will collude if optimal, we have shown that there are interesting emergent behaviors missed by prior work.

V. Discussion

Our contributions are threefold:

1. We define and motivate a novel adversarial model in which there are multiple attackers.
2. We demonstrate that the behavior of known attacks on mix networks is different under this adversarial model.
3. We present a model for analyzing attack behavior in this kind of scenario.

Of particular interest is that exact, certain attacks become inexact in this scenario, which suggests that in some cases increasing the number of attackers actually *decreases* the threat to users of the system. We only examined the most basic mixes in this work, but there is reason to believe that similar results may be shown for more advanced mix networks and even mix-like networks such as Tor.

However, it is also clear that some attacks to not become less effective as more adversaries join the system. Trickle attacks on timed mixes, for example, would suffer no such drop in effectiveness as the mix will always receive only the one message. Developing a taxonomy for the scaling of attack effectiveness with respect to the number of independent adversaries would be valuable for understanding the deeper behavior of networks under attack.

There are numerous other possible extensions of this work. Expanding the model to continuous-time Markov chains (called Markov processes) would allow modeling of mixes that include timing components (such as timed mixes, and timed pool mixes). Treating the probability of an adversary sending a message as a random variable sampled from a probability distribution rather than a fixed constant would allow a greater degree of detail in modeling adversary behavior and interaction on control points.

In conclusion, we showed that the scenario of multiple independent oblivious adversaries attacking a mix network is distinct from the scenario of a single adversary on a mix network. We further showed that under this adversarial model, attacks on mix networks have emergent behavior which does not appear in the standard single-adversary model.

References

[1] D. L. Chaum, “Untraceable electronic mail, return addresses, and digital pseudonyms,” *Commun. ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[2] A. Serjantov, R. Dingledine, and P. Syverson, “From a trickle to a flood: Active attacks on several mix types,” in *Information Hiding*, Springer, 2003, pp. 36–52.

[3] A. Serjantov and G. Danezis, “Towards an information theoretic metric for anonymity,” in *Privacy Enhancing Technologies*, Springer, 2003, pp. 41–53.

[4] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Soc., 2009.

[5] P. Paruchuri, J. P. Pearce, M. Tambe, F. Ordonez, and S. Kraus, “An efficient heuristic approach for security against multiple adversaries,” in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, ACM, 2007, p. 181.

[6] J. Pita, M. Jain, J. Marecki, F. Ordóñez, C. Portway, M. Tambe, C. Western, P. Paruchuri, and S. Kraus, “Deployed ARMOR protection: The application of a game theoretic model for security at the Los Angeles International Airport,” in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: Industrial track*, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 125–132.

[7] V. Conitzer and T. Sandholm, “Computing the optimal strategy to commit to,” in *Proceedings of the 7th ACM conference on Electronic commerce*, ACM, 2006, pp. 82–90.

[8] Z. Yin, D. Korzhyk, C. Kiekintveld, V. Conitzer, and M. Tambe, “Stackelberg vs. Nash in security games: Interchangeability, equivalence, and uniqueness,” in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1-Volume 1*, International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 1139–1146.

[9] M. Rennhard and B. Plattner, “Introducing MorphMix: Peer-to-peer based anonymous Internet usage with collusion detection,” in *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, ACM, 2002, pp. 91–102.

[10] N. Z. Gong, M. Frank, and P. Mittal, “SybilBelief: A Semi-Supervised Learning Approach for Structure-Based Sybil Detection,” *Inf. Forensics Secur. IEEE Trans. On*, vol. 9, no. 6, pp. 976–987, 2014.

[11] C. Piro, C. Shields, and B. N. Levine, “Detecting the sybil attack in mobile ad hoc networks,” in *Securecomm and Workshops, 2006*, IEEE, 2006, pp. 1–11.

[12] Y. Sheng, K. Tan, G. Chen, D. Kotz, and A. Campbell, “Detecting 802.11 MAC layer spoofing using received signal strength,” in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, IEEE, 2008.