An improved statistical disclosure attack

Bin Tang*

Department of Computer Science,
California State University Dominguez Hills,
Carson, CA, USA
Email: btang@csudh.edu
*Corresponding author

Rajiv Bagai and Huabo Lu

Department of Electrical Engineering and Computer Science,
Wichita State University,
Wichita, KS, USA
Email: rajiv.bagai@wichita.edu
Email: hxlu@wichita.edu

Abstract: Statistical disclosure attack (SDA) is known to be an effective long-term intersection attack against mix-based anonymising systems, in which an attacker observes a large volume of the incoming and outgoing traffic of a system and correlates its senders with receivers that they often send messages to. In this paper, we further strengthen the effectiveness of this attack. We show, by both an example and a proof, that by employing a weighted mean of the observed relative receiver popularity, the attacker can determine more accurately the set of receivers that a user sends messages to, than by using the existing arithmetic mean-based technique.

Keywords: statistical disclosure attack; SDA; anonymity; traffic analysis; security.

Reference to this paper should be made as follows: Tang, B., Bagai, R. and Lu, H. (xxxx) 'An improved statistical disclosure attack', *Int. J. Granular Computing, Rough Sets and Intelligent Systems*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Bin Tang received his BS in Physics from Peking University, China in 1997, MS in Materials Science and Computer Science from Stony Brook University in 2000 and 2002, respectively, and PhD in Computer Science from Stony Brook University in 2007. He is currently an Assistant Professor in the Department of Computer Science at California State University Dominguez Hills. His research interests include data anonymity and anonymising systems in information security, data preservation in wireless ad hoc sensor networks, and data replication and job scheduling in grid/cloud computing.

Rajiv Bagai received his BS in Computer Science from the Birla Institute of Technology and Science (BITS), Pilani, India, and MS and PhD in Computer Science from the University of Victoria, Canada. He is presently an Associate Professor in the Department of Electrical Engineering and Computer Science at the Wichita State University, USA. His current research area is web anonymity, but in the past he has worked in logic programming and paraconsistent databases.

Huabo Lu received his BS in Computer Science and Technology from the Beijing Forestry University, China, and MS in Computer Networking from the Wichita State University, USA. He is currently a PhD student in the department of Electrical Engineering and Computer Science at the Wichita State University, USA. His research interests include computer networking, web anonymity and privacy.

This paper is a revised and expanded version of a paper entitled 'On the sender cover traffic countermeasure against an improved statistical disclosure attack' presented at 8th IEEE/IFIP International Conference on Embedded and Ubiquitous Computing (EUC '10), Hong Kong, China, 11–13 December 2010.

1 Introduction

A popular technique to implement an anonymity system is as a mix network, as proposed by Chaum (1981). A mix network is a collection of proxy nodes that relay messages between senders and, possibly overlapping, receivers connected to the network. These intermediate proxies are the fundamental source of the anonymity achieved by such systems.

Several attacks by adversaries on mix-based anonymity systems, along with possible countermeasures, have been studied. Back et al. (2001) and Raymond (2001) contain detailed lists of attacks. Of these attacks, the class of long-term intersection attacks is one of the strongest. In such an attack, a passive global adversary can correlate senders with receivers that they often send messages to, by observing messages that enter and leave the mix over a long period.

The statistical disclosure attack (SDA), proposed by Danezis (2003), is a member of this class of attacks, and is directed against a single sender. The attack aims to uncover the receivers related to that sender by keeping track of, among others, the observed relative popularity of the various receivers of the system. Mathewson and Dingledine (2004) give an extended version of this attack by removing some of the restrictions on the number of messages flowing through the mix in the original version of Danezis (2003). Our paper is an improvement on the extended SDA of Mathewson and Dingledine (2004). Our attack is based upon the *weighted* mean of the observed relative popularity of the receivers over time. We show that this results in a more accurate conclusion than one obtained by the *arithmetic* mean method of Mathewson and Dingledine (2004).

The rest of this paper is organised as follows. In Section 2, we present the basic mix anonymous network model and give an overview of the SDA technique (Mathewson and Dingledine, 2004). Section 3 presents our improved technique for SDA from the attacker's perspective and gives an example. We present a formal proof about the effectiveness of the improved SDA in Section 4 and conclude in Section 5.

2 Statistical disclosure attack (Danezis, 2003; Mathewson and Dingledine, 2004)

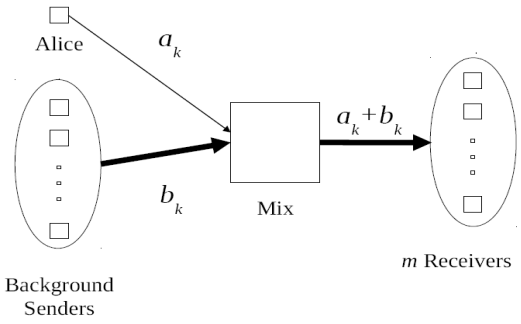
2.1 Mix anonymous network model

A mix network is a collection of proxy nodes that relay messages between a group of senders and a group of receivers. Via such a network, the messages sent by any of the senders can arrive at their receivers anonymously, thus hiding the senders' identity from the receivers. Some common attacks against such a network are to correlate the system's incoming messages with its outgoing ones by observing the entry and exit sequence numbers or by comparing bit patterns. To thwart such attacks, the mix network collects a certain number of encrypted messages in each round, decrypts them, and then transmits them simultaneously.

The SDA is targeted against a single sender, called Alice. All other senders are called *background* senders. The aim of this attack is to uncover, over a period of time, the set of receivers that *Alice* sends messages to, called Alice's *friends*. The attack is carried out by observing the number of messages sent or received by each sender and receiver in each round.

Let the total number of receivers in the system be m, and let a_k and b_k be the number of messages sent by Alice and background senders in round k, respectively. Figure 1 shows the message flow in the mix in round k. It should be noted that the attacker is unable to differentiate Alice's messages from those of the background senders received by any receiver, and can only observe the *total* number of messages received by each receiver in any round.

Figure 1 Message flow in round k



Let the $m \times t$ matrix A contain the number of Alice's messages received by each of the m receivers in the first t rounds, i.e., $A[i, j]$ is the number of Alice's messages received by receiver i in round j. Similarly, let the $m \times t$ matrix B contain the number of background senders' messages received by each of the receivers in the first t rounds. Let $R = A + B$, i.e., $R[i, j]$ is the total number of messages received by receiver i in round j. Note again that the attacker can observe only R, not A or B individually.

Let $\vec{\alpha} = (11...1)_{1\times m} \times A$ and $\vec{\beta} = (11...1)_{1\times m} \times B$ be two row vectors, which contain the total number of messages sent by Alice and background senders in each round, respectively. We have $a_k = \vec{\alpha}[k]$, and $b_k = \vec{\beta}[k]$. Let $\vec{r}_k$ be the column vector containing the number of messages received by each receiver in round k, that is, $\vec{r}_k$ is the k^{th} column

of *R*. Let $\vec{o}_k$ be the vector containing the *observed relative popularity* of each receiver in round *k*, which is defined as the fraction of the messages received by each receiver in round *k*, i.e., $\vec{o}_k = \vec{r}_k / (a_k + b_k)$.

Let $\vec{O}$ be the vector containing the *cumulative observed relative popularity* of each receiver after *t* rounds, that is, $\vec{O}$ is the average of the vectors $\{\vec{o}_k \mid 1 \le k \le t\}$:

$$\vec{O} = \frac{\sum_{k=1}^{t} \vec{o}_k}{t}.$$

SDA exploits the fact that for large values of *t*, the above actual receiver popularity approximates the expected one, which is formulated below.

Let $\bar{m}$ be the average number of messages sent by Alice in each round, i.e., $\bar{m} = \left(\sum_{k=1}^{t} a_k\right) \Big/ t$. Similarly, let $\bar{n} = \sum_{k=1}^{t} (a_k + b_k) \Big/ t$ be the average number of total messages sent in each round. Let $\vec{v}$ be the vector that contains the *relative degrees of friendship with Alice* of all the receivers, that is, the fraction of Alice's messages received by each receiver in all the rounds. The goal of SDA is to determine $\vec{v}$. Let $\vec{u}$ be the vector that contains the *observed receiver popularity with background senders* of all the receivers, i.e., $\vec{u}$ is the average of the vectors in the following set:

$$\{\vec{o}_k \mid 1 \le k \le t, a_k = 0\}.$$

In other words, $\vec{u}$ is obtained by observing the background senders' behaviour in rounds in which Alice does not send any messages. We assume that there are enough rounds in which Alice does not participate, which facilitates the computation of $\vec{u}$. This assumption is not an unreasonable one, because most senders connected to an anonymous system, such as users who browse the web, are online only some of the time and offline most of the time. If Alice is an ordinary user, $\vec{u}$ can be obtained easily during rounds that she is offline.

Recall that, the goal of SDA is to determine $\vec{v}$, which contains information about the *relative degrees of friendship with Alice* of all receivers. The *expected receiver popularity* can be expressed as:

$$\frac{\bar{m}\vec{v} + (\bar{n} - \bar{m})\vec{u}}{\bar{n}}.$$

The above expression is based upon the premise that of the *n* average number of messages sent in a round, $\bar{m}$ messages from Alice should reach receivers according to their degrees of friendship with Alice in $\vec{v}$, and the remaining $\bar{n} - \bar{m}$ messages from background senders should reach them according to their degrees of friendship, with all the background senders as a whole, in $\vec{u}$.

According to the Law of Large Numbers (Grimmett and Stirzaker, 1992), when *t* is large enough, the above expected popularity approximates $\vec{O}$, the observed one, i.e.,

$$\vec{O} \approx \frac{\bar{m}\vec{v} + (\bar{n} - \bar{m})\vec{u}}{\bar{n}}.$$

By rearranging, we get

$$\vec{v} \approx \frac{\bar{n}\vec{O} - (\bar{n} - \bar{m})\vec{u}}{\bar{m}}.$$

All values on the right side of the above equation can be obtained by observing the mix over time, thus making possible a reasonable estimate of the receivers' degrees of friendship with Alice.

3 An improved SDA using weighted mean

In the basic SDA proposed by Danezis (2003), the mix outputs a fixed number of messages in each round, exactly one of which is sent by Alice. Computing $\vec{O}$ as an arithmetic mean of the $\vec{o}_k$ vectors is all that is needed for that model. Also, in that model $\vec{u}$ corresponds to uniform distribution over all receivers, which is fixed and does not need to be computed.

The model presented in Section 2 is an extension developed by Mathewson and Dingledine (2004), in that the number of messages transmitted by the mix in each round can vary, Alice is allowed to send any number of messages in each round, and $\vec{u}$ need not be uniform. However, the SDA of Mathewson and Dingledine (2004) continues to employ the same arithmetic mean method for computing $\vec{O}$, which can be made more accurate by instead employing a weighted mean based upon the total number of messages output by the mix.

As an example, suppose *A* and *B* are the only receivers in the system. If in Round 1, *A* receives 1 message and *B* receives 3 messages, then $\vec{o}_1 = \langle 0.25, 0.75 \rangle$. Now, if in Round 2, A receives 300 messages and *B* receives 100 messages, then $\vec{o}_2 = \langle 0.75, 0.25 \rangle$. An arithmetic mean of these two vectors gives $\vec{O} = \langle 0.5, 0.5 \rangle$. On the other hand, a mean weighted by the total number of messages in each round would result in $\vec{O} = \left\langle \frac{1+100}{4+400}, \frac{3+100}{4+400} \right\rangle = \langle 0.745, 0.255 \rangle$, which better reflects the portion of the *total* number of messages received by the two receivers so far. As the intuition behind $\vec{O}$ is the *cumulative* observed relative popularity of receivers so far, its computation based upon weighted average is more in line with that intuition. In other words, $\vec{O}[i]$ should be calculated as the fraction of total messages received by receiver i in all the rounds.

We thus propose the following definition of $\vec{O}$:

$$\vec{O}[i] = \frac{\sum_{k=1}^{t} (a_k + b_k)\vec{o}_k[i]}{\sum_{k=1}^{t} (a_k + b_k)},$$

which can be simplified to be

$$\vec{O}[i] = \frac{\sum_{k=1}^{t} \vec{r}_k[i]}{\sum_{k=1}^{t} (a_k + b_k)}, \text{ for any receiver } i.$$

In order to better study the effectiveness of SDA by the above weighted mean method, we extend the example to a total of seven rounds, as shown in Table 1. The table shows the number of messages sent by Alice and the background senders to the two receivers, *A* and *B*, in each of the seven rounds. While such information is not available to the attacker, we use it to compare the effectiveness of the attack according to the old and new definitions of $\vec{O}$.

Table 1 Messages sent to receivers A and B

Round number	*Alice to A*	*Alice to B*	*Background to A*	*Background to B*
1	0	1	1	2
2	200	0	100	100
3	4	2	104	105
4	80	21	1,172	1,160
5	0	0	1,000	992
6	202	70	1,080	1,090
7	2	6	12	12
Total	488	100	3,469	3,461

The values $\bar{m}$ and $\bar{n}$ can be determined from Table 1 to be 84 and 1,074, respectively. From round 5, in which Alice does not send any messages, $\vec{u}$ is estimated to be ⟨0.502, 0.498⟩. By using these values of $\bar{m}, \bar{n},$ and $\vec{u}$ in equation (3), along with the value of $\vec{O}$ as the arithmetic mean of the $\vec{o}_k$ vectors, we get $\vec{v} = \langle 0.44, 0.56 \rangle$. The above value of $\vec{v}$ is misleading as it suggests *B* being more likely to be Alice's friend than A. On the other hand, our new definition of $\vec{O}$ as a weighted mean results in $\vec{v} = \langle 0.81, 0.19 \rangle$, which is much closer to its actual value $\left\langle \frac{488}{488+100}, \frac{100}{488+100} \right\rangle$ from these seven rounds, which is ⟨0.83, 0.17⟩.

4 Proof of the effectiveness of the improved SDA using weighted mean

We begin with the following definitions:

- $\overrightarrow{r_k^A}$: the vector containing the number of Alice's messages received by each receiver in round *k*.
- $\overrightarrow{r_k^B}$: the vector containing the number of background senders' messages received by each receiver in round *k*.
- $\vec{v}_{real}[i]$: the receiver *i*'s actual relative degree of friendship to Alice after *t* rounds:

$$\vec{v}_{real}[i] = \frac{\sum_{k=1}^{t} r_k^A[i]}{\sum_{k=1}^{t} a_k}.$$

- $\vec{O}_w[i]$ and $\vec{O}_a[i]$: the cumulative observed relative popularity for receiver *i* by weighted mean and arithmetic mean, respectively. From Sections 2 and 3, we have:

$$\vec{O}_w[i] = \frac{\sum_{k=1}^{t} \vec{r}_k[i]}{\sum_{k=1}^{t} (a_k + b_k)}, \tag{1}$$

$$\vec{O}_a[i] = \frac{\sum_{k=1}^{t} \vec{o}_k[i]}{t} = \frac{\sum_{k=1}^{t} \frac{\vec{r}_k[i]}{a_k + b_k}}{t}. \tag{2}$$

- $\vec{u}_w[i]$ and $\vec{u}_a[i]$: the cumulative observed relative popularity with the background senders, in rounds when Alice does not send messages, by weighted mean and arithmetic mean respectively.
- $\vec{v}_w[i]$ and $\vec{v}_a[i]$: the receiver *i*'s relative degree of friendship to Alice after *t* rounds, obtained by weighted mean and arithmetic mean respectively:

$$\vec{v}_w[i] \approx \frac{\bar{n}\vec{O}_w[i] - (\bar{n} - \bar{m})\vec{u}_w[i]}{\bar{m}}, \tag{3}$$

$$\vec{v}_a[i] \approx \frac{\bar{n}\vec{O}_a[i] - (\bar{n} - \bar{m})\vec{u}_a[i]}{\bar{m}}. \tag{4}$$

By substituting (1) into (3) and (2) into (4), we get

$$\vec{v}_w[i] \approx \frac{\sum_{k=1}^{t} \vec{r}_k[i] - \sum_{k=1}^{t} b_k \times \vec{u}_w[i]}{\sum_{k=1}^{t} a_k}, \tag{5}$$

$$\vec{v}_a[i] \approx \frac{\sum_{k=1}^{t} (a_k + b_k) \sum_{k=1}^{t} \frac{\vec{r}_k[i]}{a_k + b_k} - \sum_{k=1}^{t} b_k \times \vec{u}_a[i]}{\sum_{k=1}^{t} a_k}. \tag{6}$$

To show that by employing a weighted mean of the observed relative receiver popularity, the attacker can determine more accurately the set of receivers that a user sends messages to than using existing arithmetic mean-based one, we show that this is the case for each receiver *i*.

Theorem 1: For any receiver *i*, $|\vec{v}_{real}[i] - \vec{v}_w[i]| \le |\vec{v}_{real}[i] - \vec{v}_a[i]|$.

Proof: We need to show that $(\vec{v}_{real}[i] - \vec{v}_w[i])^2 \le (\vec{v}_{real}[i] - \vec{v}_a[i])^2$, i.e.,

$$2 \times \vec{v}_{real}[i] \times (\vec{v}_a[i] - \vec{v}_w[i]) \le (\vec{v}_a[i] - \vec{v}_w[i]) \times (\vec{v}_a[i] + \vec{v}_w[i]). \tag{7}$$

We first show that $\vec{v}_a[i] - \vec{v}_w[i] > 0$, and then we will only need to prove $2 \times \vec{v}_{real}[i] \le \vec{v}_a[i] + \vec{v}_w[i]$.

From (6) and (5),

$$\begin{aligned}
&\vec{v}_a[i]-\vec{v}_w[i] \\
&= \frac{\left(\sum_{k=1}^{t}(a_k+b_k)\sum_{k=1}^{t}\frac{\vec{r}_k[i]}{a_k+b_k}-\sum_{k=1}^{t}b_k\times\vec{u}_a[i]\right)-\left(\sum_{k=1}^{t}\vec{r}_k[i]-\sum_{k=1}^{t}b_k\times\vec{u}_w[i]\right)}{\sum_{k=1}^{t}a_k} \\
&= \frac{\left(\sum_{k=1}^{t}(a_k+b_k)\sum_{k=1}^{t}\frac{\vec{r}_k[i]}{a_k+b_k}-\sum_{k=1}^{t}\vec{r}_k[i]\right)+\sum_{k=1}^{t}b_k\times(\vec{u}_w[i]-\vec{u}_a[i])}{\sum_{k=1}^{t}a_k}
\end{aligned}$$

We have that

$$\begin{aligned}
\sum_{k=1}^{t}(a_k+b_k)\sum_{k=1}^{t}\frac{\vec{r}_k[i]}{a_k+b_k} &= \sum_{k=1}^{t}(a_k+b_k)\times\frac{\vec{r}_1[i]}{a_1+b_1}+\sum_{k=1}^{t}(a_k+b_k) \\
&\quad\times\frac{\vec{r}_2[i]}{a_2+b_2}+\cdots+\sum_{k=1}^{t}(a_k+b_k)\times\frac{\vec{r}_k[i]}{a_k+b_k} \\
&> \vec{r}_1[i]+\vec{r}_2[i]+\cdots+\vec{r}_t[i]=\sum_{k=1}^{t}\vec{r}_k[i].
\end{aligned}$$

Furthermore, since the observed receiver popularity with background senders does not depend on whether the attacker uses the weighted mean or arithmetic mean, we can assume that $\vec{u}_w[i]=\vec{u}_a[i]$. Therefore, we have that $\vec{v}_a[i]-\vec{v}_w[i]>0$, and from (7), we only need to prove $\vec{v}_a[i]+\vec{v}_w[i]\geq 2\times\vec{v}_{real}[i]$, shown as below:

$$\begin{aligned}
&\vec{v}_a[i]+\vec{v}_w[i] \\
&= \frac{\left(\sum_{k=1}^{t}(a_k+b_k)\sum_{k=1}^{t}\frac{\vec{r}_k[i]}{a_k+b_k}-\sum_{k=1}^{t}b_k\times\vec{u}_a[i]\right)+\left(\sum_{k=1}^{t}\vec{r}_k[i]-\sum_{k=1}^{t}b_k\times\vec{u}_w[i]\right)}{\sum_{k=1}^{t}a_k} \\
&> \frac{\left(\sum_{k=1}^{t}\vec{r}_k[i]+\sum_{k=1}^{t}\vec{r}_k[i]\right)-\sum_{k=1}^{t}b_k\times(\vec{u}_w[i]+\vec{u}_a[i])}{\sum_{k=1}^{t}a_k} \\
&= 2\times\frac{\sum_{k=1}^{t}\vec{r}_k[i]-\sum_{k=1}^{t}b_k\times\vec{u}_w[i]}{\sum_{k=1}^{t}a_k} \\
&= 2\times\frac{\sum_{k=1}^{t}\vec{r}_k[i]-\sum_{k=1}^{t}r_k^B[i]}{\sum_{k=1}^{t}a_k} \\
&= 2\times\frac{\sum_{k=1}^{t}r_k^A[i]}{\sum_{k=1}^{t}a_k} \\
&= 2\times\vec{v}_{real}[i].
\end{aligned}$$

■

5 Conclusions

We have presented an improved SDA by employing a weighted mean of the attacker's observations and proved that this makes the attack more accurate than that of Mathewson and Dingledine (2004), which employs arithmetic mean of the attacker's observations.

Acknowledgements

The research described in this paper was partially supported by the United States Navy Engineering Logistics Office contract no. N41756-08-C-3077.

References

Back, A., Möller, U. and Stiglic, A. (2001) 'Traffic analysis attacks and tradeoffs in anonymity providing systems', in *Proceedings of the 4th International Workshop on Information Hiding*, pp.245–257.

Chaum, D. (1981) 'Untraceable electronic mail, return addresses, and digital pseudonyms', *Communications of the ACM*, Vol. 24, No. 2, pp.84–88.

Danezis, G. (2003) 'Statistical disclosure attacks: traffic confirmation in open environments', in *Proceedings of Security and Privacy in the Age of Uncertainty (SEC 2003)*, pp.421–426.

Grimmett, G.R. and Stirzaker, D.R. (1992) *Probability and Random Processes*, 2nd ed., Clarendon Press, Oxford.

Mathewson, N. and Dingledine, R. (2004) 'Practical traffic analysis: extending and resisting statistical disclosure', in *Proceedings of the 4th Privacy Enhancing Technologies Workshop*, pp.17–34.

Raymond, J-F. (2001) 'Traffic analysis: protocols, attacks, design issues, and open problems', in *Proceedings of International Workshop on Design Issues in Anonymity and Unobservability*, pp.10–29.