Understanding the Effects of Real-World Behavior in Statistical Disclosure Attacks

Simon Oya*[1], Carmela Troncoso†[2] and Fernando Pérez-González*†[3]
*Signal Theory and Communications Dept., University of Vigo
Vigo 36310, Spain
[1]simonoya@gts.uvigo.es [3]fperez@gts.uvigo.es
†Gradiant (Galician R&D Center in Advanced Telecommunications)
Vigo 36310, Spain
[2]ctroncoso@gradiant.org

***Abstract*—High-latency anonymous communication systems prevent passive eavesdroppers from inferring communicating partners with certainty. However, disclosure attacks allow an adversary to recover users' behavioral profiles when communications are persistent. Understanding how the system parameters affect the privacy of the users against such attacks is crucial. Earlier work in the area analyzes the performance of disclosure attacks in controlled scenarios, where a certain model about the users' behavior is assumed. In this paper, we analyze the profiling accuracy of one of the most efficient disclosure attack, the least squares disclosure attack, in realistic scenarios. We generate real traffic observations from datasets of different nature and find that the models considered in previous work do not fit this realistic behavior. We relax previous hypotheses on the behavior of the users and extend previous performance analyses, validating our results with real data and providing new insights into the parameters that affect the protection of the users in the real world.**

***Index Terms*—anonymity, mixes, performance analysis**

I. Introduction

Mixes aim at providing anonymity in communication networks by acting as routers that hide the correspondence between senders and receivers of messages. These anonymous communication channels operate by gathering the messages they receive, changing their appearance cryptographically and outputting them in batches, in what are called *rounds* of mixing. However, providing perfect anonymity through mixes is not possible in practice, due to constraints in the bandwidth of the communication channel and the delay tolerated by users. Because of this, an adversary observing the system in the long-term may infer the frequency with which a certain sender communicates with a certain receiver by means of a *disclosure attack* [1], [2], [3], [4], [5]. One of these strategies, called the Least Squares Disclosure Attack (LSDA) [5], [6], [7], has been proven to outperform previous statistical variants [8] while keeping its computational cost much lower than more sophisticated approaches, such as [4]. One advantage of LSDA is that it is particularly suitable for analysis, due to the availability of closed-form expressions for its prediction error in terms of the system parameters. Such performance analysis is of paramount importance since it helps the designer of mix-based anonymous communication systems to understand how to improve the protection of the users.

Previous works analyze the prediction error of LSDA in mix-based systems [5], [6], [7], [8] under specific assumptions on the users' and mix behavior. However, these results have only been confirmed by computer-generated observations and therefore it is not clear whether they apply in real-world scenarios. In this document, we delve into how users behave in reality. We gather data from real databases of different nature, which we then use to show that previous analyses of the attack fall short when tested against real data. We analyze the hypotheses that are needed for the performance analysis of LSDA to be applicable in real-world scenarios and develop a new generalized closed-form expression for the attacker's error when estimating the relationships between users in mixes, which we then evaluate with real traffic. Real-world datasets have been used in other works to compare between different disclosure attacks [9] or to analyze the properties of real traffic [10]. Our approach is different, as we are interested in understanding the effects of real-world user behavior on the performance of the least squares disclosure attack.

The document is structured as follows: we describe the least squares attack in the following section, together with the system model and notation we use in the paper. In Sect. III, we study the statistical properties of real-world behavior in our system. We carry out and evaluate a new performance analysis of LSDA in Sect. IV, and conclude in Sect. V.

II. The Least Squares Disclosure Attack

The Least Squares Disclosure Attack (LSDA), introduced by Pérez-González and Troncoso in [5], estimates the intensity of the communication between each sender-receiver pair in a mix-based anonymous channel by solving a least squares problem. This intensity is represented by the *transition probabilities* $p_{j,i}$, which model the *average probability* that a message sent by sender $i \in \{1, 2, \cdots, N\}$ is addressed to receiver $j \in \{1, 2, \cdots, M\}$. These probabilities are commonly grouped per sender in the so-called *sending profiles*, $\mathbf{q}_i \doteq [p_{1,i}, \cdots, p_{M,i}]^T$. An attacker that observes the number of messages sent and received during ρ communication rounds obtains the LSDA estimator by solving

$$\hat{\mathbf{P}} = \left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{Y}\,, \qquad (1)$$

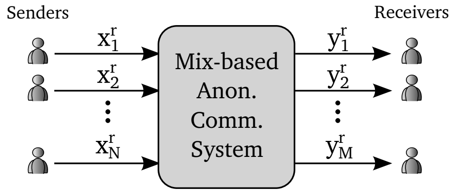


Fig. 1: System model during the communication *round* r.

where $\hat{\mathbf{P}}$ is a $N \times M$ matrix containing the estimation of the transition probability $\hat{p}_{j,i}$ in its i,jth entry, $\mathbf{U}$ is a $\rho \times N$ matrix containing the amount of messages sent by sender i in round r, denoted x_i^r, in its r,ith entry, and $\mathbf{Y}$ is a $\rho \times M$ matrix with the number of messages received by receiver j in round r, denoted y_j^r, in its r,jth entry. Figure 1 shows an example of the system and notation employed. The estimator in (1) was proven to be unbiased and asymptotically efficient, in the sense that its variance approaches zero as the length of the observation window ρ increases [5], [7], in mix-based systems where all messages leave the mix in each round.

Denoting the jth column of $\mathbf{P}$ by $\mathbf{p}_j \doteq [p_{j,1}, \cdots, p_{j,N}]^T$, and the jth column of $\mathbf{Y}$ by $\mathbf{y}_j \doteq [y_j^1, \cdots, y_j^\rho]$, (1) can be decoupled as

$$\hat{\mathbf{p}}_j = \left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{y}_j\,. \tag{2}$$

This latter formulation is specially useful to carry out a performance analysis of the attack.

III. Modeling Real-World Behavior

In this section, we study real-world user behavior from observations generated with real traffic, showing that previous performance analyses of LSDA are not valid in this scenario because the assumptions they are based on are rather unrealistic. We propose alternative hypotheses that are adequate to model real-world user behavior, which we then use in Sect. IV to assess the performance of the LSDA estimator.

A. Generating real-world observations

In order to analyze real-world behavior, we have chosen to generate observations by taking real traffic from datasets of different nature, whose users could have relied on mix-based systems to enhance their privacy, and anonymize this traffic using different mix configurations. We work with three datasets, whose basic information is summarized in Table I:

1) ***Email*:** This dataset contains around 220 000 emails sent from 294 different email addresses, which have been extracted from the Enron corpus.[1] Messages with multiple recipients are treated as different messages sent simultaneously, one for each recipient.
2) ***Location*:** This dataset contains around 400 000 location check-ins taken from the 500 most active users of Gowalla social networking website.[2] Users checking-in are considered as the senders, while the locations form the set of receivers. We consider only the 500 most active users for computational reasons: LSDA works with large-size matrices which grow with the number of senders and receivers of the system.

TABLE I: Basic information of the datasets.

Dataset	No. messages	Duration (hours)	Senders	Receivers
Email	220 032	32 416.8	294	17 017
Location	406 484	4 344.0	500	559
MailingList	178 937	76166.4	500	510

3) ***MailingList*:** we have processed the public mailing lists of Indimedia,[3] obtaining almost 180 000 messages from the 500 most active senders. Each mailing list is considered as a receiver, while users posting to these mailing lists are senders.

We anonymize the traces from these datasets using two types of mixes, which differ in the event that triggers the flushing of messages:

1) **Threshold mix:** this mix gathers messages until it has stored t of them, and then forwards each one to its correspondent recipient.
2) **Timed mix:** this mix stores the messages it receives and, after a period of time τ, outputs each one to their recipients.

To generate the adversary's observations, we choose values of t and τ that provide an acceptable degree of anonymity while keeping the delay of messages under a reasonable bound. We adopt the following criteria: in the threshold mix, we choose a value $t = 100$ and, in the timed mix, we select values of τ to ensure that ≈ 100 messages are mixed on average per round, while also considering that a delay of more that 24 hours is intolerable for users. This makes $\tau = 12$ hours for *Email*, $\tau = 1$ hour for *Location* and $\tau = 24$ hours for *MailingList*, with an average of ≈ 100 messages per round in the first two, and ≈ 57 in the latter. The result of this anonymization is a set of observations from $\{X_i^r\}$ and $\{Y_j^r\}$.

B. Modeling the input process

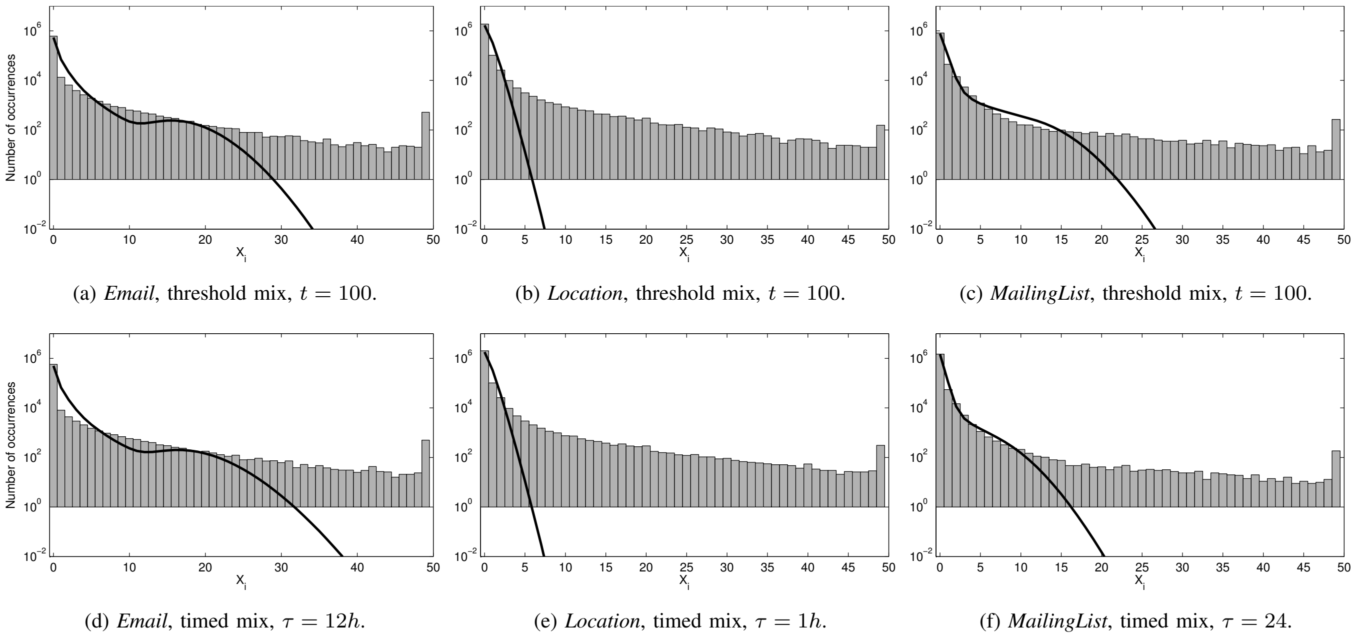
The input process, $\{X_i^r\}$, which models the amount of messages from each user arriving to the mix in each round, is determined by the frequency with which users send messages and by the firing condition of the mix. When the anonymization channel is a threshold mix, previous analyses [5], [6], [8] assume that the input process follows a *multinomial distribution*, and, when the channel is a timed mix, authors in [7] assume that the number of messages each user sends to the mix can be independently modeled as a *Poisson process*.

In Fig. 2, we compare the histogram of the inputs $\{X_i^r\}$, obtained using the observations generated with our datasets, with the theoretical values given by the multinomial and Poisson models (in the threshold and the timed mixes, respectively). Here, the last bin of the histogram contains all occurrences of $X_i^r \geq 50$. We conclude that the theoretical models fit the

[1] http://www.cs.cmu.edu/∼./enron/

[2] http://snap.stanford.edu/data/loc-gowalla.html

[3] http://lists.indymedia.org/



(a) *Email*, threshold mix, $t = 100$.

(b) *Location*, threshold mix, $t = 100$.

(c) *MailingList*, threshold mix, $t = 100$.

(d) *Email*, timed mix, $\tau = 12h$.

(e) *Location*, timed mix, $\tau = 1h$.

(f) *MailingList*, timed mix, $\tau = 24$.

Fig. 2: Histograms representing the amount of messages each user sends in each round, compared to the approximation given by the theoretical models assumed in previous works (line). The last bin contains all occurrences of $X_i^r \geq 50$.

histogram for low number of messages X_i, but fail at capturing the large values.

In the analysis in this document, we do not assume a specific distribution for $\{X_i^r\}$, but consider that it is a generic *stationary* process that satisfies the relation

$$\text{Cov}\left\{X_k, X_m\right\} \ll \text{Var}\left\{X_k\right\} \qquad \forall k, m \quad k \neq m \tag{3}$$

and, additionally,

$$\text{Cov}\left\{X_k, X_m X_n\right\} \ll \text{Cov}\left\{X_k^2, X_k\right\} \tag{4}$$

$$\text{Cov}\left\{X_k^2, X_m X_n\right\} \ll \text{Cov}\left\{X_k^2, X_k^2\right\} \tag{5}$$

for all k, m, n except when $k = m = n$. These assumptions mean, in other words, that the participation of a user in a given round is uncorrelated with the participation of each other user in that round. We have validated these hypotheses by computing the different sample covariances from our datasets, as shown in Table II.

C. Modeling the output process

A crucial point when carrying out a performance analysis of disclosure attacks on mixes is selecting a model for the distribution $\{Y_j^r | X_1^r, \cdots, X_N^r\}$, which represents how users choose the recipients of their messages in each round. A known property of this distribution, given by the definition of sending profiles, is that $\text{E}\left\{\mathbf{Y}|\mathbf{U}\right\} = \mathbf{U} \cdot \mathbf{P}$. However, this is true for many distributions. Every previous analysis of LSDA assumes that the choice of recipients is *stationary* and that $\{Y_j^r | X_1^r, \cdots, X_N^r\}$ follows a *multinomial model*, i.e.,

$$\{Y_1^r, \cdots, Y_M^r | \mathbf{U}\} \sim \sum_{k=1}^{N} \text{Multi}\left(x_k^r, \mathbf{q}_k\right) . \tag{6}$$

This model is adequate in scenarios where users choose the recipients of each of their messages in each round independently. However, when users tend to focus on a single receiver in each round, (6) is not suitable to model the output distribution.

In this work, we assume *two models* for $\{Y_j^r | X_1^r, \cdots, X_N^r\}$ that are examples of how users can distribute their messages among the receivers while satisfying $\text{E}\left\{\mathbf{Y}|\mathbf{U}\right\} = \mathbf{U} \cdot \mathbf{P}$:

1) A *multinomial model*, given by (6), as an example of users that cause low variance output.
2) A *maximum variance model*, given by

$$\{Y_1^r, \cdots, Y_M^r | \mathbf{U}\} \sim \sum_{k=1}^{N} x_k^r \cdot \text{Multi}\left(1, \mathbf{q}_k\right) . \tag{7}$$

When using these distributions, we are implicitly assuming that the choices of recipients of different senders within the same round are uncorrelated, and that the choice of recipients of the same user between rounds can be also considered uncorrelated. Our experiments in Sect. IV-2 confirm that the results we obtain with these approximations are accurate.

To illustrate how users' behavior changes between scenarios, we have computed the average number of recipients each sender chooses in each round of the observations generated with our datasets, as a function of the number of messages sent. This is displayed in Table III. As a reference, the average number of senders' contacts in each dataset is 125.7 in *Email*, 16 in *Location* and 9.6 in *MailingList*. These results show that users in the *Email* dataset tend to spread their messages among their contacts, behaving close to (6), while users in *Location* and *MailingList* focus on a single recipient in each round, as in (7).

TABLE II: Average values for different sample covariances of the input process in the datasets.

	Email		*Location*		*MailingList*	
	$t=100$	$\tau=12$	$t=100$	$\tau=1$	$t=100$	$\tau=24$
$\lvert\mathrm{Cov}\left\{X_k,X_k\right\}\rvert$	7.1	20.1	2.0	2.4	2.9	4.6
$\lvert\mathrm{Cov}\left\{X_k,X_m\right\}\rvert$	0.1	0.2	0.0	0.0	0.0	0.0
$\lvert\mathrm{Cov}\left\{X_k^2,X_k\right\}\rvert$	415.9	5815.0	57.2	159.9	171.4	2076.1
$\lvert\mathrm{Cov}\left\{X_kX_m,X_k\right\}\rvert$	1.0	15.1	0.3	0.5	0.2	0.5
$\lvert\mathrm{Cov}\left\{X_k^2,X_m\right\}\rvert$	1.9	12.8	0.3	0.5	0.5	0.7
$\lvert\mathrm{Cov}\left\{X_kX_m,X_n\right\}\rvert$	0.0	0.3	0.0	0.0	0.0	0.0
$\lvert\mathrm{Cov}\left\{X_k^2,X_k^2\right\}\rvert$	32943.2	2974803.8	2412.7	39132.0	13666.8	1381640.2
$\lvert\mathrm{Cov}\left\{X_k^2,X_kX_m\right\}\rvert$	30.1	4038.3	6.6	27.2	7.0	189.6
$\lvert\mathrm{Cov}\left\{X_k^2,X_m^2\right\}\rvert$	51.2	786.2	4.5	10.4	8.9	24.8
$\lvert\mathrm{Cov}\left\{X_k^2,X_mX_n\right\}\rvert$	0.7	16.2	0.1	0.2	0.1	0.2

TABLE III: Average number of recipients chosen by the senders in each round, as a function of the number of messages sent.

# messages (X_i)		$=2$	$=3$	$=4$	$=5$	≥ 6
Email	$t=100$	1.85	2.71	3.53	4.40	13.56
	$\tau=12h$	1.85	2.69	3.49	4.32	14.02
Location	$t=100$	1.03	1.05	1.06	1.08	1.11
	$\tau=1h$	1.10	1.14	1.17	1.18	1.26
MailingList	$t=100$	1.29	1.46	1.53	1.53	1.57
	$\tau=24h$	1.28	1.49	1.56	1.55	1.71

IV. Extended Performance Analysis of the Least Squares Disclosure Attack

We now assess the profiling accuracy of the Least Squares Disclosure Attack with the assumptions in the input and output processes proposed in the previous section, which we have validated with traffic from real-world scenarios. The profiling accuracy is measured as the Mean Squared Error (MSE) between the attacker's estimation of the *sending profiles* of the users and their real values, i.e., $\mathrm{MSE}_i \doteq \sum_{j=1}^{N} |p_{j,i} - \hat{p}_{j,i}|^2$. This analysis generalizes previous ones [5], [6], [8], [7], accommodating different types of mixes and being able to model real-world behavior, at the expense of accuracy.

1) Theoretical approximation of the average MSE: Our goal is to obtain an approximation of the *average* MSE_i when using (1) to estimate the sending profiles, where this average is computed over all the realizations of $\mathbf{U}$ and $\mathbf{Y}$ obtained with users' average behavior $\mathbf{P}$. For simplicity, we omit the conditioning on $\mathbf{P}$ in the derivations below.

For the analysis in this section, we introduce additional notation regarding the statistics of the input and output processes. We use $\mu(i)$ to refer to the expected value of X_i, and $\mu_n(i)$ is its nth central moment. Vector $\boldsymbol{\mu}$ contains all $\mu(i)$ for each sender, i.e., $\boldsymbol{\mu} \doteq [\mu(1), \cdots, \mu(N)]^T$. Matrix $\mathbf{M}$ contains these values arranged in its main diagonal, i.e., $\mathbf{M} \doteq \mathrm{diag}\{\mu(1), \cdots, \mu(N)\}$ and, similarly, $\mathbf{M}_n \doteq \mathrm{diag}\{\mu_n(1), \cdots, \mu_n(N)\}$. We use the parameter $s_{j,i} \doteq p_{j,i}(1-p_{j,i})$, which is closely related to the variance of the outputs, and the diagonal matrix $\mathbf{S}_j \doteq \mathrm{diag}\{s_{j,1}, \cdots, s_{j,N}\}$. Finally, we define the *uniformity* of the sending profile of user i as $\upsilon_i \doteq 1 - \sum_{j=1}^{M} p_{j,i}^2$. The uniformity gives an idea of how random the behavior of a user is, and ranges from 0, when sender only has one contact, to $(M-1)/M$, when this user sends messages to all the receivers with the same probability during the observation period. Note that $\sum_{j=1}^{M} s_{j,i} = \upsilon_i$.

We start the derivations by showing that the LSDA estimator is unbiased. This is straightforward from the fact that, given a matrix of input messages $\mathbf{U}$ and the average behavior of the senders $\mathbf{P}$, the expected value of the output is

$$\mathrm{E}\left\{\mathbf{Y}|\mathbf{U}\right\} = \mathbf{U}\cdot\mathbf{P} \tag{8}$$

where $\mathrm{E}\left\{\cdot\right\}$ is taken along all the possible assignments of the messages in $\mathbf{U}$ to the receivers, following $\mathbf{P}$. Using (8) together with (1), we get $\mathrm{E}\{\hat{\mathbf{P}}\} = \mathbf{P}$ (alternatively, $\mathrm{E}\left\{\hat{\mathbf{p}}_j\right\} = \mathbf{p}_j$). This property allows to write, using the law of total variance,

$$\boldsymbol{\Sigma}_{\hat{\mathbf{p}}_j} = \mathrm{E}\left\{\boldsymbol{\Sigma}_{\hat{\mathbf{p}}_j|\mathbf{U}}\right\} = \mathrm{E}\left\{(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{Y}_j|\mathbf{U}}\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\right\} \tag{9}$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}_j|\mathbf{U}} \doteq \mathrm{E}\left\{(\mathbf{Y}_j - \mathrm{E}\left\{\mathbf{Y}_j|\mathbf{U}\right\})(\mathbf{Y}_j - \mathrm{E}\left\{\mathbf{Y}_j|\mathbf{U}\right\})^T|\mathbf{U}\right\}$.

Since we have assumed that the input process is stationary, using the Law of Large Numbers and considering that the number of rounds observed ρ is large enough, we approximate

$$\lim_{\rho\to\infty} \mathbf{U}^T\mathbf{U}/\rho \to \mathbf{R}_x \tag{10}$$

where $\mathbf{R}_x$ is the autocorrelation matrix of the input process, i.e., an $N\times N$ symmetric matrix whose m,nth element is $\mathrm{E}\left\{X_mX_n\right\}$. Using (3), we write this matrix as

$$\mathbf{R}_x \approx \boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{M}_2. \tag{11}$$

The inverse of (11) can be computed applying the Sherman-Morrison formula [11], which gives us

$$\mathbf{R}_x^{-1} \approx \mathbf{M}_2^{-1}\left(\mathbf{I}_N - \gamma\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{M}_2^{-1}\right) \tag{12}$$

where $\gamma \doteq 1/(1+\boldsymbol{\mu}^T\mathbf{M}_2^{-1}\boldsymbol{\mu})$. Therefore, when the number of rounds observed is large, (9) can be approximated as

$$\boldsymbol{\Sigma}_{\hat{\mathbf{p}}_j} \approx \frac{1}{\rho}\mathbf{R}_x^{-1}\mathbf{R}_{xyx}\mathbf{R}_x^{-1}. \tag{13}$$

where the middle term is $\mathbf{R}_{xyx} \doteq \frac{1}{\rho}\mathrm{E}\left\{\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{Y}_j|\mathbf{U}}\mathbf{U}\right\}$. In order to compute the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}_j|\mathbf{U}}$, we analyze separately the two scenarios (6) and (7) we consider for the distribution of the output process given the inputs.

a) Multinomial model: Using (6) together with our assumptions, we approximate the middle term of (13) as

$$\mathbf{R}_{xyx} \approx \left(\sum_{k=1} \mu(k) s_{j,k} \right) \cdot \left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{M}_2 \right) + \mathbf{M}_3 \mathbf{S}_j. \qquad (14)$$

Finally, plugging (12) and (14) into (13) and performing matrix multiplications we obtain $\boldsymbol{\Sigma}_{\hat{\mathbf{p}}_j}$. Then, taking the i-th diagonal element of this matrix, which is Var $\{\hat{p}_{j,i}\}$, adding this element along j, and further considering $\sum_{k=1,k\neq i}^{N} \mu^2(k)/\mu_2(k) \gg 0$ for all $i \in \{1, \cdots, N\}$, we obtain:

$$\text{MSE}_i^- \approx \frac{1}{\rho} \cdot \frac{1}{\mu_2(i)} \left(\sum_{k=1}^{N} \mu(k) \cdot \upsilon_k + \frac{\mu_3(i)}{\mu_2(i)} \cdot \upsilon_i \right) \qquad (15)$$

b) Maximum variance model: We now analyze the performance of the LSDA estimator when the output distribution is (7). In that case, the middle term of (13) becomes

$$\mathbf{R}_{xyx} \approx \left(\sum_{k=1}^{N} (\mu(k)^2 + \mu_2(k)) s_{j,k} \right) \cdot \left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{M}_2 \right) + \mathbf{M}_4 \mathbf{S}_j. \qquad (16)$$

Operating as explained before to obtain the MSE in the estimation of the sending profile of user i, we get

$$\text{MSE}_i^+ \approx \frac{1}{\rho} \cdot \frac{1}{\mu_2(i)} \left(\sum_{k=1}^{N} \left(\mu(k)^2 + \mu_2(k) \right) \cdot \upsilon_k + \frac{\mu_4(i)}{\mu_2(i)} \cdot \upsilon_i \right) \qquad (17)$$

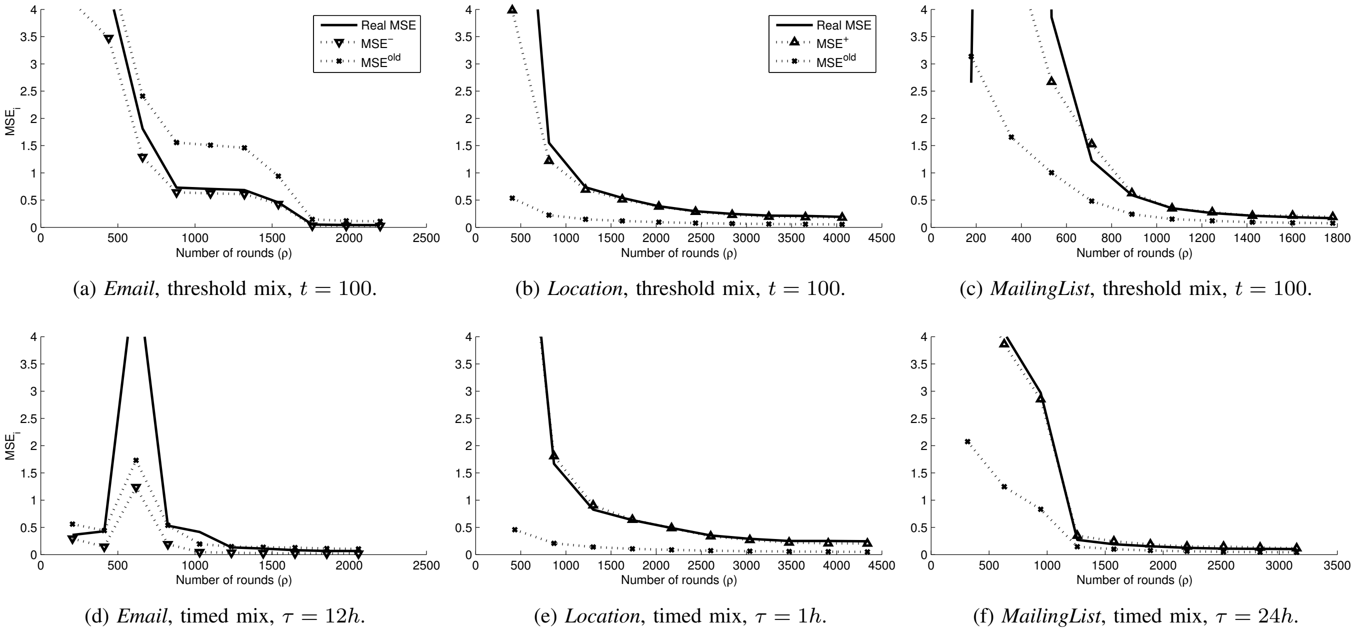
The formulas (15) and (17) provide new insights into how LSDA's error depends on the system parameters. This error decreases with ρ, since it becomes easier for the attacker to estimate the behavior of the users as more observations are available. The *variance of the input process X_i decreases* the estimation error of $\mathbf{q}_i$, i.e., it is easier to separate the sending behavior of a user from the others when we have rounds where that user participates a lot as well as rounds where that user is not present. The MSE$_i$ also *increases* with the contribution of all senders to the *output variance*, more strongly when users behave as in (7) than as in (6). The role of the uniformity of the profiles υ_k in the MSE is also very relevant: estimating the sending profiles is a much easier task when users only contact very few receivers (i.e., low υ_k) than when they distribute their messages among a larger population (i.e., υ_k close to 1).

2) Evaluation: We now evaluate our formulas, applying LSDA to the anonymized traces of real traffic. For each dataset, mix configuration, and number of rounds observed $\rho \in \{0.1\rho_{max}, 0.2\rho_{max}, \cdots, \rho_{max}\}$, where ρ_{max} is the total number of rounds in the observations, we perform LSDA and compute the real MSE$_i$. We then represent the average MSE$_i$ of those users i that meet three conditions: they are among the 40% most active users, they belong to the 40% users that remain active for the largest number of rounds, and furthermore they participate before $0.3\rho_{max}$ rounds have been observed. We do this to avoid sporadic peaks in the average MSE$_i$, which are the result of estimating the sending profile of a user that barely participates in the system, and to be able to see the trend of the MSE with clarity.

Figure 3 shows this average MSE$_i$ for the *Email*, *Location* and *MailingList* datasets, together with the theoretical formulas MSE$_i^-$ and MSE$_i^+$ in (15) and (17). We only plot the theoretical approximation that better suits each scenario: MSE$_i^-$ in the *Email* experiments and MSE$_i^+$ in the *Location* and *MailingList* experiments. We also plot the theoretical MSE from previous works, denoted by MSE$^{\texttt{old}}$, which has been taken from [8] and [7] for the threshold and the timed mix experiments, respectively. We set the limits of the vertical axis to the same value in all figures to ease the comparison between them. These limits make early values of the MSE (low ρ) to fall outside the plot, but allow to see with more detail the performance of the attack for large values of ρ. We do this on purpose: we are predicting the asymptotic MSE$_i$ of the attack, so the results for low values of ρ are not significant in our evaluation.

We see that our approximations improve those given by previous work, especially in those scenarios where the multinomial model for the choice of recipients is not appropriate (*Location* and *MailingList*). We note that the number of rounds we can generate with the *Email* database in Fig. 3d is not large enough to appreciate this improvement, due to the spike we observe in that experiment at early values of ρ. This sudden increase of the MSE, as well as the one in Fig. 3c, happens for two reasons: first, when the number of rounds observed is small, it is easier for the matrix $\mathbf{U}^T\mathbf{U}$ to be ill-conditioned, which results in a poor estimation of the sending profiles (cf. (1)), and therefore in a large MSE in the realization. This spike is not predicted by our theoretical formulas, since they approximate the *average* MSE. On the other hand, in the *Email* dataset, most of the users whose MSE$_i$ we average start sending messages when the adversary has observed around 30% of the total number of rounds. This causes an increase in the MSE$_i$ at around $\rho = 600$, as we are adding users to the average MSE$_i$ that have barely participated in the system. The average MSE$_i$ stabilizes as the number of rounds observed increases since the number of users used for the computation of the average MSE$_i$ we represent remains unchanged.

In all cases, the MSE$_i$ decreases as the number of observed rounds ρ increases, as predicted by our formulas, except for the spikes in Figs. 3d and 3c whose origin we have already explained. Due to these spikes, comparing the results of the experiments in the *Email* and *MailingList* datasets is not possible. However, we can see that the MSE in the experiments with *Location* is stable, and always larger in the threshold mix scenario (Fig. 3b). The reason for this is the following: the variance of the the input process in a threshold mix is smaller than that in a timed mix for the same average number of messages sent per round. This is the the case in the *Location* experiments, since the number of rounds we generate in the threshold and timed mix experiments is approximately the same. As predicted by our theoretical formulas, a system with lower input variance provides more protection against the LSDA attacker.



(a) *Email*, threshold mix, $t = 100$.

(b) *Location*, threshold mix, $t = 100$.

(c) *MailingList*, threshold mix, $t = 100$.

(d) *Email*, timed mix, $\tau = 12h$.

(e) *Location*, timed mix, $\tau = 1h$.

(f) *MailingList*, timed mix, $\tau = 24h$.

Fig. 3: Average MSE$_i$ evolution with ρ using the *Email*, *Location* and *MailingList* datasets, for different types of mixes.

V. Conclusions

We have analyzed the effects of real-world user behavior in the performance of the least squares disclosure attack [5] in mix-based anonymous communication systems, considering mixes that do not delay messages between communication rounds. To validate our work, we have obtained real traffic observations from three publicly available datasets of different nature: emails sent between the employees of a company, location check-ins from an online social network, and users' posts to mailing lists. By studying these data, we confirm that the hypotheses upon which former analyses of the least squares disclosure attack are based [5], [6], [7] are not adequate to model real-world behavior, and hence we formulate new ones. Based on these new assumptions, we develop a generalized performance analysis of the attack, which we validate with our datasets, confirming that it accurately models the estimation error of the attacker in the considered realistic scenarios. This analysis accommodates a wide variety of mix and users' behavior, and provides new insights into the statistics that affect the protection of the users: the variability in the participation of the users in the system contributes to the attacker's success, while the variability in the messages received by users worsens the attacker's estimation.

Acknowledgement

This work was partially supported by the Spanish Government and the European Regional Development Fund under project TACTICA, by the Galician Regional Government under project Consolidation of Research Units GRC2013/009, and by the EU 7th Framework Programme (FP7/2007-2013) under grant agreements 610613 (PRIPARE) and 285901 (LIFTGATE).

References

[1] 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., IFIP TC11. Athens: Kluwer, May 2003, pp. 421–426.

[2] N. Mathewson and R. Dingledine, "Practical traffic analysis: Extending and resisting statistical disclosure," in *4th Workshop on Privacy Enhancing Technologies*, ser. LNCS, D. Martin and A. Serjantov, Eds., vol. 3424. Springer, 2004, pp. 17–34.

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

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

[5] F. Pérez-González and C. Troncoso, "Understanding statistical disclosure: A least squares approach," in *Privacy Enhancing Technologies - 12th Symposium*, ser. LNCS, vol. 7384. Springer-Verlag, 2012, pp. 38–57.

[6] F. Pérez-González and C. Troncoso, "A least squares approach to user profiling in pool mix-based anonymous communication systems," in *IEEE Workshop on Information Forensics and Security*, 2012, pp. 115–120.

[7] S. Oya, C. Troncoso, and F. Pérez-González, "Do dummies pay off? limits of dummy traffic protection in anonymous communications," in *14th Symposium on Privacy Enhancing Technologies*, 2014.

[8] S. Oya, C. Troncoso, and F. Pérez-González, "Meet the family of statistical disclosure attacks," *IEEE Global Conference on Signal and Information Processing*, p. 4p, 2013.

[9] G. Danezis and C. Troncoso, "You cannot hide for long: De-anonymization of real-world dynamic behaviour," in *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, ser. WPES '13. ACM, 2013, pp. 49–60.

[10] K. Malinka, P. Hanáček, and D. Cvrček, "Analyses of real email traffic properties," *Radioengineering*, vol. 18, no. 4, p. 7, 2009.

[11] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.