Two-Sided Statistical Disclosure Attack

George Danezis, Claudia Diaz, and Carmela Troncoso

K.U. Leuven, ESAT/COSIC,
Kasteelpark Arenberg 10,
B-3001 Leuven-Heverlee, Belgium
{George.Danezis,Claudia.Diaz,Carmela.Troncoso}@esat.kuleuven.be

Abstract. We introduce a new traffic analysis attack: the Two-sided Statistical Disclosure Attack, that tries to uncover the receivers of messages sent through an anonymizing network supporting anonymous replies. We provide an abstract model of an anonymity system with users that reply to messages. Based on this model, we propose a linear approximation describing the likely receivers of sent messages. Using simulations, we evaluate the new attack given different traffic characteristics and we show that it is superior to previous attacks when replies are routed in the system.

1 Introduction

Anonymous communications systems have been studied since 1981, when David Chaum first proposed the mix [2]. Yet, it has been known for some time that anonymity systems, not offering full unobservability, are insecure against long term Disclosure [1] and Statistical Disclosure Attacks [3] (SDA).

In this work, we extend Statistical Disclosure Attacks [3] in order to model user's behavior that deviates from the standard model considered so far in the literature. We consider that users not only *send messages* to a list of contacts, but also *reply to received messages* with some probability. Despite the real-world significance of modeling systems that allow anonymous replies, this is the first in-depth study of their security.

An adversary deploying our *Two-sided Statistical Disclosure Attack* (TS-SDA) takes into account the fact that some messages sent by a target user Alice are replies, in order to infer information on the set of Alice's contacts, and trace individual messages more effectively. This is done by combining information from sender and receiver anonymity sets when tracing replies.

We show through simulations that the Two-sided Statistical Disclosure Attacks give much better results than the traditional Statistical Disclosure Attacks, when tracing anonymized traffic that contains replies. We also evaluate how the effectiveness of our attacks is influenced by users' behavior (e.g., how often users reply, or how long it takes them to reply).

This paper is organized as follows: We review the relevant previous work concerning Disclosure Attacks in Sect. 2. Section 3 describes our model of the network and the users' behavior. Section 4 introduces our attacks, which are

N. Borisov and P. Golle (Eds.): PET 2007, LNCS 4776, pp. 30–44, 2007.

evaluated through simulations in Sect. 5. Finally, some thoughts on extending the attacks are discussed in Sect. 6, and we offer our conclusions in Sect. 7.

2 Background and Related Work

The field of anonymous communications started in 1981 with David Chaum's mix [2]. A mix is a relaying router that ensures, through cryptography and reordering techniques, that input messages cannot be linked to output messages, therefore providing anonymity. Based on these ideas, specialized cryptographic communication protocols exist for 'remailing' email messages, and the latest standard, Mixminion [4], allows users not only to send, but also to anonymously reply to email messages.

Despite the level of protection that mix networks provide, they still leak some information. An external observer is able to find out the identities (or at least network addresses) of mix users sending or receiving messages, as well as the exact time messages are sent and received. We can find in the literature a powerful family of *Disclosure Attacks* [1], first proposed by Kesdogan *et al.* [8]. These attacks allow an observer to learn the correspondents of each user and, in the long run, de-anonymize their messages. To counter these attacks, there is new research towards unobservable mix networks, such Nonesuch [7], where the users send their messages to the anonymity system as stegotext hidden inside Usenet postings.

The Disclosure Attack relies on a simple model for anonymous communications and user behavior. The target user, Alice, communicates only with her contacts (a subset of all possible recipients), while the other users send to all possible recipients with uniform probability. All users send their messages through a simple threshold mix [2]. This type of mix collects a certain number of messages (the threshold), and sends them to their destinations in a random order. An adversary only learns the public parameters of they system, and, in each round, who is sending and receiving messages. With no further information, the adversary can learn the set of contacts of Alice.

The two key shortcomings of the Disclosure Attack are its reliance on solving an NP-complete problem, and its sensitivity to deviations from the simple user behavior and communication models considered. The computational efficiency of the attack has been reduced by the Hitting Set Attack [9] where simple heuristics are used to evaluate the most likely set of Alice's contacts, which are tested to see if they are acceptable solutions. This leads to quick and exact results, yet the Hitting Set attack is still sensitive to even slight changes in the model. Allowing flexible models for user behavior and communication is key to understanding the security of real-world anonymous systems, since neither the systems nor the users' behavior fit perfectly idealized models.

A different style of attack, the Statistical Disclosure Attack (SDA) [3], considers the same user behavior and communication model, but reduces the computation complexity by using statistical models and approximations to reveal the same information to an attacker. The Statistical Disclosure Attack has been

extended to situations where the anonymity system is a pool mix, instead of a simple threshold mix [5]. This demonstrates that its underlying principles provide enough flexibility to successfully model complex anonymity systems. Even more complex models were evaluated by simulation in [10]. In this paper, we present a variant of the Statistical Disclosure Attack to de-anonymize traffic containing replies.

3 Mix Networks with Anonymous Replies

Building systems that allow full bi-directional anonymity, as first suggested by David Chaum in 1981, has been a key goal for anonymous communication designers. The latest remailer, Mixminion, offers this feature through the use of *single use reply blocks* (SURBs), cryptographic tokens that can be used to anonymously route back reply messages through a mix network. One of the key requirements of the Mixminion reply mechanism was to make replies indistinguishable from normal messages: an adversary observing a message leaving a user is not able to tell, from the bit string of the message or the processing that is applied to it in the first few mixes, if it is a reply or a normal message.

Our objective is to study the anonymity of messages in a network, such as Mixminion, that allows anonymous replies. For this reason, we modify the user behavior model of the Disclosure Attacks to accommodate replies, while considering that they are semantically indistinguishable from normal forward messages. In our new models, users send messages to the anonymity network either to initiate a discussion with one of their contacts, or to reply to a message they have received.

Following the spirit of previous Statistical Disclosure Attacks we describe many aspects of the system, such as the choice of conversation partners, the fact of replying to a message, and the time taken to send replies, as being sampled from *independent* probability distributions. We model users' initiation of new discussions as a Poisson process, and their choice of conversation partners is a sample out of a distribution of contacts. Messages are replied to with a known probability, and the time it takes to send the reply is exponentially distributed.

3.1 A Formal Model for Message Replies

We assume that there are N users in the system that send and receive messages. Each user $n \in [0, N-1]$ has a probability distribution $\boldsymbol{D_n}$ of sending to other users. We consider that the target user Alice has a distribution $\boldsymbol{D_A}$ of sending to a subset of her $k \in N$ contacts with uniform probability $1/k$. We have considered two models for the rest of the users: in the first case, they send with uniform probability $1/N$ to the N users. In the second case, they send to a subset $k \in N$, as Alice does. All users initiate discussions according to a Poisson process with rate λ_I. An array notation denotes the probability user n initiates a conversation with user m (i.e., $\boldsymbol{D_n}[m]$), and the distribution over all users should sum up to one (i.e., $\sum_{\forall m \in N} \boldsymbol{D_n}[m] = 1$). Figure 1 depicts our system model.

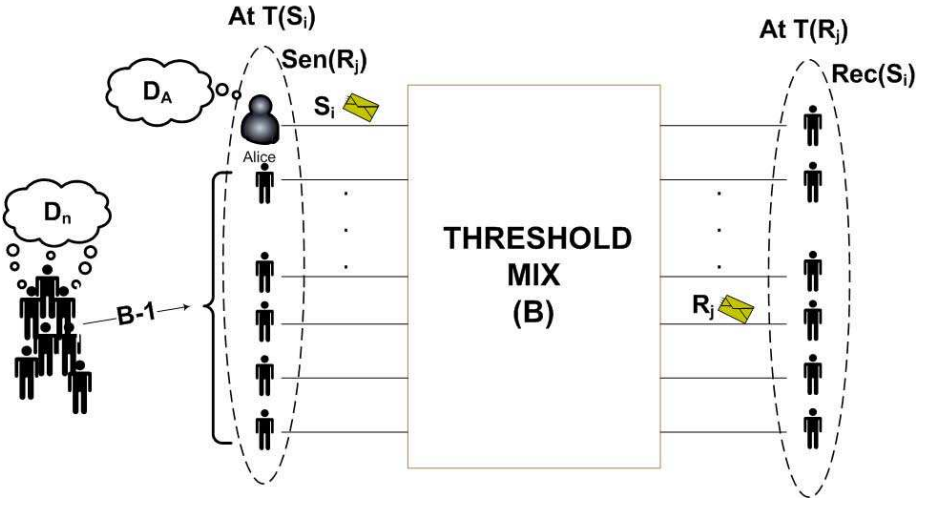

Fig. 1. System model

Alice, the target of the attack, is the only user that we initially model as replying to messages. She replies to received messages with probability r. If a reply is to be sent, it is done some time after the message it replies to was received. The reply delay is a sample from an exponential distribution with parameter (rate) λ_r. We have also considered a model in which all users reply to messages. We note that relationships are not symmetric in our system, and therefore the set of senders to which users reply to is not the same as the set of receivers considered in $\boldsymbol{D_n}$.

The adversary knows the overall replying behavior of Alice, namely the probability of a reply r and the reply delay rate λ_r. He also knows the number of users in the system N and the rate λ_I at which discussions are initiated by them. The objective of the adversary is to uncover $\boldsymbol{D_A}$.

A passive observer of the system can see when users send or receive messages. In our analysis, we only look at the messages received and sent by Alice. We denote K_s and K_r the total number of messages sent and received by Alice, respectively, within the time window $[0, t_{\max}]$ in which the system is observed. An adversary has accurate information about the time each message was sent or received, denoted as $T(S_i)$ and $T(R_j)$ for sent message i and received message j, respectively.

We consider that the adversary is observing all messages going in and out of the anonymity system, and can therefore calculate the probabilities describing the likely receivers or senders of each message. We denote the distribution over all N potential senders for a received message R_j as $\mathbf{Sen}(\boldsymbol{R_j})$, and over the receivers of a sent message S_i as $\mathbf{Rec}(\boldsymbol{S_i})$. We use an array notation to denote the probability of individual senders or receivers (e.g., $\mathbf{Sen}(\boldsymbol{R_j})[n]$ for the probability

that user n sent the message R_j received by Alice). As expected, the probabilities over all possible users should sum up to one ($\sum_{\forall n \in N} \mathbf{Sen}(\boldsymbol{R_j})[n] = 1$).

Aside from the sender and receiver distributions for messages, the attacker needs to know the relative contribution of Alice's messages to the anonymity sets. By contribution we mean the extent to which inputs S_i from Alice in the mix affect the receiver distribution $\mathbf{Rec}(\boldsymbol{S_i})$. Assuming that Alice sends α_r messages in round r, we denote the relative contribution messages from Alice as $\frac{\alpha_r}{B}$, and the contributions for others as $\frac{B-\alpha_r}{B}$, as this is the contribution of her α_r messages input into a threshold mix with parameter (i.e., threshold) B. Note that an equivalent quantity can be calculated without difficulty for other types of anonymity systems such as pool mixes.

4 The Two-Sided Statistical Disclosure Attack

Before presenting the Two-sided Statistical Disclosure Attack we will present the standard Statistical Disclosure Attack in terms of our formal model. A summary of all the notation is given in Table 1:

Table 1. Variables used in the model and the attacks

Name	Description
N	Number of users in the system
$\boldsymbol{D_A}$	The distribution of contacts of Alice
$\boldsymbol{D_n}$	The distribution of contacts of other users
λ_I	The rate of message initiations
r	Probability a message is replied to
λ_r	The rate at which messages are replied to
B	The threshold of the mix
$t_{\max}$	The total observation time
K_s, K_r	The total number of messages Alice sends and receives
S_i, $T(S_i)$	Alice's i^{th} sent message and the time it was sent
$\mathbf{Rec}(\boldsymbol{S_i})$	The receiver distribution for message S_i
R_j, $T(R_j)$	Alice's j^{th} received message and the time it was received
$\mathbf{Sen}(\boldsymbol{R_j})$	The sender distribution for message R_j
α_r	The number of messages sent by Alice in batch round r
Z_I	The expected volume of discussion initiations for each unit of time
Z_r	The expected volume of replies for a unit of time
Z_{rj}	The expected volume of replies to R_j
$\boldsymbol{I_{ij}}$	The intersection of distributions ($\mathbf{Sen}(\boldsymbol{R_j})$ and $\mathbf{Rec}(\boldsymbol{S_i})$) of messages R_j and S_i

4.1 The 'Traditional' Statistical Disclosure Attack

The traditional Statistical Disclosure Attack (SDA) works by observing the receiver anonymity sets of all messages sent by Alice, and aggregating them to infer the probability distribution $\boldsymbol{D_A}$. The messages in the receiver anonymity set $\mathbf{Rec}(\boldsymbol{S_i})$ of each message sent by Alice are assumed to be drawn from a

distribution that is a mixture between the contacts of Alice ($\boldsymbol{D_A}$) and the contacts of everyone else $\boldsymbol{D_n}$:

$$\mathbf{Rec}(\boldsymbol{S_i}) \sim \frac{1}{B}\boldsymbol{D_A} + \frac{B-1}{B}\boldsymbol{D_n} \tag{1}$$

The distributions describing the contacts of the rest of the users are approximated by using a uniform distribution U over all the possible senders. The adversary estimates the distribution $\boldsymbol{D_A}$ after a number observations K_s as:

$$\widehat{\boldsymbol{D_A}} \approx \frac{1}{K_s} \sum_{\forall i \in [0, K_s - 1]} [B \cdot \mathbf{Rec}(\boldsymbol{S_i}) - (B-1) \cdot \boldsymbol{D_n}] \tag{2}$$

The estimation $\widehat{\boldsymbol{D_A}}$ can then be used to infer the likelihood of the receiver corresponding to Alice in each round, by calculating:

$$\mathbf{Rec}(\boldsymbol{S_i})' = \frac{\mathbf{Rec}(\boldsymbol{S_i}) \cdot \widehat{\boldsymbol{D_A}}}{|\mathbf{Rec}(\boldsymbol{S_i}) \cdot \widehat{\boldsymbol{D_A}}|} \tag{3}$$

The key advantage of the statistical versions of the Disclosure Attack is their speed. It requires $\mathcal{O}(K_s)$ vector additions to estimate $\boldsymbol{D_A}$, and a further $\mathcal{O}(K_s)$ vector inner product calculations to get the estimates for the receiver of each round. Since vectors $\mathbf{Rec}(\boldsymbol{S_i})$ are sparse, both operations can be done very efficiently, and in parallel.

The downside of statistical attacks is that they are not exact. They do not take into account the basic constraint that a message can only be sent by one sender. This may lead to wrong results if too few samples are used to estimate $\boldsymbol{D_A}$.

4.2 The Two-Sided Statistical Disclosure Attack

The *Two-sided Statistical Disclosure Attack* (TS-SDA) takes into account the messages received by Alice (and the information about their potential senders), as well as the time of reception and sending of all messages. The aim of the attack is twofold: to estimate the distribution of contacts of Alice $\boldsymbol{D_A}$, and to infer the receivers of all the messages sent by Alice (i.e., forward messages she has initialized, and replies to messages she has received).

As in the Statistical Disclosure Attack (SDA), we will consider the output of each round of mixing (i.e., the distribution of potential receivers corresponding to each message) as the outcome of a mixture distribution. The components of this mixture are: the distribution $\boldsymbol{D_A}$ of contacts of Alice, the distribution $\boldsymbol{D_n}$ of the other senders, and the potential recipients of replies. Therefore, we need to *approximate* the relative weight of the contribution of each of these distributions to compute the receiver distribution.

Weight of normal messages. Let us consider a specific message, S_i sent by Alice. What is the relative probability of it being a discussion initiated by Alice, versus the probability of being a reply? We approximate this probability Z_I by

calculating the estimated number of discussions initiated by Alice that should occur at time $T(S_i)$, which is equal to:

$$\mathbb{E}(\text{Initiated discussion at } T(S_i)) = \frac{K_s}{\lambda_I \cdot t_{\max}} \equiv Z_I \tag{4}$$

The rationale behind this approximation is the following: the adversary observes Alice sending K_s messages which are a-priori equally likely to be an initiated discussion. Given that Alice initiates messages with rate λ_I, we expect an average of $\lambda_I \cdot t_{\max}$ discussions to be initiated by her over the total observation time $t_{\max}$.

Weight of replies. Similarly, we want to estimate the expected number of replies that would be sent at time $T(S_i)$. This expectation depends on the times messages R_j have been received by Alice before $T(S_i)$, and it is approximated by:

$$\mathbb{E}(\text{Reply to } R_j \text{ at } T(S_i)) = r \cdot \frac{\exp_{\lambda_r}(T(S_i) - T(R_j))}{\sum_{\forall k.T(R_j)<T(S_k)} \exp_{\lambda_r}(T(S_k) - T(R_j))} \equiv Z_{rj} \tag{5}$$

$$\mathbb{E}(\text{Replies at } T(S_i)) = \sum_{\forall j.T(R_j)<T(S_i)} Z_{rj} \equiv Z_r \tag{6}$$

A reply to message R_j is only generated with probability r. If it is generated, then it corresponds to S_i with a certain probability Z_{rj}. This probability is computed by considering the likelihood that the reply was sent at $T(S_i)$, normalized by the likelihood of the reply corresponding to any message S_k, sent after R_j was received (i.e., $T(R_j) < T(S_k)$). We have assumed that messages are answered after a delay distributed exponentially with parameter λ_r (i.e., $\exp_{\lambda_r}(t) = \lambda_r \cdot e^{-\lambda_r t}$). Summing over all messages R_j in the past gives us the likelihood Z_r of message S_i being a reply.

Full model. If message S_i is a reply to message R_j, then we can get even more of information about its destination. We intersect the receiver distribution $\mathbf{Rec}(\boldsymbol{S_i})$ for sent message, S_i, and the sender distribution $\mathbf{Sen}(\boldsymbol{R_j})$ for received message R_j, and thus obtain a probability distribution $\boldsymbol{I_{ij}}$ which describes the likely receiver of S_i:

$$\boldsymbol{I_{ij}} = \frac{\mathbf{Rec}(\boldsymbol{S_i}) \cdot \mathbf{Sen}(\boldsymbol{R_j})}{|\mathbf{Rec}(\boldsymbol{S_i}) \cdot \mathbf{Sen}(\boldsymbol{R_j})|} \tag{7}$$

Given the different weights Z_I, Z_r and Z_{rj}, we can model the distribution of receivers corresponding to the round of a message S_i. We do so by combining the distributions $\boldsymbol{D_A}$, $\boldsymbol{D_n}$ and the intersection $\boldsymbol{I_{ij}}$ for the replies, while taking into account that Alice sends a total of α_r messages in round r:

The figure below depicts the rationale behind our model. We look at the receivers at the output of the round r of mixing when Alice sends messages $S_i \ldots S_{i-\alpha_r-1}$, and consider what information they convey about her. Each message coming out of this round of mixing could correspond a message sent by Alice,

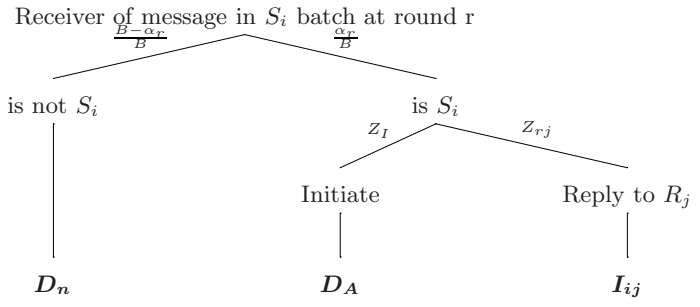

Fig. 2. An illustration of the components of equation 8

or to a message sent by another participant (drawn at random from $\boldsymbol{D_n}$). If the message corresponds to S_i, then it can either be a discussion initiation, drawn from $\boldsymbol{D_A}$, or a reply. Analytically we approximate the distribution $\mathbf{Rec}(\boldsymbol{S_i})$ as:

$$\mathbf{Rec}(\boldsymbol{S_i}) \sim \frac{\alpha_r}{B}\frac{Z_I \cdot \boldsymbol{D_A} + \sum_j Z_{rj}\boldsymbol{I_{ij}}}{Z_I + Z_r} + \frac{B-\alpha_r}{B}\boldsymbol{D_n} \tag{8}$$

The probabilities $\frac{\alpha_r}{B}$ and $\frac{B-\alpha_r}{B}$ describe the relative weight of Alice's α_r messages, versus the messages of the other senders (modeled by the distribution $\boldsymbol{D_n}$). The distributions $\boldsymbol{I_{ij}}$ are the normalized intersection of the potential set of senders of R_j with the set of possible receivers of S_i. Each of the I_{ij} are weighted by the factor Z_{rj}, which describes the likelihood S_i is indeed a reply to R_j.

As in the Statistical Disclosure Attack, we solve (8) for $\boldsymbol{D_A}$ and average the (very noisy) estimates for all sent messages S_i, in order to get the estimate $\widehat{\boldsymbol{D_A}}$:

$$\boldsymbol{D_A} \sim \frac{(B \cdot \mathbf{Rec}(\boldsymbol{S_i}) - (B-\alpha_r)\cdot \boldsymbol{D_n})(Z_I + Z_r) - \sum_j Z_{rj}\boldsymbol{I_{ij}}}{\alpha_r Z_I} \equiv \boldsymbol{C_i} \tag{9}$$

$$\widehat{\boldsymbol{D_A}} \approx \frac{1}{K_s}\sum_{\forall i} \boldsymbol{C_i} \tag{10}$$

Finally, the estimate $\widehat{\boldsymbol{D_A}}$ is in turn used to calculate the distribution of potential receivers for each message S_i:

$$\mathbf{Rec}(\boldsymbol{S_i})' \sim \left(\frac{\alpha_r}{B}\frac{Z_I \cdot \widehat{\boldsymbol{D_A}} + \sum_j Z_{rj}\boldsymbol{I_{ij}}}{Z_I + Z_r} + \frac{B-\alpha_r}{B}\boldsymbol{D_n}\right) \cdot \mathbf{Rec}(\boldsymbol{S_i}) \tag{11}$$

Our best guess for the actual receiver of message S_i is the intersection of the a-priori distribution of senders (given the volume of normal messages and replies sent by Alice) as well as their timing (the first term of (11)) and the actual receiver anonymity set for the round, $\mathbf{Rec}(\boldsymbol{S_i})$.

All the quantities needed to estimate $\mathbf{Rec}(\boldsymbol{S_i})'$ are known except for the distributions $\boldsymbol{D_n}$ describing the background traffic generated by other users. The

traditional Statistical Disclosure Attack, following the model of the Disclosure Attack, considers this distribution to be uniform U ($U[i] = 1/N$). Instead, in the TS-SDA we use a technique, first proposed by Mathewson and Dingledine [10], that estimates D_n from the traffic seen in the network in the rounds when Alice is not present.

5 Evaluation

We evaluate our new Two-Sided Statistical Disclosure Attack (TS-SDA) against the traditional Statistical Disclosure Attack (SDA) under various traffic conditions. In order to compare them and understand which parameters of the system affect their performance, we define a set of *standard parameters* that are summarized in Table 2.

Table 2. Standard parameters of the experiments

Name	Value	Description
N	1000	Number of participants
k	20	Number of contacts of Alice
B	100	Threshold of the mix
$t_{\max}$	4000	Observation time
λ_I	0.1	Initiation rate
r	0.5	Reply probability
λ_r	0.5	Reply delay rate

These parameters were chosen to depict an average system: the threshold is large enough to accommodate a good fraction of senders and receivers (about 1/10) and nodes send enough messages and replies to illustrate our techniques. Note that the rate at which replies are sent ($\lambda_r = 0.5$) is higher than the rate at which discussions are initiated ($\lambda_I = 0.1$). The choice of this parameter was based on the intuition that replies are sent much faster (with respect to the time of reception of the message that originated them) than messages initiated by a user (with respect to the last initiation).

In our analysis so far we have assumed nothing about Alice's distribution $\boldsymbol{D_A}$. For the sake of simplifying our experiments, we have assumed that the probability mass is distributed equally between k contacts of Alice, meaning that Alice chooses at random between them when she wants to initiate a discussion. Again, the ratio between the number of Alice's contacts k and the total number of users N reflects values observed in a medium size systems ($k/N = 2\%$). It is important to note though, that the statistical attacks should work with arbitrary $\boldsymbol{D_A}$ (although the time needed to discard the unlikely components of $\boldsymbol{D_A}$ would be larger.)

The final output of the TS-SDA attack is a probability distribution $\mathbf{Rec}(\boldsymbol{S_i})'$ for each message S_i that Alice has sent. These distributions describe the belief of the adversary as to who is the receiver of message S_i. We evaluate our attacks

by looking at the *rank* that the real receivers of S_i have in the distributions $\mathbf{Rec}(\boldsymbol{S_i})'$. The rank is the *number of receivers* in distribution $\mathbf{Rec}(\boldsymbol{S_i})'$ that have at least the same probability as the real receiver, and would therefore mislead the adversary in its attempts to trace the message[1]. Intuitively, this is equivalent to ordering receivers according to their probabilities, and using the position of the real receiver as a metric. Low ranks show that the attack is more successful.

In each round of the attack, we have a collection of ranks, one for each message S_i the adversary wants to trace. This distribution of ranks is represented in our graphs using box plots containing information about their maximum value, first quartile (Q_1), median, third quartile (Q_3) and maximum value. The box plots also depict outliers; i.e., ranks p that are very far from the rest of the distribution ($p > Q_3 + 1.5(Q_3 - Q_1)$ or $p < Q_1 - 1.5(Q_3 - Q_1)$). The box plots and outliers give a good overview of the tails of the rank distribution, which is crucial in our evaluation.

Fig. 3, compares the performance of the Statistical Disclosure Attack (SDA) to the Two-Sided Statistical Disclosure Attack (TS-SDA), as a user is observed for more time using the standard parameters. While the accuracy of both attacks increases with time, the TS-SDA always provides better results than the SDA. After 4000 ticks the TS-SDA classifies the correct sender within the 20 first candidates 3/4 of the time (for the SDA it is within ~ 35, 3/4 of the time).

It is important to explain why the key difference between the TS-SDA and the SDA can only be seen at the tail of the distributions, while their first quartile (Q_1) and median are about the same. For this, we need to understand better the strengths of each attack. Figure 4 shows the effectiveness of both attacks in tracing discussion initiation messages and replies (in a system using the standard parameters). We we can see that the TS-SDA and the SDA perform equally well in de-anonymizing discussion initiations (the SDA often performs slightly better). However, the main strength of the TS-SDA is its effectiveness in uncovering the recipients of replies, while the simple SDA is remarkably bad at it.

This explains in Fig. 3 the fact that the two attacks have the same first quartile and median: most messages in the system are discussion initiations, and therefore the attacks perform equally well for them. The difference appears only for the minority of messages that are replies, giving the SDA distribution of ranks a much heavier tail.

As we have seen, the TS-SDA outperforms the SDA mostly in its ability to trace replies. In Fig. 5 we show the sensitivity of both the SDA and TS-SDA to the reply parameters. We can see in the first graph that in the absence of

[1] Why not use metrics based on Information Theory? The traffic analysis models we use in the (TS-)SDA are only approximations of the real-world as well as the theoretical sending models. Therefore the distributions $\mathbf{Rec}(\boldsymbol{S_i})'$ give us only partial information about the actual receivers and are sometimes (as we show) just wrong. before applying the information theoretic metrics for anonymity, one would need to look at the mutual information between $\mathbf{Rec}(\boldsymbol{S_i})'$ and the actual receivers to understand the biases in the approximations.

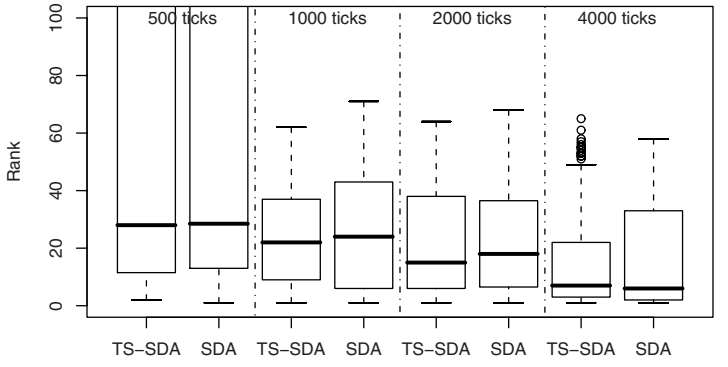

Fig. 3. Distribution of the ranks of the actual receivers after the Statistical Disclosure and Two-sided Statistical Disclosure attacks were applied. The estimation of $\widehat{\boldsymbol{D_A}}$ was after $500, 1000, 2000$ and 4000 ticks, and has a dramatic effect on the effectiveness of the attack. (Some outliers are not shown.)

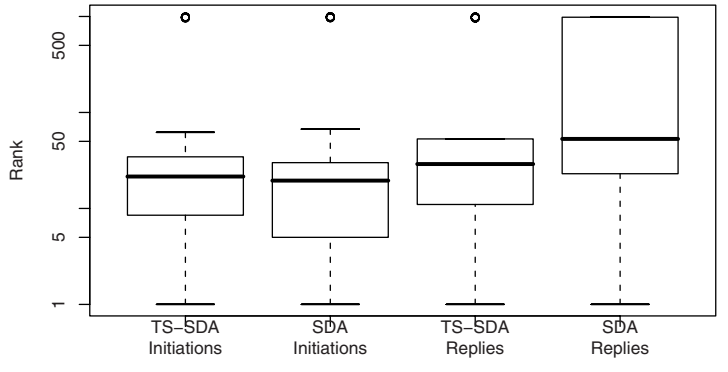

Fig. 4. (*Note the logarithmic scale.*) The effectiveness of the Two-sided Statistical Disclosure and traditional Statistical Disclosure attacks in de-anonymizing discussion initiation messages opposite to replies. The key advantage of the TS-SDA is its ability to correctly handle replies.

replies both attacks yield similar results. The graph labeled "Normal" shows the results for the standard parameters of our simulations, where our attack is more accurate that the SDA.

As messages are replied to at a slower rate ($\lambda_r = 0.025$), both attacks become less effective, since discussion initiations and replies become difficult to distinguish using timing information. The TS-SDA does not benefit any more from being able to intersect receiver and sender anonymity sets, and the standard SDA is subject to more noise. The consequences of this worsening are rather important, since it gives us a idea on how to resist the TS-SDA, and make use of the noise generated by the replies defensively.

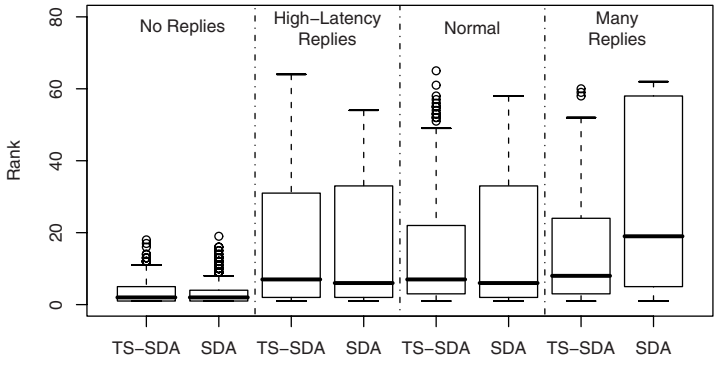

Fig. 5. The effectiveness of the Two-sided Statistical Disclosure and traditional Statistical Disclosure attacks for different types of traffic. With no replies, at high latency times ($\lambda_r = 0.025$), standard conditions and many replies ($r = 0.95$).

When we have more replies (graph "Many Replies," with probability of reply $r = 0.95$) the TS-SDA performs even better. The higher number of replies introduces noise for the SDA, worsening its performance, while the TS-SDA can de-anonymize them more effectively.

Finally, in Fig.6 we show the effects of changing the background traffic. The SDA is sensitive to other users having a non-uniform behaviour (all users send only to a limited set of $k = 20$ contacts – just as Alice does – instead of uniformly to all other $N = 1000$ users in the system), which worsens its accuracy. The TS-SDA, however, can handle this sort of traffic due to its estimation of the background noise.

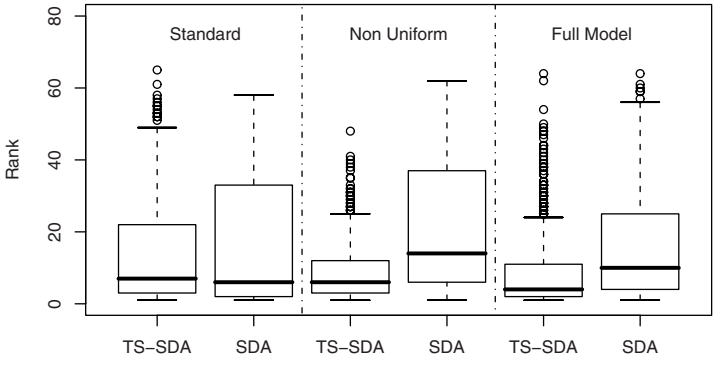

Fig. 6. The effect of the background traffic on the Two-sided Statistical Disclosure and traditional Statistical Disclosure Attacks. In the Non-Uniform case everyone sends to contacts only, and in the Full Model everyone replies.

Lastly, we show the results for the case in which all users reply to messages. In this case, Alice's correspondents reply to her, which in turn may generate further replies by her part. This symmetry increases the flow of traffic between Alice and her communication partners, thus making it easier to estimate her distribution of contacts. In the graph we can see that both the SDA and TS-SDA take benefit of this advantage, but that the impact on the latter is significantly stronger.

6 Discussion and Open Problems

The model presented is simple, allowing us to model a user replying to email and to illustrate our techniques to perform attacks. Yet its simplicity makes it deviate from real world usage in some ways.

First of all, we do not expect real users to initiate discussions in a way that can be approximated by a single Poisson process. There will definitely be fluctuations in the rate new discussions are initiated according to the time of day, the week day, the environment of the user and the user himself. The same is true for fluctuations in the reply delay and the probability of replying to messages.

Secondly, the parameters of the system are not likely to be independent of each other. Users are likely to read, write and reply to emails in bursts and not as they arrive.

Yet in some aspects the greatest shortcoming of the proposed model is that the probability of replying to an email is considered independent of the identity of the sender of the message. This is rather counterintuitive: one would expect Alice to be preferring to reply to her contacts rather than strangers (or even spammers) writing to her. Our model does not capture this aspect of two-way communication.

One way of modeling this aspect is to require contacts to be symmetric, meaning that if Alice has a certain probability to talk to Bob, then Bob should have the same probability to write to Alice (i.e. $\boldsymbol{D_A}[B] = \boldsymbol{D_B}[A]$). In this way, conversations will inevitably leak information about friendships. Yet, if this was the case, the messages sent out by Alice would basically follow her distribution $\boldsymbol{D_A}$ no matter if they were the product of a discussion initiation or a reply. This case would allow the attack to be no more complex than the simple statistical disclosure.

Instead, we allow the distributions D_n to be arbitrary, and replies to be independent of the sender of messages. This means that replies do not leak any information about the the distribution of Alice ($\boldsymbol{D_A}$), and are rather noise when the adversary attempts to estimate this distribution. We leave the specification of a model that takes into account the contacts of Alice when replying as an open problem for future work.

In our analysis we have assumed, for simplicity, that a threshold mix is used. Since only the probability distributions describing the likely senders and receivers of messages contribute to the attack, it should be possible to extend our analysis to any other anonymity system (not offering full unobservability).

Finally, our model presupposes that there can be only one reply per message, forcing the total volume of replies to be at most a fraction of the discussion initiations. This may not be the case for many users that have reactive rather than proactive email habits, and prefer to reply to messages rather than initiating a discussion. In certain environments (like tech-support desks) this may be a more appropriate model. Again, extending the model to take into account such traffic patterns is an open problem.

7 Conclusions

The Two-sided Statistical Disclosure Attack (TS-SDA) is the first traffic analysis attack to be explicitly targeted at anonymous communication systems that allow anonymous, and indistinguishable, replies. It takes into account the existence of replies and the timing of messages to estimate the correspondents of a target user and to trace the messages they send. The attack we show is very fast, as it operates in time linear in the number of messages ($O(K_s)$) and only requires simple operations on vectors. It is also possible to execute it in parallel or in specialised hardware very efficiently.

We have assessed the effectiveness of the TS-SDA under different conditions. It performs best when the volume of replies is high, and the time it takes users to reply is short. In this case, it uses the timing correlations between the received messages and sent replies to de-anonymize them. On the other hand, it performs as well as the Statistical Disclosure Attack (SDA) when few or no replies are present.

An important observation is that the timing of the replies is critical to the security of the anonymity system. When users send replies long time after receiving a message, it is difficult to correlate them with the originating message. This means that the replies act as cover traffic for the discussion initiation and both TS-SDA and SDA perform worse than in the absence of replies. Therefore, our key conclusion is that secure anonymity systems should make replies not only cryptographically indistinguishable from normal messages, but also difficult to correlate in time with the messages that are being replied.

User contacts in our study are not symmetric: Alice initiating discussions with Bob, does not mean that Bob also initiates discussions with Alice. Yet real social networks are likely to exhibit such symmetries. In this case, replies leak information about a user's contacts that contribute to the success of both the TS-SDA but also the simple SDA.

Acknowledgements

This work was partially supported by the IWT SBO ADAPID project (Advanced Applications for e-ID cards in Flanders), the Concerted Research Action (GOA) Ambiorics 2005/11 of the Flemish Government and by the IAP Programme P6/26 BCRYPT of the Belgian State (Belgian Science Policy). George Danezis is funded by a research grant of the Katholieke Universiteit Leuven.

References

1. Agrawal, D., Kesdogan, D.: Measuring anonymity: The disclosure attack. IEEE Security & Privacy 1(6), 27–34 (2003)
2. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Commun. ACM 24(2), 84–88 (1981)
3. Danezis, G.: Statistical disclosure attacks. In: Gritzalis, D., De Capitani di Vimercati, S., Samarati, P., Katsikas, S.K. (eds.) SEC of IFIP Conference Proceedings, vol. 250, pp. 421–426. Kluwer, Dordrecht (2003)
4. Danezis, G., Dingledine, R., Mathewson, N.: Mixminion: Design of a type iii anonymous remailer protocol. In: IEEE Symposium on Security and Privacy, pp. 2–15. IEEE Computer Society Press, Los Alamitos (2003)
5. Danezis, G., Serjantov, A.: Statistical disclosure or intersection attacks on anonymity systems. In: Fridrich [6], pp. 293–308
6. Fridrich, J. (ed.): IH 2004. LNCS, vol. 3200, pp. 23–25. Springer, Heidelberg (2004)
7. Heydt-Benjamin, T.S., Serjantov, A., Defend, B.: Nonesuch: a mix network with sender unobservability. In: 2006 Workshop on Privacy in the Electronic Society, ACM Press, New York (2006)
8. Kesdogan, D., Agrawal, D., Penz, S.: Limits of anonymity in open environments. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 53–69. Springer, Heidelberg (2003)
9. Kesdogan, D., Pimenidis, L. : The hitting set attack on anonymity protocols. In: Fridrich [6], pp. 326–339
10. Mathewson, N., Dingledine, R.: Practical traffic analysis: Extending and resisting statistical disclosure. In: Martin, D., Serjantov, A. (eds.) PET 2004. LNCS, vol. 3424, pp. 17–34. Springer, Heidelberg (2005)