STATISTICAL DISCLOSURE ATTACKS

Traffic Confirmation in Open Environments

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

Abstract An improvement over the previously known *disclosure attack* is presented that allows, using statistical methods, to effectively deanonymize users of a mix system. Furthermore the *statistical disclosure attack* is computationally efficient, and the conditions for it to be possible and accurate are much better understood. The new attack can be generalized easily to a variety of anonymity systems beyond mix networks.

Keywords: Statistical disclosure attack, traffic analysis, anonymity

1. Introduction

Since the concept of a mix network was introduced in (Chaum, 1981) the field of anonymous communications has been growing as new systems and attacks are proposed. All mix systems require that messages to be anonymized should be relayed through a sequence of trusted intermediary nodes. These nodes, called mixes, hide the correspondence between their input and output messages.

Although originally it was proposed that all participants should act as mixes, subsequent systems developed and deployed (Möller and Cottrell, 2000; Gulcu and Tsudik, 1996; Danezis et al., 2002) make a distinction between clients simply using the network, and mix nodes that form its core. This distinction is observable by an adversary, that sets as his goal to trace the ultimate recipient of messages injected in the network or trace back the originators of messages coming out of the network. Using information present at the edges of the mix network, where messages are injected or received, the attacker can try to link senders and receivers.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35691-4_52
D. Gritzalis et al. (eds.), *Security and Privacy in the Age of Uncertainty*

Such attacks are sometimes called *traffic confirmation* attacks since they do not rely on tracing messages through the network.

Another family of well known attacks against mix systems are intersection attacks (Berthold et al., 2000). These rely on the fact that different messages use the same route through the network to perform traffic analysis. Kesdogan presents an interesting variant of this attack in (Kesdogan et al., 2002), where it is applied to a whole anonymity system. He assumes that a particular user, Alice, sends messages only to a restricted set of recipients. He then observes that it is possible by observing the recipient anonymity sets attributed to her messages to extract information about their ultimate recipients. The attack is generalized by viewing mix networks or other systems providing anonymity as abstract mixes, since the attack does not rely upon any particular properties of mixing other than the unlinkability it provides.

In this paper we are going to briefly describe the disclosure attack as originally presented. A more efficient attack, the statistical disclosure attack, will then be presented. It requires less computational effort by the attacker and yields the same results. An analysis of the applicability and efficiency of the statistical disclosure attack, and a discussion of its relevance to other systems beyond the formal model is included.

2. The Disclosure Attack Revisited

The formal model on which the disclosure attack is based is quite simple. A single mix is used by b participants each round, one of them always being Alice, while the other $(b-1)$ are choosen randomly out of a total number of $N-1$ possible ones. The threshold of the mix is b so it fires after each of the round's participants has contributed one message. Alice chooses the recipient of her message to be a random member of a fixed set of m recipients. Each of the other participants sends a message to a recipient chosen uniformly at random out of N potential recipients. We assume that the other senders and Alice choose the recipients of their messages independently from each other. The attacker observes $R_1, \ldots, R_t$ the recipient anonymity sets corresponding to t messages sent out by Alice during t different rounds of mixing. The attacker then tries to establish which out of all potential recipients, each of Alice's messages was sent to.

The original attack as proposed by Kesdogan (Kesdogan et al., 2002) first tries to identify mutually disjoint sets of recipients from the sequence of recipient anonymity sets corresponding to Alice's messages. This operation is the main bottleneck for the attacker since it takes time exponential in the number of messages to be analyzed. The under-

lying method used is equivalent to solving the Constrains Satisfaction Problem which is well known to be NP-complete.

The second phase of the algorithm proposed intersects the disjoint sets found with anonymity sets of messages. When this intersection generates a set of only one element it is assumed that it is a correspondent of Alice.

3. The Statistical Disclosure Attack

We wish to use the same model as above to show that a statistical attack is possible that yields the set of potential recipients of Alice. In turn this set can be used to find the recipients of particular messages sent out by Alice.

We define as $\vec{v}$, the vector with N elements corresponding to each potential recipient of a messages in the system. We also set the values corresponding to the m recipients that might receive messages by Alice to $\frac{1}{m}$ and the others to zero, therefore requiring $|\vec{v}| = 1$. Observe that $\vec{v}$ is the probability distribution that is used by Alice to choose the recipient of its message for each round of the abstract mixing as described in the formal model above.

We also define $\vec{u}$ to be equal to the uniform distribution over all potential recipients N. Therefore all elements of $\vec{u}$ are set to be equal to $\frac{1}{N}$ with $|\vec{u}| = 1$. This vector represents the probability distribution used by all other senders to select their recipients' for each round of mixing.

The information provided to the attacker is a sequence of vectors $\vec{o_1}, \ldots, \vec{o_t}$ representing the recipient anonymity sets observed corresponding to the t messages sent by Alice. Each of $\vec{o_i}$ is the probability distribution assigning potential recipients to Alice's message during round i. The adversary will therefore try to use this information in order to infer $\vec{v}$ that, as described above, is closely linked to the set of recipients that Alice communicates with.

The principal observation underlying the statistical disclosure attack is that for a large enough set of observations t it holds true that (by using the Law of Large Numbers):

$$\bar{O} = \frac{\sum_{i=1\ldots t} \vec{o_i}}{t} = \frac{\vec{v} + (b-1)\vec{u}}{b} \tag{1}$$

It is therefore possible, just from the knowledge of the observations $\vec{o_1}, \ldots, \vec{o_t}$, the batch size b of the mix and the model $\vec{u}$ of other senders to calculate $\vec{v}$, the set of recipients of Alice:

$$\vec{v} = b\frac{\sum_{i=1\ldots t} \vec{o_i}}{t} - (b-1)\vec{u} \tag{2}$$

When the vector $\vec{v}$ is reconstructed by the adversary it can then be used to give an indication on the particular communications partners of Alice in a round k. The attacker simply multiplies each element of the $\vec{v}$ vector with each element of the observation $\vec{o_k}$ of round k, and normalizes the resulting vector.

$$\vec{r_k} = \frac{\vec{v} \cdot \vec{o_k}}{|\vec{v} \cdot \vec{o_k}|} \tag{3}$$

The elements with highest probability out of $\vec{r_k}$ are the most likely recipients of Alice's message k.

The statistical disclosure attack therefore allows an attacker to identify all possible recipients m of Alice's messages and even further to establish the precise recipients of particular messages in the formal model, with an arbitrary degree of confidence that, as we will see, depends on the number of observations t.

3.1 Applicability and Efficiency of the Statistical Disclosure Attack

The main drawback of the original disclosure attack was its reliance on solving an NP-complete problem. The statistical disclosure attack only relies on collecting observations and performing trivial operations on vectors, and therefore is computationally cheap and scales very well. Therefore we foresee the collection of observations, and the calculation of anonymity sets corresponding to messages to be the main computational bottleneck of an attacker.

It is important to establish the limits of the statistical disclosure attack and calculate the number of observations that are necessary in order to reliably perform it. We observe that extracting the vector $\vec{v}$ is a typical signal detection problem. The problem therefore is to differentiate the signal of Alice from the noise introduced by the other senders. In this case the signal strength of Alice is $\frac{1}{m}t$ versus the noise strength of the other senders that is equivalent to $\frac{b-1}{N}t$. For the signal to noise ratio to be larger than one we require:

$$\frac{\text{Alice's Signal}}{\text{Noise Strength}} = \frac{\frac{1}{m}t}{\frac{1-b}{N}t} > 1 \Rightarrow m < \frac{N}{b-1} \tag{4}$$

The above bound on m provides the necessary condition for a mix system following the formal model to be susceptible to the statistical disclosure attack. It is interesting that Kesdogan arrives to the same result in (Kesdogan et al., 2002), but proves it in a different way, which

means that any system that is vulnerable to the disclosure attack is also susceptible to the attack presented here.

Given that the signal to noise ratio allows for the statistical disclosure attack to be performed, it is important to calculate how many observations t are necessary to reliably retrieve $\vec{v}$. This depends on the variance of the signal $\vec{v}$ and the noise $(b-1)\vec{u}$.

The observations in $\bar{O}$ corresponding to Alice's recipients have a mean proportional to $\mu_{\text{Alice}} = \frac{1}{m}t$ and a corresponding variance of $\sigma^2_{\text{Alice}} = \frac{m-1}{m^2}t$ while the noise has a mean of $\mu_{\text{Noise}} = \frac{1}{N}(b-1)t$ and a variance of $\sigma^2_{\text{Noise}} = \frac{N-1}{N^2}(b-1)t$. We should require a number of observations t large enough for the mean of the signal to be larger than the sum of the standard deviations, multiplied by an appropriate factor to provide us with a satisfactory confidence interval.

$$\mu_{\text{Alice}} - l\sigma_{\text{Alice}} > l\sigma_{\text{Noise}} \tag{5}$$

$$t > \left[ml\left(\sqrt{\frac{m-1}{m^2}} + \sqrt{\frac{N-1}{N^2}(b-1)}\right)\right]^2 \tag{6}$$

With $l = 2$ we have a 95% confidence of correct classification, when determining if a recipient is associated with Alice or not, while $l = 3$ increases the confidence to 99%.

4. Conclusions

The statistical disclosure attack does not simply provide a computational improvement over the disclosure attack, but also presents important new features. The conditions for it to be possible can be expressed in closed algebraic form, as presented above, and therefore no simulations are required to decide when it is applicable and effective.

An important improvement over the previous work is also the fact that the statistical disclosure attack can be applied when the probability distributions described by $\vec{v}$, $\vec{u}$ and $\vec{o_i}$ are not uniform, but are skewed. This extends the attack from being applicable to anonymity systems that create discrete anonymity sets, to probabilistic systems that provide anonymity described by the entropy of the anonymity sets, as presented in (Serjantov and Danezis, 2002). As a result the entropy of the vector $\vec{r_k}$ represents the anonymity that a message still has after the attack has been performed. Therefore the statistical disclosure attack is more general than the simple disclosure attack, and can be applied to other models beyond the formal model presented here.

More work can be done on modeling different senders and their corresponding vectors $\vec{u}$, to construct a more realistic formal model. Even if

all other senders besides Alice have their own small sets of recipients we foresee the statistical disclosure attack to still be applicable if different senders are involved in each round.

Acknowledgements. The author would like to thank Andrei Serjantov and Richard Clayton for their comments on this work.

References

Berthold, O., Pfitzmann, A., and Standtke, R. (2000). The disadvantages of free MIX routes and how to overcome them. In *Designing Privacy Enhancing Technologies, LNCS Vol. 2009*, pages 30–45. Springer-Verlag. http://www.tik.ee.ethz.ch/~weiler/lehre/netsec/Unterlagen/anon/disadvantages_berthold.pdf.

Chaum, D. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 4(2). http://www.eskimo.com/~weidai/mix-net.txt.

Danezis, G., Dingledine, R., Mathewson, N., and Hopwood, D. (2002). Mixminion: Design of a Type III Anonymous Remailer Protocol. Manuscript. http://seul.org/~arma/minion-design.ps.

Gulcu, C. and Tsudik, G. (1996). Mixing E-mail with Babel. In *Network and Distributed Security Symposium - NDSS '96*. IEEE. http://citeseer.nj.nec.com/2254.html.

Kesdogan, D., Agrawal, D., and Penz, S. (2002). Limits of anonymity in open environments. In *Information Hiding, 5th International Workshop*, Noordwijkerhout, The Netherlands. Springer Verlag.

Möller, U. and Cottrell, L. (2000). Mixmaster Protocol — Version 2. Unfinished draft. Available online on the World Wide Web. http://www.eskimo.com/~rowdenw/crypt/Mix/draft-moeller-mixmaster2-protocol-00.txt.

Serjantov, A. and Danezis, G. (2002). Towards an information theoretc metric for anonymity. In *Proceedings of the Privacy Enhancing Technologies Workshop 2002*, San Francisco, CA.