Meet the Family of Statistical Disclosure Attacks

Simon Oya*, Carmela Troncoso† and Fernando Pérez-González*†
*Signal Theory and Communications Dept., University of Vigo
†Gradiant (Galician R&D Center in Advanced Telecommunications)

***Abstract*—Disclosure attacks aim at revealing communication patterns in anonymous communication systems, such as conversation partners or conversation frequency. In this paper, we propose a framework to compare between the members of the statistical disclosure attack family. We compare different variants of the Statistical Disclosure Attack (SDA) in the literature, together with two new methods; as well as show their relation with the Least Squares Disclosure Attack (LSDA).**

We empirically explore the performance of the attacks with respect to the different parameters of the system. Our experiments show that i) our proposals considerably improve the state-of-the-art SDA and ii) confirm that LSDA outperforms the SDA family when the adversary has enough observations of the system.

***Index Terms*—anonymity, mixes, disclosure attacks**

I. Introduction

Mixes constitute the basic building block of high-latency anonymous communication systems [1]. They act as a channel that hides the correspondence between incoming and outgoing messages, thus preventing a potential adversary from unveiling users' communication patterns (e.g. friendships, frequency).

There exist a wide variety of attacks that compromise the anonymity provided by mixes. In this paper, we revisit a particularly efficient family of attacks which is based on the Statistical Disclosure Attack (SDA) [2] and propose a framework that allows us to easily compare the attacks when performed on threshold mixes. We revisit Mathewson and Dingledine's generalization of the SDA and propose two new variants that outperform previous work. We also illustrate the relation between the SDA and the Least Squares Disclosure Attack (LSDA).

Additionally, we improve the theoretical analysis of the LSDA in [3] and extend it to one of the proposed variants of the SDA, which helps us understand the tradeoffs in performance versus complexity when attacking mixes.

The rest of the paper is organized as follows: we start with a brief overview of the current attacks on threshold mixes in Sect. II. In Sect. III, we introduce our system model and notation and then proceed with our revision of statistical disclosure attacks in Sect. IV. We perform a theoretical analysis of the attacks in Sect. V and validate our results in Sect. VI. Finally, we conclude in Sect. VII.

II. Previous work

The Disclosure Attack [4] relies on Graph Theory to reveal the exact set of friends of a user (Alice), seeking for mutually disjoint sets of receivers. This attack is known to be NP-complete but there exist other implementations that speed up the search [5].

Danezis proposed the Statistical Disclosure Attack (SDA) [2] as a faster alternative to the Disclosure Attack, which is based on the idea that it is possible to statistically isolate Alice's sending behavior after observing a large amount of her message's sets of receivers. The original SDA is limited to a specific scenario and was extended later to a more general user model and more complex mixing algorithms [6].

The Least Squares Disclosure Attack (LSDA) [3] models profiling as a least squares problem, minimizing the error between the actual number of output messages and a prediction based on the input messages.

In this work, we present an analysis of the family of statistical disclosure attacks [2], [6] and the LSDA [3], which share the goal of estimating the sending behavior of the users by combining the set of observations in an appropriate way. Other approaches that we leave out of our work are the Two-Sided SDA (TS-SDA) [7] and the Reversed SDA (RSDA) [8], which assume that users reply to messages; the Perfect Matching Disclosure Attack (PMDA) and the Normalized Statistical Disclosure Attack (NSDA) [9], which exploit that the relationship between sent and received messages is one-to-one; and the Bayesian inference-based approach, Vida [10].

III. System Model and Notation

Throughout the text, we will represent vectors using boldface lowercase characters and matrices using boldface capital letters. We will also use $\mathbf{1}_N$ to refer to the column vector whose N elements are equal to 1, and $\mathbf{1}_{N\times M}$ to the all-ones matrix of size $N \times M$. The superscript T will denote the transposing operation.

a) System Model: Our system consists of a population of N users, designated by index $i \in \{1, 2, ...N\}$, which communicate using a threshold mix. The system works as follows: every time a user i in our population wants to send a message to another user j, she encrypts the message and sends it to the mix. The mix receives and stores the messages until it has gathered t of them. Then, it transforms the messages cryptographically to change their appearance and outputs them in a random order; hence hiding the correspondence between incoming and outgoing messages. We call this process a *round* of mixing, and t is the *threshold* of the mix.

We denote the number of messages user i sends in round r by u_i^r. We define the column vector containing all the messages sent by user i up to round ρ as $\mathbf{u}_i = [u_i^1, u_i^2, ..., u_i^\rho]^T$, and the matrix of all observed inputs to the mix as $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N)$. Likewise, we denote the number of messages user j receives in round r by y_j^r and define $\mathbf{y}_j =$

$[y_j^1, y_j^2, ..., y_j^\rho]^T$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N)$. Additionally, we define $\tilde{u}_i^r$ as a binary representation of u_i^r, denoting whether there is at least one message sent by user i in round r ($\tilde{u}_i^r = 1$) or not ($\tilde{u}_i^r = 0$). We also define, $\tilde{\mathbf{u}}_i = [\tilde{u}_i^1, \tilde{u}_i^2, ..., \tilde{u}_i^\rho]^T$.

User i sends messages to their recipients according to her *sender profile* and her *sender frequency*. We define the sender profile of user i as $\mathbf{q}_i = [p_{1,i}, p_{2,i}, ..., p_{N,i}]^T$, where $p_{j,i}$ models the probability that user i sends a message to user j. The sender frequency f_i models the probability that a message arriving to the mix comes from user i ($f_i \geq 0$ for $i = 1, 2, ..., N$ and $\sum_{i=1}^N f_i = 1$). We also define the vector $\mathbf{p}_j = [p_{j,1}, p_{j,2}, ..., p_{j,N}]^T$ which shall come in handy later. We make no assumptions on the distribution of each sender profile, other than $p_{j,i} \geq 0$ for $i, j = 1, 2, ..., N$ and $\sum_{j=1}^N p_{j,i} = 1$ for $i = 1, 2, ..., N$.

We define the *uniformity* of the sender profile of user i as $\mu_i = 1 - \sum_{j=1}^N p_{j,i}^2$. The uniformity μ_i ranges from 0, when user i always sends messages to the same contact (i.e. $p_{k,i} = 1$, $p_{j,i} = 0$ for $k \in \{1, ..., N\}$ and $j \neq k$, $j = 1, ..., N$), to $\frac{N-1}{N}$, when she sends messages to all the other users equiprobably.

Finally, we define the *background traffic* of a user i as an aggregate of the traffic generated by all users but i. This way, vector $\mathbf{u}_b$ contains the messages sent by all users but i, $\mathbf{u}_b = \sum_{\substack{k=1 \\ k \neq i}}^N \mathbf{u}_k = \mathbf{1}_\rho \cdot t - \mathbf{u}_i$. The *background profile* is $\mathbf{q}_b = [p_{1,b}, p_{2,b}, ..., p_{N,b}]^T$ where $p_{j,b} = \sum_{\substack{k=1 \\ k \neq i}}^N \frac{f_k}{1 - f_i} \cdot p_{j,k}$ and the uniformity of this sender profile is denoted by μ_b. In all cases, user i will be clear from the context.

b) Adversary Model: We consider a global passive adversary that observes the system during ρ rounds. The adversary observes the identity of the users communicating through the mix and knows all the parameters of the system. We also assume that the adversary is not able to link any messages by their content, i.e. the cryptographic transformations do not leak information.

The goal of the adversary is to infer the sending behavior of the users in the system from the observations, i.e. to obtain an estimator $\hat{p}_{j,i}$ of $p_{j,i}$ given the input and output observations $\mathbf{U}$ and $\mathbf{Y}$.

IV. Revisiting the Family of Disclosure Attacks

A. The Original Statistical Disclosure Attack

Danezis introduced the original Statistical Disclosure Attack ($\texttt{SDA}_d$) in [2], which provides an estimator of $p_{j,i}$ under the assumptions that the user i does not send more than one message each round and the background traffic for that user is uniform, i.e. $p_{j,b} = \frac{1}{N}$ for $j = 1, 2, ..., N$.

Danezis claims that, by using the Law of Large Numbers, the mean of the observations y_j^r in the rounds where i has sent at least one message can be written as

$$\frac{\tilde{\mathbf{u}}_i^T \mathbf{y}_j}{\tilde{\mathbf{u}}_i^T \mathbf{1}_\rho} \approx p_{j,i} + (t-1) \cdot p_{j,b}\,, \tag{1}$$

and therefore an estimator for $p_{j,i}$ is

$$\hat{p}_{j,i}^{\texttt{SDA}_d} = \frac{\tilde{\mathbf{u}}_i^T \mathbf{y}_j}{\tilde{\mathbf{u}}_i^T \mathbf{1}_\rho} - (t-1) \cdot \hat{p}_{j,b}\,, \text{ with } \hat{p}_{j,b} = \frac{1}{N}\,. \tag{2}$$

In order to compare $\texttt{SDA}_d$ with its variants, note that we can write (1) as

$$\tilde{\mathbf{u}}_i^T \mathbf{y}_j \approx \tilde{\mathbf{u}}_i^T \mathbf{1}_\rho \cdot p_{j,i} + \tilde{\mathbf{u}}_i^T \left(\mathbf{1}_\rho \cdot t - \mathbf{1}_\rho\right) \cdot p_{j,b}\,. \tag{3}$$

B. Generalized Statistical Disclosure Attack

Mathewson and Dingledine extended Danezis' attack in [6], allowing user i to send multiple messages in a round and estimating the background from the observations.

Using this extension, (3) becomes

$$\tilde{\mathbf{u}}_i^T \mathbf{y}_j \approx \tilde{\mathbf{u}}_i^T \mathbf{u}_i \cdot p_{j,i} + \tilde{\mathbf{u}}_i^T \mathbf{u}_b \cdot p_{j,b}\,, \tag{4}$$

where we have just replaced the $\mathbf{1}_\rho$s which referred to the number of messages sent by user i in each round in (3) with the actual number of messages sent by i, $\mathbf{u}_i$, and $\mathbf{1}_\rho \cdot t - \mathbf{u}_i = \mathbf{u}_b$.

The background profile is estimated by computing the average number of messages received by j in the rounds where i does not participate and dividing by the total number of messages exiting the mix each round (t),

$$\hat{p}_{j,b} = \frac{1}{t} \cdot \frac{\left(\mathbf{1}_\rho - \tilde{\mathbf{u}}_i\right)^T \mathbf{y}_j}{\left(\mathbf{1}_\rho - \tilde{\mathbf{u}}_i\right)^T \mathbf{1}_\rho}\,. \tag{5}$$

We denote this attack by SDA0, whose estimator is

$$\hat{p}_{j,i}^{\texttt{SDA0}} = \frac{\tilde{\mathbf{u}}_i^T \mathbf{y}_j}{\tilde{\mathbf{u}}_i^T \mathbf{u}_i} - \frac{\tilde{\mathbf{u}}_i^T \mathbf{u}_b}{\tilde{\mathbf{u}}_i^T \mathbf{u}_i} \cdot \hat{p}_{j,b}\,. \tag{6}$$

C. Improvements in the Generalized SDA

The attack described in the previous section performs an average of the outputs in those rounds where user i sends at least one message in order to compute $\hat{p}_{j,i}^{\texttt{SDA0}}$, giving the same value to those outputs regardless of the actual participation of user i. We propose a new estimator, which we denote SDA1, that counts the outputs once for every message sent by user i, therefore giving more weight to those rounds where the number of messages sent by i is larger.

Using this approach, (4) becomes

$$\mathbf{u}_i^T \mathbf{y}_j \approx \mathbf{u}_i^T \mathbf{u}_i \cdot p_{j,i} + \mathbf{u}_i^T \mathbf{u}_b \cdot p_{j,b}\,, \tag{7}$$

where we have replaced the vector we used to select the rounds we were taking into account, $\tilde{\mathbf{u}}_i$, by the vector with the actual number of messages sent by i in each round, $\mathbf{u}_i$.

From (7), we get the following estimator,

$$\hat{p}_{j,i}^{\texttt{SDA1}} = \frac{\mathbf{u}_i^T \mathbf{y}_j}{\mathbf{u}_i^T \mathbf{u}_i} - \frac{\mathbf{u}_i^T \mathbf{u}_b}{\mathbf{u}_i^T \mathbf{u}_i} \cdot \hat{p}_{j,b}\,, \tag{8}$$

where $\hat{p}_{j,b}$ is estimated as in (5).

Note that the idea behind this estimator appears in [6] applied to other mixing algorithms. The analysis of SDA in [6] also features the idea of exploiting observations from rounds where user i appears as a sender in order to compute $\hat{p}_{j,b}$.

The latter idea inspires our second variant, denoted SDA2, which uses the observations from *all rounds* to get the background estimation. Following (7), we can write

$$\begin{cases} \mathbf{u}_i^T \mathbf{y}_j = \mathbf{u}_i^T \mathbf{u}_i \cdot \hat{p}_{j,i} + \mathbf{u}_i^T \mathbf{u}_b \cdot \hat{p}_{j,b} \\ \mathbf{u}_b^T \mathbf{y}_j = \mathbf{u}_b^T \mathbf{u}_i \cdot \hat{p}_{j,i} + \mathbf{u}_b^T \mathbf{u}_b \cdot \hat{p}_{j,b}\,. \end{cases} \tag{9}$$

If we define the $\rho \times 2$ matrix $\mathbf{U_{i,b}} = (\mathbf{u}_i, \mathbf{u}_b)$, the new estimator $\hat{p}_{j,i}^{\texttt{SDA2}}$ can be obtained by solving

$$\begin{pmatrix} \hat{p}_{j,i}^{\texttt{SDA2}} \\ \hat{p}_{j,b} \end{pmatrix} = \left(\mathbf{U_{i,b}}^T \mathbf{U_{i,b}}\right)^{-1} \mathbf{U_{i,b}}^T \mathbf{y}_j \,. \tag{10}$$

D. The Least Squares Disclosure Attack

The estimator in (10) uses the information from all outputs when estimating both $p_{j,i}$ and $p_{j,b}$. However, users' profiles are solved independently, compressing information in matrices $\mathbf{U_{i,b}}$. We can extend the idea in (9) considering that, when computing the sender profile of i, the background is formed by all the users but i. In that case, we would have N equations with N unknowns, which are

$$\mathbf{u}_i^T \mathbf{y}_j = \mathbf{u}_i^T \sum_{k=1}^{N} \left(\mathbf{u}_k \cdot \hat{p}_{j,k}\right) \text{ , for } i = 1, ..., N \,. \tag{11}$$

Presenting this system in matricial form, we have

$$\mathbf{U}^T \mathbf{y}_j = \mathbf{U}^T \mathbf{U} \hat{\mathbf{p}}_j \,. \tag{12}$$

Therefore, if $\mathbf{U}^T\mathbf{U}$ is not singular, we obtain the Least Squares Disclosure Attack (LSDA) estimator in [3],

$$\hat{\mathbf{p}}_j^{\texttt{LSDA}} = \left(\mathbf{U}^T\mathbf{U}\right)^{-1} \mathbf{U}^T \mathbf{y}_j \,. \tag{13}$$

V. Performance Analysis

In this section, we aim at deriving a theoretical expression for the *Mean Squared Error of sender profile* i, which we define as $\text{MSE}_i \doteq ||\mathbf{q}_i - \hat{\mathbf{q}}_i||^2 = \sum_{j=1}^{N} (p_{j,i} - \hat{p}_{j,i})^2$, for the described estimators. Due to space limitations, we reduce our analysis to SDA2 and LSDA.

We start by deriving an expression of MSE_i in LSDA. In order to do so, we first show, by using the law of total expectation together with $\text{E}\left\{\mathbf{y}_j|\mathbf{U}\right\} = \mathbf{U} \cdot \mathbf{p}_j$, that this estimator is unbiased, since

$$\text{E}\{\hat{\mathbf{p}}_j\} = \text{E}\left\{\text{E}\left\{\hat{\mathbf{p}}_j|\mathbf{U}\right\}\right\} = \text{E}\left\{\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T\text{E}\left\{\mathbf{y}_j|\mathbf{U}\right\}\right\} = \mathbf{p}_j \tag{14}$$

Using this fact, along with the law of total variance, we can write the covariance matrix of $\mathbf{p}_j$ as

$$\mathbf{\Sigma}_{\mathbf{p}_j} = \text{E}\left\{\mathbf{\Sigma}_{\mathbf{p}_j|\mathbf{U}}\right\} = \text{E}\left\{\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{\Sigma}_{\mathbf{y}_j|\mathbf{U}}\mathbf{U}\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\right\} \tag{15}$$

We model $\{u_1^r, ..., u_N^r\}$ together as a multinomial distribution with t trials and probabilities $\{f_1, ..., f_N\}$. In order to compute (15), we first assume that the number of observations is large enough, so that we can approximate $\mathbf{U}^T\mathbf{U} \approx \text{E}\{\mathbf{U}^T\mathbf{U}\} = \mathbf{R}_u \cdot \rho$, where $\mathbf{R}_u$ is the autocorrelation matrix of the input process,

$$\mathbf{R}_u = t\left[\mathbf{F} + (t-1)\,\mathbf{F}\mathbf{1}_{N\times N}\mathbf{F}\right] \tag{16}$$

where $\mathbf{F} = \text{diag}\{f_1, ..., f_N\}$.

Applying the matrix inversion lemma, we can write the inverse of this autocorrelation matrix as

$$\mathbf{R}_u^{-1} = \frac{1}{t}\left[\mathbf{F}^{-1} - \left(1 - \frac{1}{t}\right)\mathbf{1}_{N\times N}\right] \,. \tag{17}$$

Now that, using $\mathbf{U}^T\mathbf{U} \approx \mathbf{R}_u \cdot \rho$, the only term remaining inside $\text{E}\{\cdot\}$ in (15) is $\text{E}\{\mathbf{U}^T\mathbf{\Sigma}_{\mathbf{y}_j|\mathbf{U}}\mathbf{U}\}$. We model $y_j^r|\mathbf{U}$ as the sum of N binomial processes with u_i^r trials and probabilities $p_{j,i}$, for $i = 1, 2, ..., N$. Let $s_{j,k} = p_{j,k} \cdot (1 - p_{j,k})$ and $\mathbf{S}_j = \text{diag}\{s_{j,1}, ..., s_{j,N}\}$. Then, $\mathbf{\Sigma}_{\mathbf{y}_j|\mathbf{U}}$ is a diagonal matrix whose (r,r)-th element is $\left(\mathbf{\Sigma}_{\mathbf{y}_j|\mathbf{U}}\right)_{r,r} = \sum_{k=1}^{N} u_k^r s_{j,k}$. Operating,

$$\begin{aligned} &\text{E}\{\mathbf{U}^T\mathbf{\Sigma}_{\mathbf{y}_j|\mathbf{U}}\mathbf{U}\} = \\ &\rho\left\{\mathbf{F}\left(\eta_j t^{(3)}\mathbf{1}_{N\times N} + \mathbf{S}_j\mathbf{1}_{N\times N}t^{(2)} + \mathbf{1}_{N\times N}\mathbf{S}_j t^{(2)}\right)\mathbf{F}\right\} \\ &+ \rho\left\{\left(\eta_j t^{(2)}\mathbf{I}_{N\times N} + t\mathbf{S}_j\right)\mathbf{F}\right\} \end{aligned} \tag{18}$$

where $\eta_j = \sum_{k=1}^{N} f_k s_{j,k}$ and $t^{(n)} = t \cdot (t-1) \cdot ... \cdot (t-n+1)$.

Plugging (17) and (18) into (15) we get an approximation of $\mathbf{\Sigma}_{\mathbf{p}_j}$. Now, taking each of the diagonal elements of this matrix, which are $\text{Var}\{\hat{p}_{j,i}\}$ for $i = 1, ..., N$ and adding them along j to obtain $\text{MSE}_i = \sum_{j=1}^{N} \text{Var}\{\hat{p}_{j,i}\}$, we finally get

$$\text{MSE}_i^{\texttt{LSDA}} \approx \frac{1}{\rho}\left\{\left(f_i^{-1} - 1\right)\left(1 - \frac{1}{t}\right)\bar{\mu}_{\texttt{LSDA}} + \frac{f_i^{-1}}{t} \cdot \mu_i\right\} \tag{19}$$

where $\bar{\mu}_{\texttt{LSDA}} = \sum_{k=1}^{N} f_k \mu_k$ is the average uniformity of the sender profiles.

Following a similar approach, it can be shown for SDA2 that, when the number of observed rounds is large enough,

$$\text{MSE}_i^{\texttt{SDA2}} \approx \frac{1}{\rho}\left\{\left(f_i^{-1} - 1\right)\left(1 - \frac{1}{t}\right)\bar{\mu}_{\texttt{SDA2}} + \frac{f_i^{-1}}{t} \cdot \mu_i\right\} \tag{20}$$

where $\bar{\mu}_{\texttt{SDA2}} = f_i\mu_i + (1 - f_i)\mu_b$ is the average uniformity considering that there are only two users in the system: the user i and her background.

Note that the only approximations made to derive (19) and (20) were $\mathbf{U}^T\mathbf{U} \approx \text{E}\{\mathbf{U}^T\mathbf{U}\} = \mathbf{R}_u \cdot \rho$ and its equivalent with matrix $\mathbf{U_{i,b}}$. Therefore, these MSE estimators are more accurate as the number of observed rounds is large.

Given the definition of the background sending profile in Sect. III, it is easy to see that $\bar{\mu}_{\texttt{SDA2}} \geq \bar{\mu}_{\texttt{LSDA}}$, and therefore $\text{MSE}_i^{\texttt{SDA2}} \geq \text{MSE}_i^{\texttt{LSDA}}$, where the equality holds only when all users have the same sending profile. This proves that LSDA will eventually outperform SDA2 in terms of MSE when the attacker observes the system indefinitely.

VI. Evaluation

We evaluate the performance of the attacks in Sect. IV in terms of MSE_i, simulating a threshold mix system as described in Sect. III.[1] We exclude $\texttt{SDA}_d$ from this evaluation and use its generalization SDA0 instead.

We vary the number of users in the population N, the threshold t, the sending frequencies f_i, the number of rounds observed by the attacker ρ and the uniformity of the sending profiles μ_i.

[1] The simulator, written in Matlab, will be available upon request.

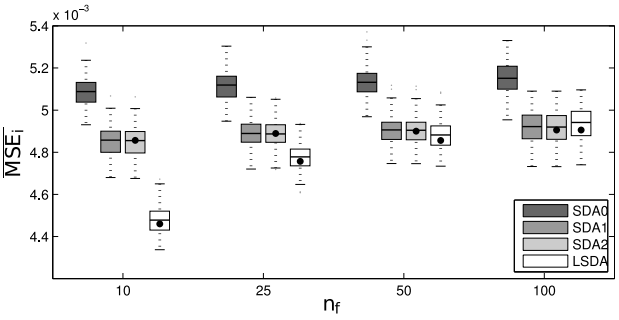

Fig. 1. Average MSE for the different attacks, as a function of the number of friends n_f of each user ($\rho = 20000$, $N = 100$, $f_i = 1/N$, $t = 10$).

A. Performance with respect to the uniformity μ_i

As we have shown in Sect. V, the uniformity of the sender profiles is a key parameter to show the difference in performance between SDA2 and LSDA. For simplicity, we assume that each user i has n_f friends to whom she sends messages uniformly, which are users $\bmod(i+k, N)$ for $k = 0, ..., n_f - 1$. This allows us to vary the uniformity of the sender profile of each user with a single parameter: $\mu_i = \frac{1-n_f}{n_f}$. We choose the number of friends n_f from $\{10, 25, 50, 100\}$ and, for each value, perform 100 repetitions of the experiment.

Figure 1 shows a box-and-whiskers plot of the average MSE per sender profile, $\overline{\text{MSE}_i}$. On the boxes, the central mark is the mean and the edges are the 25th and 75th percentiles. The black circles • represent the theoretical asymptotic values of the $\overline{\text{MSE}_i}$, from (19) and (20). Since ρ is finite, $\overline{\text{MSE}_i}$ does not coincide exactly with its theoretical value, although (19) and (20) reliably describe the accuracy of the attacks. As expected, when the uniformity of the sender profiles is low and the background uniformity μ_b is large, LSDA outperforms the other estimators, but as the uniformity of each user increases and therefore becomes closer to the background uniformity, the advantage of LSDA decreases. Also, note that the proposed estimators SDA1 and SDA2 outperform SDA0.

B. Performance with respect to the other parameters

Due to space limitations, we are not able to plot the results obtained when varying all the other parameters. We summarize the basic results next and refer to [3] for further information about LSDA. First, the MSE_i decreases with $1/\rho$ in each of these attacks, as in (19) and (20). Also, in every attack, the MSE_i is approximately proportional to the inverse of the sending frequency f_i^{-1}, due to the increasing difficulty of estimating the sender profile of a user when she rarely participates in the system. The threshold t has little influence on the MSE_i of SDA2 and LSDA but does, however, decrease the number of rounds that can be used to estimate the background (5) in SDA0 and SDA1, thus increasing the MSE_i in these estimators. Finally, we note that increasing N adds an extra error in LSDA which is not predicted by (19) and that stems from the matrix inversion in (13). This error can be reduced by increasing the number of rounds observed. This can be seen in Fig. 1, where the mean values of $\overline{\text{MSE}_i}$ obtained for LSDA are slightly above their asymptotic value.

The improvements in performance achieved by the more sophisticated versions of statistical disclosure come at the price of an increase in the computational cost. While SDA0 adds the observations where the user whose profile is being estimated has participated, SDA1 needs to perform an additional multiplication for each of these rounds. SDA2 has a higher computational cost since it requires solving a system of two equations for each user, and LSDA requires solving a linear system of N equations with N unknowns.

VII. Conclusions

In this work, we have introduced a framework to model the different attacks of the statistical disclosure family, showing how better results can be achieved when performing more complex operations with the observations from the system. We have formalized two new variants of the SDA, which we called SDA1 and SDA2, and showed that they significantly improve the state-of-the-art SDA in threshold mixes, SDA0. Furthermore, we have shown that the LSDA, introduced in [3], can be seen as an upgraded version of statistical disclosure that solves the problem jointly for all users.

We have also improved the previous theoretical analysis on LSDA and derived for the first time an expression which accurately approximates the error of SDA2. Our experiments confirm these theoretical results.

Acknowledgements

This research was supported by the European Union under project LIFTGATE (Grant Agreement Number 285901), the European Regional Development Fund (ERDF) and the Spanish Government under projects DYNACS (TEC2010-21245-C02-02/TCM) and COMONSENS (CONSOLIDER-INGENIO 2010 CSD2008-00010), and the Galician Regional Government under projects Consolidation of Research Units 2009/62, 2010/85.

References

[1] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–90, 1981.
[2] G. Danezis, "Statistical disclosure attacks," in *Security and Privacy in the Age of Uncertainty (SEC2003)*, 2003, pp. 421–426.
[3] F. Pérez-González and C. Troncoso, "Understanding statistical disclosure: A least squares approach," in *Privacy Enhancing Technologies*. Springer, 2012, pp. 38–57.
[4] D. Agrawal and D. Kesdogan, "Measuring anonymity: The disclosure attack," *Security & Privacy, IEEE*, vol. 1, no. 6, pp. 27–34, 2003.
[5] D. Kesdogan and L. Pimenidis, "The hitting set attack on anonymity protocols," in *Information Hiding*. Springer, 2005, pp. 326–339.
[6] N. Mathewson and R. Dingledine, "Practical traffic analysis: Extending and resisting statistical disclosure," in *Privacy Enhancing Technologies*. Springer, 2005, pp. 17–34.
[7] G. Danezis, C. Diaz, and C. Troncoso, "Two-sided statistical disclosure attack," in *Privacy Enhancing Technologies*. Springer, 2007, pp. 30–44.
[8] N. Mallesh and M. Wright, "The reverse statistical disclosure attack," in *Information Hiding*. Springer, 2010, pp. 221–234.
[9] C. Troncoso, B. Gierlichs, B. Preneel, and I. Verbauwhede, "Perfect matching disclosure attacks," in *Privacy Enhancing Technologies*. Springer, 2008, pp. 2–23.
[10] G. Danezis and C. Troncoso, "Vida: How to use bayesian inference to de-anonymize persistent communications," in *Privacy Enhancing Technologies*. Springer, 2009, pp. 56–72.