Statistical Disclosure or Intersection Attacks on Anonymity Systems

George Danezis and Andrei Serjantov

University of Cambridge, Computer Laboratory,
William Gates Building, 15 JJ Thomson Avenue,
Cambridge CB3 0FD, United Kingdom.
George.Danezis@cl.cam.ac.uk Andrei.Serjantov@cl.cam.ac.uk

Abstract. In this paper we look at the information an attacker can extract using a statistical disclosure attack. We provide analytical results about the anonymity of users when they repeatedly send messages through a threshold mix following the model of Kesdogan, Agrawal and Penz [7] and through a pool mix. We then present a statistical disclosure attack that can be used to attack models of anonymous communication networks based on pool mixes. Careful approximations make the attack computationally efficient. Such models are potentially better suited to derive results that could apply to the security of real anonymous communication networks.

1 Introduction

Intersection attacks take advantage of persistent communications between two parties to compromise the anonymity offered to them by anonymous communication systems. While it is possible to manage their impact within the anonymous communication infrastructure, they can be devastating when the totality of the anonymous communication system is abstracted as a single mix and attacked. In this case the adversary observes a victim sending messages and notes all their potential receivers. By aggregating and processing such information, Berthold, Pfitzmann and Standtke [2] observe, that an attacker is able to deduce some information about who is communicating with whom.

We are extending the previous work done on a simple model of an anonymity system — a threshold mix. Such a model was used by Kesdogan, Agrawal and Penz [7, 1] to mount an exact, but expensive, attack. Similarly Danezis [5] proposes an approximate but less computationally demanding attack, that still provides very good results. We will revisit the original model proposed and present new analytical results about the information that can be inferred by observing the mix.

Anonymous communication systems cannot in many cases be modelled as abstract threshold mixes, since a set of messages is likely to remain in the network across any chosen division in rounds. We therefore propose a statistical attack that applies to an anonymous communication channel modelled as a pool mix [11]. Such a mix retains a number of messages every round that are mixed with the messages injected in the network in the following round. This model can be used more effectively to study the limit of how much anonymity anonymous communication networks can provide. The attack presented is very efficient, and allows the adversary to judge the confidence of the results. The set of careful approximations that make this attack very efficient are explained as part of this work.

2 Previous Work

Anonymous communications over information networks were introduced in his seminal paper by David Chaum [3]. The basic building block that such systems use to provide the required anonymity properties is the *mix*, a node that takes a batch of input messages and outputs them all in such a way that their correspondence is hidden. Cryptographic techniques are used to hide the correlation between the input and output message bit patterns, and reordering of the messages is used to disrupt the timing patterns within each batch of messages. This mixing strategy is called a *threshold mix*. Other mix strategies have also been suggested that may make the mix node more resilient to active attacks [8, 11], and a body of work has concentrated on measuring the anonymity they provide [10, 6, 12].

Although the mix was originally conceived as a real network node, Kesdogan, Agrawal and Penz model [7] observe that any anonymity system that provides unlinkability (rather than unobservability) to its participants could be modelled as an abstract threshold mix. They then examine the anonymity offered by such a network to a sender that uses the mix across many rounds to communicate with a set of recipients. He describes the *disclosure attack* that can be used to deduce the set of recipients of a target sender. An analysis of the performance of the attack is further investigated by Agrawal, Kesdogan and Penz [1].

Such attacks were previously described as *intersection attacks* [2] or *partitioning attacks*, both when applied to single mixes and when performed against the whole anonymous network. When applied to single mixes, the attack can be eliminated by requiring each message travelling through the network to follow a different path, as originally proposed by Chaum [3], or by restricting the routes that messages can take out of each node [4]. On the other hand, given that senders will be communicating with a persistent set of parties, such attacks will always yield information when applied to the whole network. The Onion Routing project was the first to draw attention to such attacks performed at the edges of the network, and named them *traffic confirmation* attacks [9].

The main disadvantage of the disclosure attack is that its exact nature makes it computationally very expensive. Danezis [5] proposed a statistical attack based on a set of carefully selected approximations that allows an attacker observing the same model of a network to estimate a victim's set of receivers. As we will see, one of the main advantages of the statistical disclosure attack is that it can be generalised and applied against other anonymous communication network models. In particular [7] assumes that an anonymous network can be abstracted as a large threshold mix where batches of messages are anonymized together and sent out to their respective recipients. We will illustrate how the statistical disclosure attack can be generalised to anonymous communication mechanisms that can be modelled as pool mixes, or in other words where some messages are fed forward to the next mixing rounds of the model.

3 Formal Account of the Attack on the Threshold Mix

We follow the model considered by Danezis in [5]. The anonymity system is considered as a threshold mix with threshold $B+1$. Thus, at each round $B+1$ messages are processed. The victim of the attack, Alice, is known to the adversary to send one message at every round to a receiver chosen uniformly at random from a set M. Naturally, if Alice does not send a message during a round, we simply ignore it altogether. The other B senders whom we collectively call Steves send one message each to a receiver chosen independently and uniformly at random from a set N, $M \subseteq N$. The attacker knows $|M|$ (and $|N|$), and wishes to determine M.

We now define some notation. Let p_r be the probability that one of the other senders, a Steve, sends a message to a particular receiver r. Naturally, if they pick their recipients uniformly, $p_r = \frac{1}{|N|}$. Let $q_r = 1 - p_r$, the probability Bob does not send a message to r.

Let us now start with some very simple cases and build up a technique for analysing how much information the attacker gains from observing Alice send messages via the anonymity system modelled as a threshold mix.

3.1 One Round, Alice Sends to One Receiver

Suppose $M = \{r\}$. Now consider the attacker observing one round of communication. The probability that we see r receiving exactly one message is q_r^B — Alice definitely sends her message to r, the other senders must send their messages to other receivers. The probability of any other receiver $r'(r' \in N \setminus \{r\})$ receiving exactly one message is $Bp_rq_r^{B-1}$.

Now define event X as "A particular user k receives one message" and an event Y as "$M = \{k\}$", i.e. k is the user Alice sends messages to. The event $Y|X$ is then "k is Alice's receiver given that k receives one message". Now note that what we calculated above is the probability of k receiving a message if he was Alice's receiver and the probability of k receiving a message if he was not. Thus, $\Pr[X|Y] = q_r^B$. Let us now look at the probability of Y being true. For this we need to consider what the adversary knows about the set M. We stated above that the attacker knows how many elements there are in M. If he knows nothing else, it is reasonable that he regards all possible sets of $|M|$ elements as equally likely. Thus, in our example here $\Pr[Y] = \frac{1}{|N|}$.

Now,

$$\Pr[X] = \Pr[X|Y]\Pr[Y] + \Pr[X|\neg Y]\Pr[\neg Y] = q_r^B \frac{1}{|N|} + Bp_rq_r^{B-1}\frac{|N|-1}{|N|}$$

We can now use Bayes' theorem to work out $\Pr[Y|X]$.

$$\Pr[Y|X] = \frac{\Pr[X|Y]\Pr[Y]}{\Pr[X]} = \frac{q_r^B \frac{1}{|N|}}{q_r^B \frac{1}{|N|} + Bp_r q_r^{B-1} \frac{|N|-1}{|N|}} =$$
$$= \frac{q}{q + Bp(|N|-1)} = \frac{1 - \frac{1}{|N|}}{1 - \frac{1}{|N|} + B - B\frac{1}{|N|}} = \frac{1}{1+B}$$

This is, of course, exactly what one would expect — after all, the attacker knew that M contains one receiver out of N with equal probability, and then observed that during one round of the mix (in which he knows Alice has participated) some particular receiver r has received one message. Without taking any further information into account (notably without knowing where all the other messages went), he can say that the probability that r is Alice's receiver is $\frac{1}{B+1}$.

A similar derivation shows that if all the messages during a round went to different receivers, the probability of any of them being Alice's receiver is still, as expected, $\frac{1}{B+1}$.

Now let us consider how much information the attacker gets if he observes someone receiving c messages.

The probability that r receives exactly c messages is

$$\binom{B}{c-1} p_r^{c-1} q_r^{B-c+1}$$

Note that c can be as high as $B+1$ requiring all the messages to go to the receiver r.

The probability of any other receiver $r'(r' \in N \setminus \{r\})$ receiving exactly c messages is:

$$\binom{B}{c} p_r^c q_r^{B-c}$$

Note that this becomes zero in the case of $c = B+1$ – the receiver who is not r cannot possibly receive all the messages from the mix as Alice sends her message to r. We calculate the probability that k who receives c messages is Alice's receiver r. From above:

$$\Pr[X|Y] = \binom{B}{c-1} p_r^{c-1} q_r^{B-c+1}$$

$$\Pr[X] = \binom{B}{c-1} p_r q_r^{B-c+1} \frac{1}{|N|} + \binom{B}{c} p_r^c q_r^{B-c} \frac{|N|-1}{|N|}$$

$$\Pr[Y|X] = \frac{\Pr[X|Y]\Pr[Y]}{\Pr[X]} = \frac{\binom{B}{c-1}}{\binom{B}{c-1} + \binom{B}{c}}$$

For example, if we have a system with ten potential receivers and B=10, i.e. the mix processes 11 messages during a round, then if the attacker sees two messages being sent to Bob during a round can deduce that Alice sent a message to Bob with probability $\frac{1}{11}$.

3.2 Several Rounds, Alice Sends to One Receiver

We now generalise this to any number of rounds l.

From before, we know that $\Pr[X|Y] = q_r^B$. Now, for many independent rounds (let X_l be "k receives exactly one message during each of the l rounds"), $\Pr[X_l|Y] = q_r^{Bl}$ and $\Pr[X_l|\neg Y] = B^l p_r^l q_r^{(B-1)l}$. A derivation very similar to above yields:

$$\Pr[Y|X_l] = \frac{q_r^l}{q_r^l + B^l p^l (|N|-1)} = \frac{(|N|-1)^{l-1}}{(|N|-1)^{l-1} + B^l}$$

This, of course, subsumes (and is consistent with) the above case for $l = 1$.

An example is in order. If everyone chooses uniformly from 10 different receivers (and Alice always sends to the same person), then just from the fact that Alice participated in two rounds of a threshold mix with threshold of five and Bob receives exactly one message during each of the two rounds, the attacker can deduce that Alice is talking to Bob with probability 0.36.

Of course, we have merely given the probability of Y given a very specific event X_l, but it is clear that the probability of Y given any event Z_l can be computed by merely multiplying the probabilities of Y given the event corresponding to each round. This is justified as the rounds are independent.

3.3 Several Rounds, Alice sends to Many Receivers

If Alice may send messages to more than one receiver, the situation changes slightly. We define the event X to be "there is a set K such that exactly one member of K receives one message during every round" and the event Y to be "Alice's set of receivers is the same as K or $M = K$". If the attacker knows the size of the set M then the number of possible sets K is $\binom{|N|}{|M|}$.

Now a simple derivation shows:

$$\Pr[Y|X] = \frac{q_r^l}{q_r^l + B^l p^l \left(\binom{|N|}{|M|} - 1\right)} = \frac{(|N|-|M|)^l}{(|N|-|M|)^l + B^l (|M|)^l \left(\binom{|N|}{|M|} - 1\right)}$$

Note that because M contains more than one element, $\Pr[Y]$ is $\frac{1}{\binom{|N|}{|M|}}$. The set of Alice's receivers is equally likely to be any of the sets of that size. Of course, if the attacker knew nothing about the size of M, the situation would have been rather different. The reader is invited to consider it[1].

We have shown how to calculate the probability of any set K of being Alice's receiver set, or, in other words, a probability distribution over all possible K. This can be used to compute the anonymity of M *as a whole* – following [10], one just computes the entropy of this probability distribution.

Modifying the example from the previous section shows us what effect increasing the size of M has. If Alice sends to one of two people at each round, then the probability of Alice's receiver set being $\{r, r'\}$ where r got a message during the first round and r' got a message during the second round is merely 0.009!

3.4 Some Generalisations and Remarks

The reader may have observed that confining Alice to choosing her receivers from a uniform distribution over M and the other senders – a uniform distribution over N is rather restrictive. Indeed, as long as all the other senders (Steves) choose their receivers using *the same* probability distributions, we may substitute different values for p_r and q_r in the equations above.

If the Steves send messages to receivers picked from different probability distributions (which are known to the attacker) the situation becomes more complicated. We consider it for the case of the pool mix in Section 4.

The attacker may well know more or fewer things about Alice's receiver set M. As we mentioned above, he may not know $|M|$, but assume that every possible M is equally likely. Alternatively, he may know a set N' such that $M \subseteq N' \subseteq N$. This knowledge too can be incorporated into the above calculations (but is a tedious exercise).

We have now given an account of the statistical disclosure attack on a anonymity system modelled by the threshold mix formally, giving a rigorous analysis underlying the attacks presented by Danezis [5] and Kesdogan et al [7, 1]. We go on to show how similar techniques can be used to derive similar results for a pool mix.

4 Formal Account of the Attack on the Threshold Pool Mix

We now turn our attention to the pool mix. During each round b of messages are input into the mix from the previous round. We call these messages the pool. A number B of messages are input from the senders. Out of the $B + b$ messages in the mix a random subset of size B is sent to their respective receivers. The remaining b messages stay in the pool for the next round.

[1] Naturally, the probability of any particular set K being Alice's set of receivers decreases and one might like to consider the probability that a receiver r is a member of Alice's set of receivers. We leave this for future work.

Unlike in the case of a threshold mix, the rounds of a pool mix are not independent. Therefore we must consider a complete run of the pool mix as one observation and try to extract information from it. A complete run starts when the mix comes online and its pool is empty and finishes when the mix is about to be shut down and has sent out all the messages in its pool out to the receivers.

We follow our running example of Alice choosing her receivers uniformly at random from M (call this probability distribution[2] $\boldsymbol{v}$) and all the other senders choosing uniformly from N, call this $\boldsymbol{u}$, $M \subseteq N$.

We make several assumptions:

- The messages which are in the pool at the beginning of the operation of the mix are distributed according to $\boldsymbol{u}$. We may think of the mix operator inserting these messages.
- The attacker is able to observe an entire run of the pool mix, from the very first round, 0, to the very last, k (when no messages remain in the pool). This may seem unrealistic; indeed any real attack of this form will rely on a smaller run and will necessarily yield an approximation to the results presented below. We take the "pure" case merely as an illustration.

First of all, let us define an observation of a pool mix over l rounds. Call O_i (for outputs) the multisets of receivers of round i and S_i the set of senders of round i[3]. One of the senders is Alice. Define S_0 to include all the initial messages in the pool and O_0 to include all the messages which ended up in the pool in the last round and got set out to receivers. Observe that $|S_0| = |O_l| = B + b$ and $i \neq 0 \Rightarrow |S_i| = B$ and $j \neq l \Rightarrow |O_j| = B$. Now construct $O = \{r_i | r \in O_i\}$ and $S = \{s_i | s \in S_i\}$. Given an observation $\text{Obs} = (S, O)$, there are many possible scenarios of what happened inside the anonymity system which would have been observed as Obs by the attacker. Indeed, a *possible scenario* λ is a relation on $S \times O$ such that each member of the S and O occurs in the relation exactly once and $(s_i, r_j) \in \lambda \Rightarrow i \leq j$. The relation λ represents a possible way senders could have sent messages to receivers which is consistent with Obs.

We illustrate this with a simple example. Suppose we have a pool mix with a threshold of two messages and a pool of one message which functioned for two rounds. The message which was in the pool initially came from the sender m, the mix itself, the other two messages came from A (Alice) and q. Thus, $S_0 = \{m, A, q\}$. $O_0 = \{r, r'\}$. At the next round which happens to be the last, messages from Alice and s arrived and messages for r, r' and r''' were sent, leaving the mix empty. Hence, $S_1 = \{A, s\}$, $O_1 = \{r, r', r'''\}$, $S = \{m_0, A_0, q_0, A_1, s_1\}$ and $O = \{r_0, r'_0, r_1, r''_1, r'''_1\}$. A possible scenario λ consistent with the observation (S, O) is: $\lambda = \{(m_0, r'''_1), (A_0, r_0), (q_0, r'_0), (A_1, r_1), (s_1, r''_1)\}$.

We can now compute the set of all possible scenarios which are compatible with the observation Obs. Call this set Λ. Take a $\lambda \in \Lambda$ and a set K such that

[2] Bold will consistently be used to indicate that the quantity is a vector describing a probability distribution.

[3] Until now we have not distinguished individual senders as all but Alice sent messages to receivers chosen according to the same probability distribution.

$|K| = |M|$. Define event Y as "$M = K$". If the possible scenario λ happened, then the attacker observes Obs — λ was observed by the attacker as Obs by definition — hence $\Pr[\text{Obs}|\lambda, K] = 1$. What is the probability of the possible scenario λ occurring if K was Alice's set of receivers? The possible scenario occurs if two things hold: if all the senders involved in this scenario picked their receivers in the same way as specified in λ and the mixing happened is such a way that the messages are sent to the receivers in accordance to λ. Hence

$$\Pr[\lambda|Y] = \left(\prod_{s\in S} p_s\right) \frac{1}{\binom{B+b}{b}^l}$$

where p_s is the probability of sender s sending a message to the receiver r such that $(s,r) \in \lambda$. Naturally, in the case we are considering above, $p_s = \frac{1}{|N|}$ if $s \neq$ Alice or $p_s = \frac{1}{|M|}$ if $s =$ Alice $\wedge\, r \in M \wedge (s,r) \in \lambda$ or $p_s = 0$ if $s =$ Alice $\wedge\, r \notin M \wedge (s,r) \in \lambda$. However, this approach is also applicable if the senders have different probability distributions p_s over N which are known to the attacker.

Having obtained $\Pr[\lambda|M]$, we can calculate $\Pr[\text{Obs}|M]$ and then, using Bayes' theorem as above, $\Pr[M|\text{Obs}]$. First,

$$\Pr[\text{Obs}|Y] = \sum_{\lambda\in\Lambda} \Pr[\text{Obs}|\lambda, M] \times \Pr[\lambda|M] = \sum_{\lambda\in\Lambda} \Pr[\lambda|M]$$

$$\Pr[\text{Obs}] = \sum_{K s.t. |K|=|M|} \sum_{\lambda\in\Lambda} \Pr[\text{Obs}|\lambda, Y] \Pr[Y]$$

Now,

$$\Pr[Y|\text{Obs}] = \frac{\Pr[\text{Obs}|Y]\Pr[Y]}{\Pr[\text{Obs}]} = \frac{\sum_{\lambda\in\Lambda} \Pr[\lambda|M] \binom{|N|}{|M|}}{\sum_{K s.t. |K|=|M|} \sum_{\lambda\in\Lambda} \Pr[\lambda|Y] \binom{|N|}{|M|}}$$

This enables us to compute the probability of a set K being Alice's receiver set. Unfortunately, this calculation requires generating all the possible scenarios, Λ. The number of these is clearly at least exponential in Bk. Hence a calculation which is based on all possible scenarios which could have happened inside the mix is not feasible for any practical run of a pool mix. In the next section we make some simplifying assumptions and show that it is possible to extract some information out of this scenario efficiently.

5 Efficient Statistical Attack on the Pool Mix

This attack is a modification of the attack presented in [5] to apply in the case of the pool mix. It is worth noting that the threshold mix is a special example of a pool mix, with no messages feeding forward to the next mixing round. Figure 1 illustrates the model used for the attack.

As before, one of the senders, Alice, is singled out to be the victim of the attack. Each time she has to send a message, she selects a recipient randomly out of a probability distribution described by the vector $\boldsymbol{v}$ over all possible N receivers in the system. Alice does not send in each round (as was the case in the model described in [5]) but only sends at rounds described by the function $s(k)$. Depending on whether it is a round when Alice sends or not, $B-1$ or B other senders respectively, send a message. They each choose the recipient of their messages, each independently, according to a probability distribution $\boldsymbol{u}$ over all possible recipients N. The initial b messages present in the pool at round 1 are also destined to recipients chosen independently according to the same probability distribution $\boldsymbol{u}$.

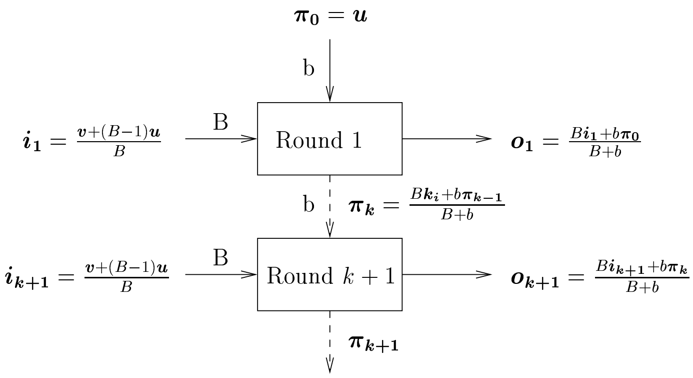


Fig. 1. The pool mix model and the probability distributions defined

6 Approximating the Model

We are going to define a series of approximations. These approximations distance the generalised statistical disclosure attack from other exact attacks, but allow the adversary to make very quick calculations and to decrease the anonymity of Alice's set of recipients.

We will first model the input distribution $\boldsymbol{i_k}$ of recipient of messages of each round k as being a combination of the distributions $\boldsymbol{u}$ and $\boldsymbol{v}$. Depending on whether Alice sends a message or not the component $\boldsymbol{v}$ will be present.

$$\boldsymbol{i_k} = \begin{cases} \frac{\boldsymbol{v}+(B-1)\boldsymbol{u}}{B} & \text{if } s(k) = 1 \\ \boldsymbol{u} & \text{if } s(k) = 0 \end{cases} \tag{1}$$

$\boldsymbol{i_k}$ is a vector modelling the distribution of messages expected after a very large number of rounds with the input characteristic of input round k. Depending on whether Alice is sending at round k, ($s(k)$ being equal to one), the appropriate distribution is used to model this input.

At the same time we model the output of each round k, and name it $\boldsymbol{o_k}$. This output is the function of the input distribution at the particular round k and the distribution of recipients that is forwarded to the present round via the pool. We call the distribution of recipients that are in the pool $\boldsymbol{\pi_{k-1}}$. The output distribution of each round can then be modelled as

$$\boldsymbol{o_k} = \frac{B\boldsymbol{i_k} + b\boldsymbol{\pi_{k-1}}}{B + b} \tag{2}$$

By definition $\boldsymbol{\pi_0} = \boldsymbol{u}$ and for all other rounds the distribution that represents the pool has no reason to be different from the distribution that represents the output of the round. Therefore $\boldsymbol{\pi_k} = \boldsymbol{o_k}$.

The attacker is able to observe the vector $\boldsymbol{s}$ describing the rounds at which Alice is sending messages to the anonymous communication channel. The adversary is also able to observe for each round the list O_k of receivers, to whom messages were addressed.

The generalised statistical disclosure attack relies on some approximations:

- The set of receivers at round O_k can be modelled as if they were each independently drawn samples from the distribution $\boldsymbol{o_k}$ as modelled above.
- The outputs of the rounds are independent from each other, and can be modelled as samples from the distribution $\boldsymbol{o_k}$.

Using the samples O_k we will try to infer the the distributions $\boldsymbol{o_k}$ and in turn infer the distribution $\boldsymbol{v}$ of Alice's recipients.

One can solve Equation 2 for a given function $s(k)$ and calculate $\boldsymbol{o_k}$ for all rounds k. Each distribution $\boldsymbol{o_k}$ is a mixture of $\boldsymbol{u}$, the other senders' recipients, and $\boldsymbol{v}$ Alice's recipients. The coefficient x_k can be used to express their relative weights.

$$\boldsymbol{o_k} = x_k\boldsymbol{v} + (1 - x_k)\boldsymbol{u} \tag{3}$$

By combining Equations 1 and 2 one can calculate x_k as:

$$x_k = \sum_{i \leq k, s(i)=1} \left(\frac{b}{B+b}\right)^{(i-1)} \frac{B}{B+b}\frac{1}{B} \tag{4}$$

This x_k expresses the relative contribution of the vector $\boldsymbol{v}$, or in other words Alice's communication, to each output in O_k observed during round k. When seen as a decision tree, each output contained in O_k has a probability $(1 - x_k)$ of being unrelated to Alice's set of recipients, but instead be drawn from another participant's distribution $\boldsymbol{u}$.

6.1 Estimating $\boldsymbol{v}$

The aim of the attack is to estimate the vector $\boldsymbol{v}$ that Alice uses to choose the recipients of her messages. Without loss of generality we will select a particular recipient Bob, and estimate the probability v_{Bob} Alice selects him as the recipient.

We can calculate the probability of Bob being the recipient of Alice for each sample we observe in O_k. We denote the event of Bob receiving message i in the observation O_k as $O_{ki} \to \text{Bob}$. Given our approximations we consider that the particular message O_{ki} was the outcome of sampling $\boldsymbol{o_k}$ and therefore by using equation 3 we can calculate the probabilities.

$$\Pr[O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k] = (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}}) \tag{5}$$

$$\Pr[\neg O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k] = 1 - (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}}) \tag{6}$$

As expected, Bob being the recipient of the message is dependent on the probability Alice sends a message v_{Bob} (that is Bob's share of $\boldsymbol{v}$), the probability others have sent a message u_{Bob} (which is Bob's share of $\boldsymbol{u}$) and the relative contributions of Alice and the other's to the round k, whose output we examine.

Now applying Bayes' theorem to Equations 5 and 6 we estimate p.

$$\begin{aligned}
&\Pr[v_{\text{Bob}}|O_{ki} \to \text{Bob}, u_{\text{Bob}}, x_k] = \\
&\frac{\Pr[O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k]\Pr[v_{\text{Bob}}|u_{\text{Bob}}, x_k]}{\int_0^1 \Pr[O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k]\Pr[v_{\text{Bob}}|u_{\text{Bob}}, x_k] \mathrm{d}v_{\text{Bob}}} \\
&d \sim (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}})\Pr[\text{Prior } v_{\text{Bob}}]
\end{aligned}$$

$$\begin{aligned}
&\Pr[v_{\text{Bob}}|\neg O_{ki} \to \text{Bob}, u_{\text{Bob}}, x_k] = \\
&\frac{\Pr[\neg O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k]\Pr[v_{\text{Bob}}|u_{\text{Bob}}, x_k]}{\int_0^1 \Pr[\neg O_{ki} \to \text{Bob}|v_{\text{Bob}}, u_{\text{Bob}}, x_k]\Pr[v_{\text{Bob}}|u_{\text{Bob}}, x_k] \mathrm{d}v_{\text{Bob}}} \\
&\sim (1 - (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}}))\Pr[\text{Prior } v_{\text{Bob}}]
\end{aligned}$$

Note that we choose to ignore the normalising factor for the moment since we are simply interested in the relative probabilities of the different values of v_{Bob}. The $\Pr[\text{Prior } v_{\text{Bob}}]$ encapsulates our knowledge about v_{Bob} before the observation, and we can use it to update our knowledge of v_{Bob}. We will therefore consider whether each message observed has been received or not by Bob and estimate v_{Bob} considering in each step the estimate of v_{Bob} given the previous data as the *a priori* distribution[4]. This technique allows us to estimate the probability distribution describing v_{Bob} given we observed R_k messages sent to Bob in each round k respectively.

$$\begin{aligned}
&\Pr[v_{\text{Bob}}|(x_1, R_1)\ldots(x_l, R_l), u_{\text{Bob}}] \\
&\sim \prod_k (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}})^{R_k}(1 - (x_k v_{\text{Bob}} + (1 - x_k)u_{\text{Bob}})))^{(B - R_k)}
\end{aligned}$$

[4] Since we are calculating relative probabilities we can discard the *a priori* since it is the uniform distribution over $[0, 1]$

The calculation above can be performed for each receiver in the system to estimate the likelihood it is one of Alice's receivers. The resulting probability distributions can be used as an indication of who Alice is communicating with, and their standard deviations can be used to express the certainty that this calculation provides.

7 Evaluation of the Attack

Figure 2 shows the set of probability distributions for 60 receivers. In this case we take the the probability distribution $\boldsymbol{u}$ to be uniform over all receivers and Alice to be choosing randomly between the first two receivers and sending messages for a thousand consecutive rounds (the mix characteristics in this case were $B = 10, b = 0$, namely it was a threshold mix). Figure 3 shows the same data for a pool mix with characteristics $B = 30, b = 15$. Note that the receivers 1 and 2 are Alice's and their respective v_1 and v_2 have different characteristics from the other receivers.

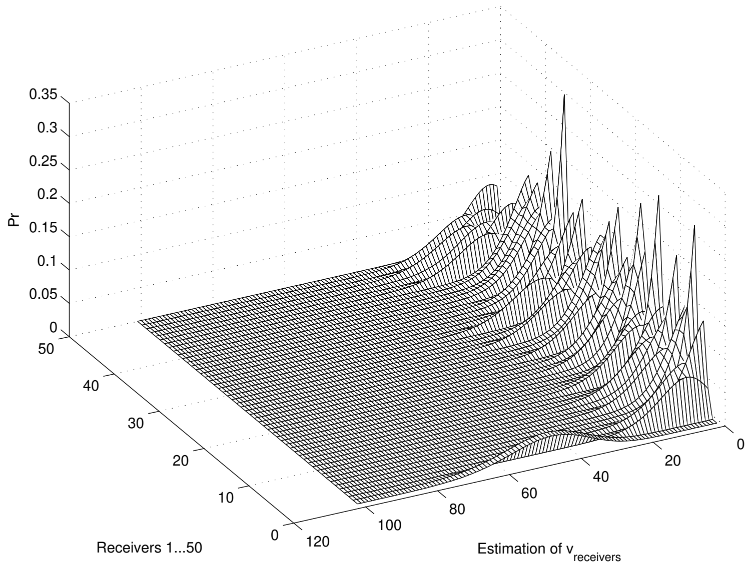


Fig. 2. Comparing the distributions of v_{receiver} for $B = 10, b = 0$

The same information can be more easily visualised if we take the average of all the distributions of receivers that do not belong to Alice, and compare them with the receivers of Alice. Figures 4(a) and 4(b) show the distributions

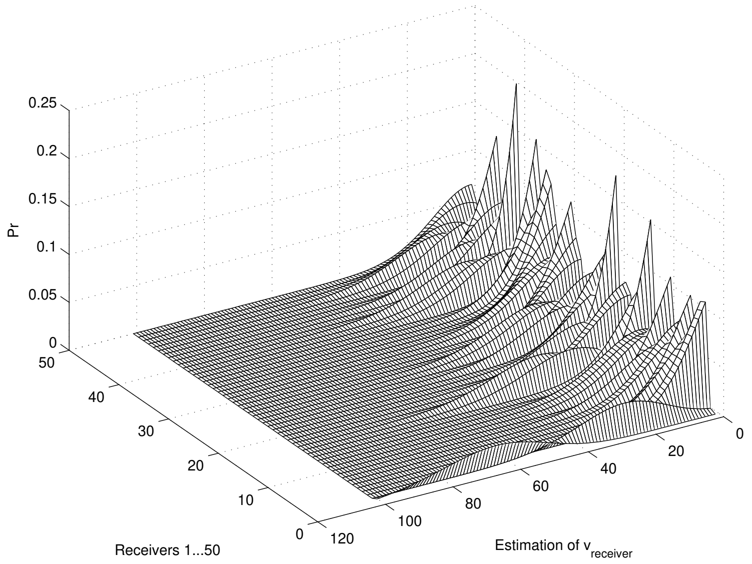


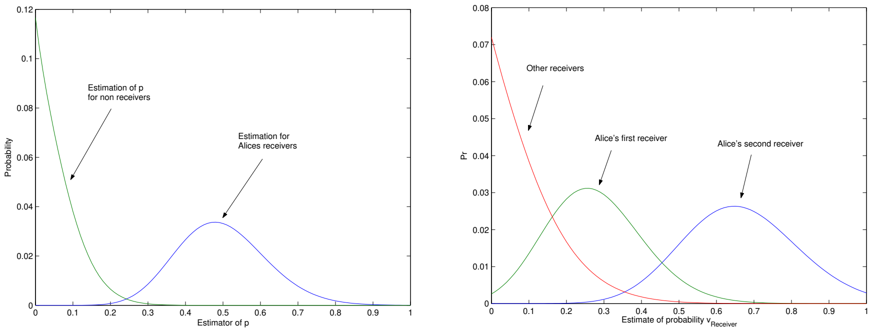
Fig. 3. Comparing the distributions of $v_{\mathrm{receiver}}$for $B = 30, b = 15$

of Alice's receivers and the averaged distributions of other receivers. The curves can be used to calculate the false positive rates, namely the probability a receiver has been attributed to Alice but is actually not in Alice's set, and false negative, namely a receiver wrongly being excluded from Alice's set of receivers.

It is unfortunate that we do not yet have analytic representations for the means and variances of the distribution describing v_{receiver}. Such representations would allow us to calculate the number of rounds for which Alice can send messages, given a particular set of mix characteristics, without being detected with any significant degree of certainty. The attack presented allows an attacker to understand where they stand, and how much certainty the attack has lead to, by numerically calculating them. On the other hand the network designer must simulate the behaviour of the network for particular characteristics to get some confidence that it does not leak information.

8 Conclusions

In this paper we presented a thorough account of attacks which consider repeated communication and the attacker's knowledge of it. First we gave some analytical results which enable the attacker to compute the probability of a set being Alice's set of receivers, and therefore the anonymity of that set of receivers. Then we presented a similar result for the pool mix. However, computing the



(a) Comparing the distributions of v1 and others. B=10, b=0

(b) Comparing the distributions of v1, v2 and others. B=30,b=15

probabilities in this case is expensive, and we resorted to using approximations to yield an efficient attack against a pool mix. The approximations were validated by simulations; the results show that the attack is powerful as well as efficient. This is an important and unfortunate result for the designers of anonymity systems.

References

1. Dakshi Agrawal, Dogan Kesdogan, and Stefan Penz. Probabilistic Treatment of MIXes to Hamper Traffic Analysis. In *Proceedings of the 2003 IEEE Symposium on Security and Privacy*, May 2003.
2. Oliver Berthold, Andreas Pfitzmann, and Ronny Standtke. The disadvantages of free MIX routes and how to overcome them. In H. Federrath, editor, *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*, pages 30–45. Springer-Verlag, LNCS 2009, July 2000.
3. David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 4(2), February 1981.
4. George Danezis. Mix-networks with restricted routes. In Roger Dingledine, editor, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2003)*. Springer-Verlag, LNCS 2760, March 2003.
5. George Danezis. Statistical disclosure attacks. In Samarati Katsikas Gritzalis, Vimercati, editor, *Proceedings of Security and Privacy in the Age of Uncertainty, (SEC2003)*, pages 421–426, Athens, May 2003. IFIP TC11, Kluwer.
6. Claudia Diaz and Andrei Serjantov. Generalising mixes. In Roger Dingledine, editor, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2003)*. Springer-Verlag, LNCS 2760, March 2003.
7. Dogan Kesdogan, Dakshi Agrawal, and Stefan Penz. Limits of anonymity in open environments. In Fabien Petitcolas, editor, *Proceedings of Information Hiding Workshop (IH 2002)*. Springer-Verlag, LNCS 2578, October 2002.

8. Dogan Kesdogan, Jan Egner, and Roland Büschkes. Stop-and-go MIXes: Providing probabilistic anonymity in an open system. In *Proceedings of Information Hiding Workshop (IH 1998)*. Springer-Verlag, LNCS 1525, 1998.
9. Michael G. Reed, Paul F. Syverson, and David M. Goldschlag. Anonymous connections and onion routing. *IEEE Journal on Selected Areas in Communication Special Issue on Copyright and Privacy Protection*, 1998.
10. Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Roger Dingledine and Paul Syverson, editors, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*. Springer-Verlag, LNCS 2482, April 2002.
11. Andrei Serjantov, Roger Dingledine, and Paul Syverson. From a trickle to a flood: Active attacks on several mix types. In Fabien Petitcolas, editor, *Proceedings of Information Hiding Workshop (IH 2002)*. Springer-Verlag, LNCS 2578, October 2002.
12. Andrei Serjantov and Richard E. Newman. On the anonymity of timed pool mixes. In *Proceedings of the Workshop on Privacy and Anonymity Issues in Networked and Distributed Systems*, pages 427–434, Athens, Greece, May 2003. Kluwer.