A Least Squares Approach to User Profiling in Pool Mix-based Anonymous Communication Systems

Fernando Pérez-González [1,2,3], Carmela Troncoso [4]

[1] *Signal Theory and Communications Dept., University of Vigo*
[2] *Gradiant (Galician R&D Center in Advanced Telecommunications)*
[3] *Electrical and Computer Engineering Dept., University of New Mexico*
[4] *K.U. Leuven/IBBT, ESAT/SCD-COSIC*
[1] fperez@gts.uvigo.es, [4] carmela.troncoso@esat.kuleuven.be

***Abstract*—Deployed high-latency anonymous communication systems conceal communication patterns using pool mixes as building blocks. These mixes are known to be vulnerable to Disclosure Attacks that uncover persistent relationships between users. In this paper we study the performance of the Least Squares Disclosure Attack (LSDA), an approach to disclosure rooted in Maximum Likelihood parameter estimation that recovers user profiles with greater accuracy than previous work. We derive analytical expressions that characterize the profiling error of the LSDA with respect to the system parameters for a threshold binomial pool mix and validate them empirically. Moreover, we show that our approach is easily adaptable to attack diverse pool mixing strategies.**

I. Introduction

High-latency anonymous communication systems aim at obfuscating communication patterns that can be exploited to infer sensitive information about users even when the content of messages is kept confidential. A common building block for these systems are mixes, relaying routers that hide the correspondence between their inputs and outputs [1]. In particular, deployed systems [2], [3] make use of *pool mixes*. Each round these mixes collects a number of messages, change their appearance cryptographically, and place them on a pool from which they are probabilistically chosen to leave the mix. Otherwise, they stay in the pool and get mixed with messages arriving in subsequent rounds.

It has been demonstrated that an adversary observing a mix-based system long enough can uncover persistent communication patterns [4] by launching a disclosure attack [5], [6], [7], [8]. This attack finds a user likely set of friends by intersecting the recipient anonymity sets [9] of this user's messages. These attacks have mostly been evaluated against threshold mixes in which, as opposed to pool mixes, mixing occurs only between messages in a given round and hence are easy to model and analyze. To the best of our knowledge, only the Statistical Disclosure Attack has been extended to attack pool mixes [10]. This is because the probabilistic nature of these mixing algorithms hampers the identification of sender and receiver anonymity sets, and hence hinders the adaptation of graph-theory based attacks [5], [7], [8] to pool-mix scenarios.

In this paper we study the applicability of the Least Squares Disclosure Attack (LSDA) [11] to pool mixes. This attack models profiling as a Least Squares problem, and yields profiles that minimize the error between the actual number of output messages and a prediction based on the input messages. We first revisit the LSDA extension to threshold binomial pool mixes proposed in [12]. We confirm that the derived analytic results that describe evolution of the estimation error with the parameters of the system closely model reality, and that the approach outperforms previous work. We finally show that the Least Squares can easily accommodate various pool mix strategies.

The rest of the paper is organized as follows: in the next section we describe our system and adversarial models. We introduce the Least Squares approach to disclosure applied to threshold binomial pool mixes in Sect. III and we validate the equations that characterize the LSDA's error in Sect. IV, further demonstrating its efficacy against different pool mix strategies. Finally, we conclude in Sect. V.

II. Notation and system model

Throughout the text we will use capital letters to denote random variables and lowercase letters to denote realizations. Vectors will be represented using boldface characters; thus, $\mathbf{x} = [x_1, \cdots, x_N]^T$ denotes a realization of random vector $\mathbf{X} = [X_1, \cdots, X_N]^T$. Matrices will be represented by boldface capital characters; whether they contain random or specific values will be clear from the context.

1) System model: We study a system in which a population of N_{users} users, designated by an index $i \in \{1, \ldots, N_{\text{users}}\}$, communicate through an high-latency anonymous communication channel, that we model as a pool mix. In this mix messages are placed on a pool with messages from previous rounds upon arrival. Then they leave the mix with a certain probability, or otherwise they stay in the pool and are mixed with messages arriving in subsequent rounds. The appearance of messages is changed cryptographically by the mix to avoid bitwise linkability between inputs and outputs.

We model with the random variable X_i^r, respectively Y_j^r, the number of messages that the ith (jth) user sends (receives) in round r; and denote as x_i^r (y_j^r) the actual number of messages i (j) sends (receives) in that round. Let $\mathbf{x}^r$ and $\mathbf{y}^r$ denote column vectors that contain as elements the number of messages sent or received by all users in round r: $\mathbf{x}^r = [x_1^r, \cdots, x_{N_{\text{users}}}^r]^T$, and $\mathbf{y}^r = [y_1^r, \cdots, y_{N_{\text{users}}}^r]^T$.

Users pick their messages recipients according to their sender profile $\mathbf{q}_i \doteq [p_{1,i}, p_{2,i}, \cdots, p_{N_{\text{users}},i}]^T$; being $p_{j,i}$ the

probability that user i chooses user j as receiver of a message. We consider that users have f friends to whom they send with probability $p_{j,i}$ ($p_{j,i} = 0$ when i is not a friend of j). We call *unnormalized receiver profile* of user j the column vector $\mathbf{p}_j \doteq [p_{j,1}, p_{j,2}, \cdots, p_{j,N_{\text{users}}}]^T$ containing the probabilities of the different senders choosing the jth user as receiver. This vector can be related to the receiver profile of user j through a simple normalization. We finally construct the vector $\mathbf{p}$ by stacking the unnormalized receiver profiles of all users, i.e., $\mathbf{p}^T \doteq [\mathbf{p}_1^T, \cdots, \mathbf{p}_{N_{\text{users}}}^T]$.

2) Adversary model: We consider a global passive adversary that knows all the parameters of the mix such as the mixing algorithm and the firing probability. As we focus on quantifying the impact of the information leaked by the mixing protocols on anonymity we assume that the cryptographic transformation performed during the mixing is perfect and thus the adversary cannot gain any information from studying the content of the messages.

The adversary monitors the system during ρ rounds observing the identity of the senders and receivers that communicate through the mix. Her goal is to uncover communication patterns from the observed flow of messages. Formally, given the observation $\mathbf{x}^r = \{x_i^r\}$ and $\mathbf{y}^r = \{y_j^r\}$, for $i, j = 1, \ldots, N_{\text{users}}$, and $r = 1, \ldots, \rho$, the adversary's goal is to obtain estimates $\hat{p}_{j,i}$ as close as possible to the probabilities $p_{j,i}$, which in turn allow for the recovery of the users' sending and receiver profiles. Fig. 1 illustrates the construction of the adversary's observation for a pool mix whose mixing algorithm we abstract, for the sake of simplicity, as simply firing messages in the pool with probability α. In round r, senders $i = 4, 8, 6$ send messages to receivers $j = 1, 2, 3$, chosen with probability $p_{j,i}$ according to their sender profiles. The messages are mixed in the pool with messages for receivers $j = 5, 7$ left from previous rounds. The mix chooses messages for receivers $j = 1, 5$ to be output; and messages for $j = 2, 3, 7$ stay in the pool until round $r+1$. Hence, in round r, the adversary's observation consists of $\mathbf{x}^r = [0, 0, 0, 1, 0, 1, 0, 1, 0, 0]^T$ and $\mathbf{y}^r = [1, 0, 0, 0, 1, 0, 0, 0, 0, 0]^T$. The same process is followed to construct $\mathbf{x}^{r+1}$, and $\mathbf{y}^{r+1}$.

We summarize the notation introduced in this section in Table II-2.

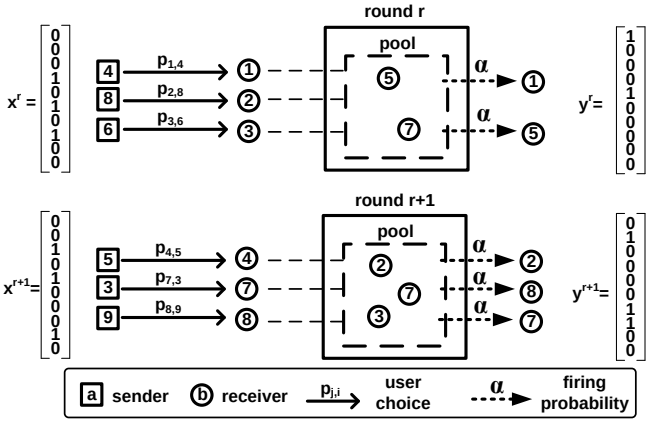

Fig. 1. System model.

TABLE I
SUMMARY OF NOTATION

Symbol	Meaning
N_{users}	Number of users, denoted by $i = \{1, \cdots, N_{\text{users}}\}$
f	Number of friends of each sender i
α	Firing probability of the pool mix
$p_{j,i}$	Probability that user i sends a message to user j
$\mathbf{q}_i$	User i's sender profile
$\mathbf{p}_j$	User j's unnormalized receiver profile
$\mathbf{p}$	Vector of unnormalized receiver profiles
ρ	Number of rounds observed by the adversary
x_i^r (y_j^r)	Number of messages i (j) sends (receives) in round r
$\mathbf{x}^r$ ($\mathbf{y}^r$)	Column vector containing elements x_i^r (y_j^r)
$\hat{p}_{j,i}$	Adversary's estimation of $p_{j,i}$
$\hat{\mathbf{q}}_i$	Adversary's estimation of $\mathbf{q}_i$
$\hat{\mathbf{p}}_j$	Adversary's estimation of $\mathbf{p}_j$
$\hat{\mathbf{p}}$	Adversary's estimation of $\mathbf{p}$

III. A Least Squares approach to Disclosure Attacks on Threshold Binomial Pool Mixes

In this section we discuss how to estimate the unnormalized receiver profiles from the observations $\mathbf{x}^r$, $\mathbf{y}^r$, $r = 1, \cdots, \rho$, focusing our analysis on a threshold binomial pool mix. This mix fires when it collects t messages (where t is called the threshold), having each message in the pool a probability α of being fired and $(1 - \alpha)$ of staying. We remark that once the unnormalized receiver profiles are known, the sender profiles can be straightforwardly obtained. To distinguish between the number of messages from the ith sender that enter and leave the mix in round r (note that some messages may stay in the pool) we will use vectors $\mathbf{x}^r$ and $\mathbf{X}_s^r$, respectively, where the vector $\mathbf{x}^r$ is observable while $\mathbf{X}_s^r$ is not. We let $\mathbf{U}^T \doteq [\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^\rho]$ and $\mathbf{U}_s^T \doteq [\mathbf{X}_s^1, \mathbf{X}_s^2, \cdots, \mathbf{X}_s^\rho]$. Since for any $i \in \{1, \cdots, N_{\text{users}}\}$, $r \in \{1, \cdots, \rho\}$, the sample $X_{s,i}^r$ is not observable, we construct its minimum mean-square predictor $\hat{x}_{s,i}^r$ given the whole set of input observations $\mathbf{U}$ as

$$\hat{x}_{s,i}^r = \alpha \sum_{k=0}^{r-1} (1-\alpha)^k x_i^{r-k} + \alpha(1-\alpha)^{r-1} m / N_{\text{users}} \qquad (1)$$

where we have assumed that at the time the adversary starts observing the system the pool contains m messages whose sender is unknown, i.e., they may correspond to any of the N_{users} users with uniform probability. For implementation purposes, a more convenient way of writing (1) is the following recursive equation

$$\hat{x}_{s,i}^{r+1} = (1-\alpha)\hat{x}_{s,i}^r + \alpha x_i^{r+1}, \quad r = 1, \cdots, N_{\text{users}} \qquad (2)$$

where $\hat{x}_{s,i}^1$ is initialized to $x_i^1 + m/N_{\text{users}}$. We will stack the predicted values for round r into vector $\hat{\mathbf{x}}_s^r \doteq [\hat{x}_{s,1}^r, \cdots, \hat{x}_{s,N_{\text{users}}}^r]^T$, and define $\hat{\mathbf{U}}_s^T \doteq [\hat{\mathbf{x}}_s^1, \hat{\mathbf{x}}_s^2, \cdots, \hat{\mathbf{x}}_s^\rho]$.

In [11] we show that a Maximum Likelihood formulation leads, after some simplifications, to the following estimate for vector $\mathbf{p}^T \doteq [\mathbf{p}_1^T, \cdots, \mathbf{p}_{N_{\text{users}}}^T]$:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p} \in \mathcal{P}} ||\mathbf{y} - \mathbf{H}\mathbf{p}||^2, \qquad (3)$$

where $\mathcal{P}$ denotes the set of valid probability vectors,[1] and $\mathbf{H} \doteq \mathbf{I}_{N_{\text{users}}} \otimes \mathbf{U}_s$, with $\otimes$ the Kronecker product and $I_{N_{\text{users}}}$ the identity matrix of size N_{users}. Problem (3) is nothing but a constrained Least Squares (LS) one. However, due to the simplicity of the analysis and the fact that the estimator so obtained is asymptotically efficient (i.e., $\hat{\mathbf{p}} \to \mathbf{p}$, when $\rho \to \infty$) we focus next on the unconstrained problem, for which the LS estimate $\hat{\mathbf{p}}_j$ for the jth unnormalized receiver profile can be decoupled from those of the other users and written as

$$\hat{\mathbf{p}}_j = (\hat{\mathbf{U}}_s^T \hat{\mathbf{U}}_s)^{-1} \hat{\mathbf{U}}_s^T \mathbf{y}_j, \quad j = 1, \cdots, N_{\text{users}}\,.$$

The performance analysis of this estimator in the case of a pool mix is carried out in Appendix A, where it is shown that the MSE is

$$\begin{aligned} \text{MSE} \quad &\doteq \textstyle\sum_{j=1}^{N_{\text{users}}} \text{tr}\left(\text{E}[(\mathbf{p}_j - \hat{\mathbf{p}}_j)(\mathbf{p}_j - \hat{\mathbf{p}}_j)^T]\right) && (4)\\ &\approx \quad \tfrac{N_{\text{users}}(N_{\text{users}}-1+\alpha_q/t)}{\rho\alpha_q} - \tfrac{(N_{\text{users}}-1)/(2-\alpha)+1/t}{\rho} && (5)\end{aligned}$$

where $\alpha_q \doteq \alpha/(2-\alpha)$. This approximation is asymptotically tight as $\rho \to \infty$. Moreover, when $\alpha = 1$ we recover the results for the threshold mix in [12]. When N_{users} is large, (5) can be approximated as

$$\text{MSE} \approx \frac{N_{\text{users}}^2}{\rho\alpha_q} \tag{6}$$

Noticing that the threshold mix corresponds to $\alpha_q = 1$, we can conclude that the pool mix requires $(2-\alpha)/\alpha$ times more rounds for the adversary to achieve the same MSE. For instance, for $\alpha = 0.5$, three times more rounds are needed to achieve the same MSE as in the threshold mix. Since α_q monotonically increases with α, the difficulty of learning the profiles is always larger in the pool mix compared to the threshold mix. Of course, this comes at the price of an increased delay; the binomial nature of the mixing process implies that the number of rounds messages stay in the pool follow a geometric distribution with parameter α, and hence the mean delay (measured in rounds) is $(1-\alpha)/\alpha$.

IV. Evaluation

We evaluate the effectiveness of the LSDA approach against synthetic anonymized traces created by a simulator written in the Python language.[2] comparing its results with those obtained performing the Statistical Disclosure Attack (SDA) [10], the only attack in the literature that has been applied to pool mixes. The SDA estimates a users sender profile by averaging the probability distributions describing the recipient anonymity set [9] of her messages. In a nutshell the sending probability $p_{j,i}$ is computed by counting in how many rounds user j has been seen receiving a message weighted by the probability that one (ore more) messages from user i leave the pool at that round, and averaging the result.

We simulate a population of N_{users} users with f contacts each, to whom they send messages with equal probability (i.e., $p_{j,i} = 1/f$ if i is friends with j, zero otherwise). For the sake of simplicity, we further fix that each receiver receives messages from the same number of senders. In the first part of the evaluation messages are anonymized using a threshold binomial pool mix where each round t messages arrive to the mix, and each message in the pool has a probability α of leaving the mix; and later on we study other mixing strategies. We consider that the adversary observes ρ rounds of mixing. The parameters' values used in our experiments, though rather unrealistic, have been chosen such that experiments could be carried out in reasonable time. We note, however, that the LSDA's results can be extrapolated to any set of parameters as long as the proportion among them is preserved.

We define the *Mean Squared Error per transition probability* (MSE_p) as the total MSE normalized by the number of elements of vector $\mathbf{p}$:

$$\text{MSE}_p = \text{MSE}/N_{\text{users}}^2.$$

The MSE_p expresses the accuracy of the attack by measuring the amount by which the output values $\hat{\mathbf{p}}$ output differ from the actual value $\mathbf{p}$ to be estimated. The smaller the MSE_p, the better is the adversary's estimation of the users' actual profiles. For each of the studied set of parameters we store the sets of senders and receivers during ρ rounds and compute the MSE_p for both SDA and LSDA. We repeat this process 20 times and plot the average of the results in our figures.

A. Results

We first evaluate the LSDA profiling performance, and the accuracy of our error predictor, when messages are anonymized using a threshold binomial pool mix. We recall that in such mix arriving messages are stored on a pool, and each round (i.e., when t messages are received) leave the mix with probability α. Otherwise, messages stay on the pool until the next round, when they are mixed with the arriving fresh messages and again probabilistically selected to be fired or not.

1) Performance with respect to delay: We recall that the mean delay in rounds is $(1-\alpha)/\alpha$. Fig. 2, top, illustrates the evolution of the LSDA's error when α varies. We see that the empirical error (represented by • in the figure) closely follows the prediction given by (6). As expected, large delays (i.e., small values of α) increase the error. The longer the delay, the more messages participate in the mixing, hindering the estimation of the sending probabilities.

Surprisingly, it seems that the SDA's MSE_p is independent from the pool mix firing probability α, and that moreover it outperforms the LSDA when the delay is large. A closer look at the estimated profiles reveals that actually the SDA's output resembles noise with mean $1/N_{\text{users}}$ (see Fig. 3). Only when the delay is minimal, i.e., when $\alpha \to 1$, and hence messages are rarely mixed with messages from other rounds, actual friends are assigned the largest probabilities in the estimated profile. When the firing probability is set to $\alpha = 0.9$ ($\rho = 10\,000$) the LSDA and SDA perform similarly, correctly

[1] Without further constraints, that may be furnished when there is partial knowledge about the transition probabilities, $\mathcal{P}$ is simply given by the constraints $0 \le p_{j,i} \le 1$ for all j, i, and $\sum_{j=1}^{N_{\text{users}}} p_{j,i} = 1$, for all i.

[2] The code will be made available upon request.

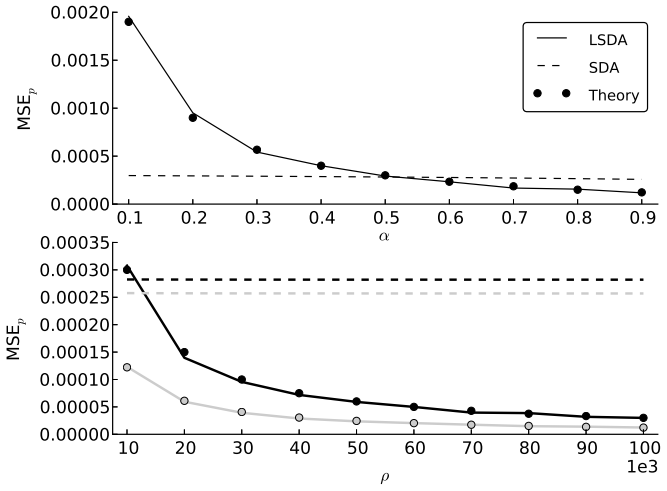

Fig. 2. MSE$_p$ evolution with α ($N = 100$, $f = 25$, $t = 10$, $\rho = 10\,000$) (top); and with ρ ($N = 100$, $f = 25$, $t = 10$) (bottom), where black lines (—) indicate $\alpha = 0.5$, and grey lines (—) indicate $\alpha = 0.9$.

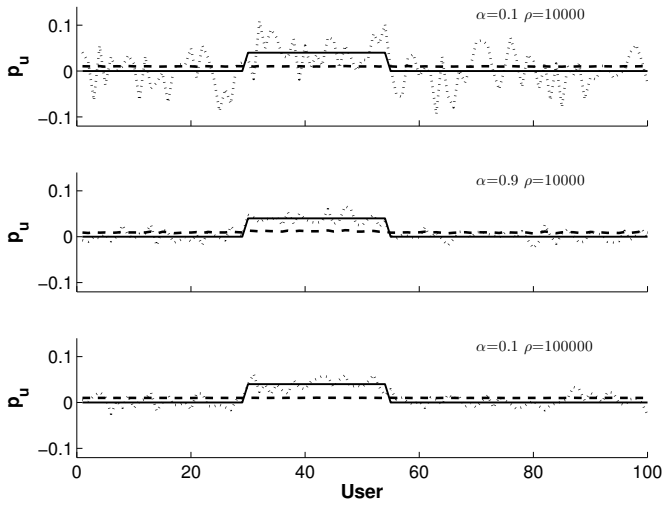

Fig. 3. Real (—), LSDA ($\cdots$), and SDA (- -) profiles for a given user depending on α and ρ.

identifying 93% and 91% of the users' friends, respectively. However, when $\alpha = 0.1$ the LSDA correctly identifies 47% of the users' friends, while the SDA only uncovers 37%.

2) Performance with respect to the number of rounds ρ: Fig. 2, bottom, shows the evolution of the MSE$_p$ with the number of rounds for two firing probabilities: $\alpha = 0.5$, and $\alpha = 0.9$ (black and grey lines, respectively). As predicted by (6), observing more rounds decreases significantly the LSDA's error. Nevertheless, the SDA's naive approach takes little advantage of the information procured by additional observations, and its output error remains virtually constant as the number of rounds grows. Looking at the number of correctly identified friends we find that for $\alpha = 0.1$ the LSDA correctly identifies 47% of the users' friends when 10 000 rounds are observed, and 75% when ρ is increased to 100 000. The SDA's performance in identifying friends, however, only improves from 37% to 55%.

3) Comparison between pool mixing strategies: So far we have tested the performance of the LSDA against threshold binomial pool mixes (denoted as "BIN"), that receive a constant number of messages, and fire a variable fraction of the messages in the pool. The attack is, however, not limited to such strategy. In this section we adapt the attack to other pool mixing strategies in order to illustrate its flexibility.

First, we consider a timed binomial pool mix (denoted as "TIME") that fires at regular intervals instead of upon the reception of t messages. Hence, not only the number of messages fired every round is variable, but also the number of arrivals. This can be easily encoded in the LSDA's formulation, as the only difference with respect to the attack described in Sect. III is that the observation vectors $\mathbf{x}^r$ do not necessarily sum up to t but may add up to a variable number (depending on the number of arrivals in a given round).

We also study a threshold pool mix in which a fixed fraction α of messages from the pool are fired every round (denoted as "%"). Hence both the number of messages arriving and leaving the mix in every round is constant (and equal to t). Adapting the LSDA to this pool mix requires us to compute the probability that a given message leaves the mix in a round. Let us denote as m the number of messages in the pool at the beginning of round r from which S_i^r belong to sender i. At the time of firing (when t new messages have entered the mix), the number of messages from the ith sender leaving the mix follow a hypergeometric distribution: t messages are selected from a total number of $m + t$ (out of which $S_i^r + x_i^r$ belong to sender i). Hence the average number of messages from the ith sender leaving the mix is

$$E\left\{\frac{t(x_i^r + S_i^r)}{(m+t)}\right\} = \frac{tx_i^r}{(m+t)} + \frac{tE\{S_i^r\}}{(m+t)} \doteq w_i^r\,. \qquad (7)$$

The first summand in the right hand side can be directly computed by the adversary; the second summand can be obtained recursively: the average number of messages from sender i remaining in the pool for round $r+1$ is $E\{S_i^{r+1}\} = E\{S_i^r\} + x_i^r - w_i^r$, that is, the average number of initial messages minus the average number of those leaving the pool. In fact, it can be seen that the attack on the "%" mix can be implemented by putting $t/(m+t)$ in place of α in Eq. (1).

We show in Fig. 4, box plots representing the distribution of the MSE$_p$ for the different mixing strategies depending on α (top) and on ρ (bottom). The parameters used in our experiments ensure that the mean number of arrivals, as well as the mean delay suffered by messages, is the same for the three mixes. The bottom figure reinforces our claim that the number of rounds has a dominant role on the profiling error, that decreases with the number of observations.

The top figure better illustrates the differences between the mixes. When α is small (i.e., there is good mixing among messages from subsequent rounds) the timed pool mix outperforms the other approaches. We conjecture that the variability at the entry and exit of the timed pool mix increase the uncertainty

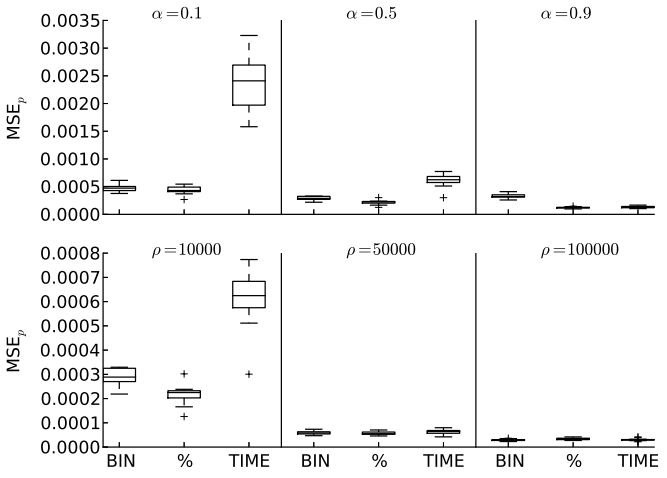

Fig. 4. MSE_p evolution for different mixing strategies with α ($N = 100$, $f = 25$, $t = 10$, $\rho = 10\,000$) (top); and with ρ ($N = 100$, $f = 25$, $t = 10$, $\alpha = 0.5$) (bottom).

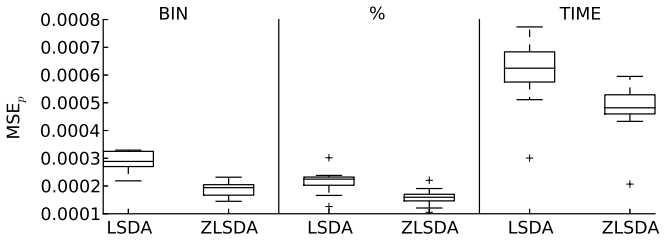

Fig. 5. LSDA vs ZLSDA MSE_p for different mixing strategies ($N = 100$, $f = 25$, $t = 10$, $\alpha = 0.5$ $\rho = 10\,000$).

of the adversary about the correspondence between input and output messages, increasing her error. The effect is opposite when α grows and messages often enter and leave the mix in the same round resembling the behavior of a threshold mix. In this case the variability would benefit the adversary that is able to infer a lot of information from rounds where few messages enter the mix. The LSDA obtains similar results against the binomial and percentage mixes, being slightly more effective when attacking the latter as there is less variation at the output of the mix than in the former.

4) Constrained profile estimation: The solution above approaches user profiling as an unconstrained problem which causes some of the estimated probabilities $\hat{p}_{j,i}$ to be negative (see Fig. 3). One can reduce the error by just setting those probabilities to zero, as we see in Fig. 5 where ZLSDA denotes this attack. Nevertheless, zeroing negative probabilities in such a straightforward manner disregards that profiles are well-defined probability distributions and hence $\sum_j p_{j,i} = 1$. The error can be further reduced by establishing constraints on Eq. (3) to ensure that the profiles recovered by the LSDA are well-defined. Such a solution can be found in [12].

V. Conclusion

We have studied the applicability of the Least Squares Disclosure Attack [11], [12] to pool mixes. Our empirical evaluation confirms that our formulas describing the error of the attack against a threshold binomial pool mix closely model reality. Furthermore, the LSDA outperforms the Statistical Disclosure Attack [10], the only attack in the literature adapted to these mixes. We have shown that the Least Squares approach is not limited to the analysis of these particular mixes, but can be adapted to account for other mixing strategies. The derivation of analytical results for mixing strategies other than the threshold binomial pool mix is left as subject for future work.

Appendix

Derivation of MSE for the pool mix. We will use the following result, which is proven in [11]. Let $\mathbf{M}$ be an arbitrary matrix of size $\rho \times \rho$, and $\mathbf{U}^T \doteq [\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^\rho]$, where the vectors X^r, $r = 1, \cdots N_{\text{users}}$, are independent and each follows a multinomial distribution with t trials and uniform probabilities $p_x \doteq 1/N_{\text{users}}$. Then, for large ρ

$$\begin{aligned}\mathbf{U}^T\mathbf{M}\mathbf{U} &\approx \text{tr}(\mathbf{M})tp_x\mathbf{I}_{N_{\text{users}}} \\ &+ \left(\text{sum}(\mathbf{M})t^2p_x^2 - \text{tr}(\mathbf{M})tp_x^2\right)\mathbf{1}_{N_{\text{users}}\times N_{\text{users}}} \quad (8)\end{aligned}$$

where $\text{tr}(\mathbf{M})$ stands for the trace of $\mathbf{M}$ and $\text{sum}(\mathbf{M})$ is the summation of all the elements of $\mathbf{M}$.

For compactness in the subsequent derivations, we will find it useful to define the following *convolution matrix*

$$\mathbf{B} \doteq \begin{bmatrix} \alpha & 0 & 0 & \cdots & 0 \\ \alpha(1-\alpha) & \alpha & 0 & \cdots & 0 \\ \alpha(1-\alpha)^2 & \alpha(1-\alpha) & \alpha & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \alpha(1-\alpha)^{\rho-1} & \alpha(1-\alpha)^{\rho-2} & \alpha(1-\alpha)^{\rho-3} & \cdots & \alpha \end{bmatrix} \quad (9)$$

Then, the predictor in (1) can be also written as

$$\hat{\mathbf{U}}_s = \mathbf{B}(\mathbf{U} + \mathbf{N}_0) \quad (10)$$

where the matrix $\mathbf{N}_0$, which accounts for the average initial state of the mix, is such that all entries in the first row take the value m/N_{users}, while all the remaining elements are zero. Notice that for the standard threshold mix, which corresponds to $\alpha = 1$, $m = 0$, we have that $\mathbf{B} = \mathbf{I}_\rho$, $\mathbf{N}_0 = \mathbf{0}$.

We can start now to derive an expression for the MSE. From (4) we can write the MSE as the sum of the traces of the covariance matrix for the estimated profile error corresponding to each user. Thus, we focus on such covariance matrix, which for the jth user becomes

$$\text{E}[(\mathbf{p}_j - \hat{\mathbf{p}}_j)(\mathbf{p}_j - \hat{\mathbf{p}_j})^T] = (\hat{\mathbf{R}}_{xs})^{-1}\hat{\mathbf{U}}_s^T\mathbf{\Sigma}_{y_j}\hat{\mathbf{U}}_s(\hat{\mathbf{R}}_{xs})^{-1} \quad (11)$$

where

$$\hat{\mathbf{R}}_{xs} \doteq \hat{\mathbf{U}}_s^T\hat{\mathbf{U}}_s = (\mathbf{U}^T + \mathbf{N}_0^T)\mathbf{B}^T\mathbf{B}(\mathbf{U} + \mathbf{N}_0) \quad (12)$$

is an estimate of the correlation matrix of $\hat{\mathbf{x}}_s^r$.

First, we compute the covariance matrix $\mathbf{\Sigma}_{y_j}$ of $\mathbf{Y}_j$, whose entries are $\text{Cov}\{Y_j^r, Y_j^l\}$, for all $r, l = 1, \cdots, \rho$. Under the assumption that each sender and receiver have exactly f friends, we can see Y_j^r as the sum of r independent binomial processes

with t trials and probabilities $p_x\alpha(1-\alpha)^k$, $k=0,\cdots,r-1$. Then,

$$\begin{aligned}\mathrm{Cov}\{Y_j^r, Y_j^l\} &= -tp_x^2\alpha^2\sum_{m=0}^{r-1}\sum_{k=0}^{l-1}(1-\alpha)^m(1-\alpha)^k, \quad r\neq l\\ \mathrm{Var}\{Y_j^r\} &= tp_x\alpha\sum_{m=0}^{r-1}(1-\alpha)^m - tp_x^2\alpha^2\sum_{m=0}^{r-1}(1-\alpha)^{2m}\end{aligned}$$

Thus, for large ρ we can write

$$\mathbf{\Sigma}_{y_j} \approx tp_x\mathbf{I}_\rho - tp_x^2\mathbf{B}\mathbf{B}^T \tag{13}$$

We also need to write the estimated correlation matrix $\hat{\mathbf{R}}_{xs}$ in a way that does not depend on the particular input. We will assume that $\mathbf{N}_0=\mathbf{0}$, as the impact of the initial conditions can be neglected for a large number of rounds. Therefore, $\hat{\mathbf{R}}_{xs}=\mathbf{U}^T\mathbf{B}^T\mathbf{B}\mathbf{U}$, so from the result at the beginning of the Appendix we need to obtain $\mathrm{tr}(\mathbf{B}^T\mathbf{B})$ and $\mathrm{sum}(\mathbf{B}^T\mathbf{B})$. For large ρ and neglecting border effects we have

$$\mathrm{tr}(\mathbf{B}^T\mathbf{B}) \approx \rho(b_k * b_{-k})|_{k=0}; \quad \mathrm{sum}(\mathbf{B}^T\mathbf{B}) \approx \rho\sum_{k=-\infty}^{\infty} b_k * b_{-k} \tag{14}$$

where $*$ denotes convolution, and $b_k \doteq \alpha(1-\alpha)^k u_k$, with u_k the unit-step function. From the definition, we find that

$$b_k * b_{-k} = \frac{\alpha}{2-\alpha}(1-\alpha)^{|k|} \tag{15}$$

from which it follows that $\mathrm{tr}(\mathbf{B}^T\mathbf{B}) = \rho\alpha/(2-\alpha) \doteq \rho\alpha_q$ and $\mathrm{sum}(\mathbf{B}^T\mathbf{B})=\rho$. Then, we can write $\hat{\mathbf{R}}_{xs} = d_{xs}\mathbf{I}_{N_{\text{users}}} + c_{xs}\mathbf{1}_{N_{\text{users}}\times N_{\text{users}}}$, where

$$c_{xs} \doteq \rho t p_x^2(t-\alpha_q); \quad d_{xs} \doteq \rho\alpha_q t p_x \tag{16}$$

From the structure of $\hat{\mathbf{R}}_{xs}$ it is possible to write

$$\hat{\mathbf{R}}_{xs}^{-1} = d_{xs}^{-1}\left(\mathbf{I}_{N_{\text{users}}} - \theta\mathbf{1}_{N_{\text{users}}\times N_{\text{users}}}\right) \tag{17}$$

$$\hat{\mathbf{R}}_{xs}^{-2} = d_{xs}^{-2}\left(\mathbf{I}_{N_{\text{users}}} - (2\theta - N_{\text{users}}\theta^2)\mathbf{1}_{N_{\text{users}}\times N_{\text{users}}}\right) \tag{18}$$

where $\theta \doteq p_x(1-\alpha_q/t)$.

With the previous derivations, the trace of (11) can be expanded as follows

$$\begin{aligned}\mathrm{tr}(\hat{\mathbf{R}}_{xs}^{-1}\mathbf{U}^T\mathbf{B}^T\mathbf{\Sigma}_{y_j}\mathbf{B}\mathbf{U}\hat{\mathbf{R}}_{xs}^{-1}) &= tp_x\mathrm{tr}(\hat{\mathbf{R}}_{xs}^{-1})\\ &- tp_x^2\mathrm{tr}(\mathbf{U}^T(\mathbf{B}^T\mathbf{B})^2\mathbf{U}\hat{\mathbf{R}}_{xs}^{-2})\end{aligned}$$

The first summand can be obtained from (17) since

$$\mathrm{tr}(\hat{\mathbf{R}}_{xs}^{-1}) = N_{\text{users}}d_{xs}^{-1} - N_{\text{users}}\theta = \frac{N_{\text{users}}}{\rho t\alpha_q}\left(N_{\text{users}} - 1 + \alpha_q/t\right) \tag{19}$$

For the second summand we use (18) together with the result at the beginning of this appendix to show

$$\begin{aligned}&\mathbf{U}^T(\mathbf{B}^T\mathbf{B})^2\mathbf{U}\hat{\mathbf{R}}_{xs}^{-2} = d_{xs}^{-2}\mathrm{tr}((\mathbf{B}^T\mathbf{B})^2)tp_x\mathbf{I}_{N_{\text{users}}} + d_{xs}^{-2}tp_x^2\\ &\cdot\ \left(\mathrm{sum}((\mathbf{B}^T\mathbf{B})^2t(1-N_{\text{users}}\theta)^2 - \mathrm{tr}((\mathbf{B}^T\mathbf{B})^2)\right)\mathbf{1}_{N_{\text{users}}\times N_{\text{users}}}\end{aligned}$$

Following the same reasoning as above, for large ρ we can show that

$$\mathrm{tr}((\mathbf{B}^T\mathbf{B})^2) = \frac{\rho\alpha}{(2-\alpha)^3}; \quad \mathrm{sum}((\mathbf{B}^T\mathbf{B})^2) = \rho \tag{20}$$

Hence

$$\begin{aligned}\mathrm{tr}(\mathbf{U}^T(\mathbf{B}^T\mathbf{B})^2\mathbf{U}\hat{\mathbf{R}}_{xs}^{-2}) &\approx d_{xs}^{-2}N_{\text{users}}\rho tp_x\\ &\cdot\ \left(\frac{\alpha}{(2-\alpha)^3}(1-p_x) + \alpha_q^2p_x\right)\end{aligned}$$

Combining all the previous results we obtain (5).

Acknowledgements. Research supported by the European Regional Development Fund (ERDF); by the Galician Regional Government under projects Consolidation of Research Units 2010/85 and SCALLOPS (10PXIB322231PR); by the Spanish Government under project COMONSENS (CONSOLIDER-INGENIO 2010 CSD2008-00010); by the Iberdrola Foundation through the Prince of Asturias Endowed Chair in Information Science and Related Technologies; by the Concerted Research Action (GOA) Ambiorics 2005/11 of the Flemish Government; and by the IAP Programme P6/26 BCRYPT. C. Troncoso is a research assistant of the Flemish Fund for Scientific Research (FWO).

REFERENCES

[1] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[2] G. Danezis, R. Dingledine, and N. Mathewson, "Mixminion: Design of a Type III Anonymous Remailer Protocol," in *IEEE Symposium on Security and Privacy*, pp. 2–15, IEEE Computer Society, 2003.

[3] U. Möller, L. Cottrell, P. Palfrader, and L. Sassaman, "Mixmaster Protocol — Version 2." IETF Internet Draft, July 2003.

[4] D. Kesdogan, D. Agrawal, and S. Penz, "Limits of anonymity in open environments," in *5th International Workshop on Information Hiding* (F. A. P. Petitcolas, ed.), vol. 2578 of *LNCS*, pp. 53–69, 2002.

[5] D. Agrawal and D. Kesdogan, "Measuring anonymity: The disclosure attack," *IEEE Security & Privacy*, vol. 1, no. 6, pp. 27–34, 2003.

[6] G. Danezis, "Statistical disclosure attacks: Traffic confirmation in open environments," in *Proceedings of Security and Privacy in the Age of Uncertainty* (Gritzalis, Vimercati, Samarati, and Katsikas, eds.), (Athens), pp. 421–426, IFIP TC11, Kluwer, May 2003.

[7] G. Danezis and C. Troncoso, "Vida: How to use Bayesian inference to de-anonymize persistent communications," in *9th Privacy Enhancing Technologies Symposium* (I. Goldberg and M. J. Atallah, eds.), vol. 5672 of *LNCS*, pp. 56–72, Springer, 2009.

[8] C. Troncoso, B. Gierlichs, B. Preneel, and I. Verbauwhede, "Perfect matching disclosure attacks," in *8th Privacy Enhancing Technologies Symposium* (N. Borisov and I. Goldberg, eds.), vol. 5134 of *LNCS*, pp. 2–23, Springer-Verlag, 2008.

[9] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *2nd International Workshop on Privacy Enhancing Technologies* (R. Dingledine and P. Syverson, eds.), vol. 2482 of *LNCS*, pp. 41–53, Springer, 2002.

[10] G. Danezis and A. Serjantov, "Statistical disclosure or intersection attacks on anonymity systems," in *6th International Workshop on Information Hiding* (J. J. Fridrich, ed.), vol. 3200 of *LNCS*, pp. 293–308, Springer, 2004.

[11] F. Pérez-González and C. Troncoso, "Understanding Statistical Disclosure: A Least Squares approach," in *12th Privacy Enhancing Technologies Symposium* (M. Wright and S. Fischer-Hubner, eds.), vol. 7384 of *LNCS*, pp. 38–57, Springer-Verlag, 2012.

[12] F. Pérez-González and C. Troncoso, "A least squares approach to the traffic analysis of high-latency anonymous communication systems," *IEEE Transactions on Information Forensics and Security*, 2012. Under Submission.