Under consideration for publication in Math. Struct. in Comp. Science

Asymptotic Information Leakage under One-Try Attacks[†]

Michele Boreale[1], Francesca Pampaloni[2], Michela Paolini[2]

[1] *Università di Firenze – Dipartimento di Sistemi ed Informatica, Viale Morgagni 65, 50134 Firenze (IT).*

[2] *IMT Lucca Institute for Advanced Studies - Piazza S. Ponziano 6, 55100 Lucca (IT).*

Received January 2011

We study the asymptotic behaviour of (a) information leakage and (b) adversary's error probability in information hiding systems modelled as noisy channels. Specifically, we assume the attacker can make a single guess after observing n independent executions of the system, throughout which the secret information is kept fixed. We show that the asymptotic behaviour of quantities (a) and (b) can be determined in a simple way from the channel matrix. Moreover, simple and tight bounds on them as functions of n show that the convergence is exponential. We also discuss feasible methods to evaluate the rate of convergence. Our results cover both the Bayesian case, where a prior probability distribution on the secrets is assumed known to the attacker, and the maximum-likelihood case, where the attacker does not know such distribution. In the Bayesian case, we identify the distributions that maximize the leakage. We consider both the min-entropy setting studied by Smith and the additive form recently proposed by Braun et al., and show the two forms do agree asymptotically. Next, we extend these results to a more sophisticated eavesdropping scenario, where the attacker can perform a (noisy) observation at each state of the computation and the systems are modelled as hidden Markov models.

Keywords: security, quantitative information leakage, information theory, Bayes risk, hidden Markov models.

1. Introduction

In recent years there has been much interest in formal models to reason about quantitative information leakage in computing systems (Clark *et al.* 2001; Chatzikokolakis *et al.* 2008; Boreale 2009; Köpf and Basin 2007; Backes and Köpf 2008; Smith 2009; Standaert *et al.* 2009). A general situation is that of a program, protocol or device carrying out computations that depend probabilistically on a secret piece of information, such as a password, the identity of a user or a private key. We collectively designate these as *information hiding systems*, following a terminology established in (Chatzikokolakis *et al.* 2008). During the computation, some observable

[†] Extended version of (Boreale *et al.* 2011). Corresponding author: Michele Boreale, Università di Firenze, Dipartimento di Sistemi e Informatica, Viale Morgagni 65, I-50134 Firenze, Italy. E-mail: `boreale@dsi.unifi.it`. Work partially supported by the EU project ASCENS under the FET open initiative in FP7.

information related to the secret may be disclosed. This might happen either by design, e.g. if the output of the system is directly related to the secret (think of a password checker denying access), or for reasons depending on the implementation. In the latter case, the observable information may take the form of physical quantities, such as the execution time or the power consumption of the device (think of timing and power attacks on smart cards (Kocher 1996; Kocher *et al.* 1999)). The observable information released by the system can be exploited by an eavesdropper to reconstruct the secret, or at least to limit the search space. This is all the more true when the eavesdropper is given the ability of observing several executions of the system, thus allowing her/him to mount some kind of statistical attack.

A simple but somehow crucial remark due to Chatzikokolakis et al. (Chatzikokolakis *et al.* 2008) is that, for the purpose of quantifying the amount of secret information that is leaked, it is useful to view an information hiding system as a *channel* in the sense of Information Theory: the inputs represent the secret information, the outputs represent the observable information and the two sets are related by a conditional probability matrix. This remark suggests a natural formalization of leakage in terms of Shannon entropy based metrics, like mutual information and capacity. In fact, by a result due to Massey (Massey 1994), these quantities are strongly related to the resistance of the system against *brute-force* attacks. Specifically, Shannon entropy is related to the average number of questions of the form "is the secret equal to x?" an attacker has to ask an oracle in order to identify the secret *with certainty*. In a recent paper, Smith (Smith 2009) objects that, even if the number of such questions is very high, the attacker might still have a significant chance of correct guess in just one or very few attempts. Smith demonstrates that *min-entropy* quantities, based on error probability (a.k.a. *Bayes risk*), are more adequate to express leakage in this *one-try* scenario. Whatever the considered attack scenario, brute-force or one-try, the analytic computation of leakage is in general difficult or impossible. Henceforth, a major challenge is being able to give simple and tight bounds on leakage in general, or exact expressions that exploit specific properties of a system (e.g. symmetries in the channel matrix) in some special cases.

In the present paper, we tackle these issues in a scenario of one-try attacks and system re-execution. More precisely, we assume the attacker makes his guess after observing several, say n, independent executions of the system, throughout which the secret information is kept fixed. In real-world situations, re-execution may happen either forced by the attacker (think of an adversary querying several times a smart card), or by design (think of routing paths established repeatedly between a sender and a receiver in anonymity protocols like Crowds (Reiter and Rubin 1998)). Since the computation is probabilistic, in general the larger the number n of observed executions, the more information will be gained by the attacker. Therefore, it is important to asses the resistance of a system in this scenario.

Our goal is to describe the asymptotic behaviour of the adversary's error probability and of information leakage as n goes to ∞. We show that the asymptotic values of these quantities can be determined in a simple way from the channel matrix. Moreover, we provide simple and tight bounds on error probability and on leakage as functions of n, showing that the convergence is exponential. We also discuss feasible methods for evaluating the rate of convergence. Our results cover both the Bayesian case (MAP rule), where a prior probability distribution on the secrets is assumed known to the attacker, and the maximum-likelihood case (ML rule), where the attacker does not know such distribution. In the Bayesian case, we identify the distributions that maximize leakage. We consider both the min-entropy leakage studied by Smith (Smith 2009) and the additive form recently proposed by Braun et al. (Braun *et al.* 2009), and show the two forms do agree asymptotically.

We next consider a more sophisticated scenario, where computations of the system may take several steps to terminate, or even not terminate at all. In any case, to each state of the computation there corresponds one (in general, noisy) observation on the part of the attacker. Hence, to each computation there corresponds a sequential *trace* of observations. The attacker may collect multiple such traces, corresponding to multiple independent executions of the system. Like in the simpler scenario, the secret is kept fixed throughout these executions. This set up is well suited to describe situations where the attacker collects information from different sources at different times, like in a coalition of different local eavesdroppers. An instance of this situation in the context of an anonymous routing application will be examined. We formalize this scenario in terms of discrete-time *Hidden Markov Models* (Rabiner 1989) and then show that the results established for the simpler scenario carry over to the new one.

Throughout the paper, we illustrate our results with a few examples: the modular exponentiation algorithm used in public-key cryptography, the Crowds anonymity protocol, S-boxes in block ciphers and onion routing protocols (Goldschlag *et al.* 1998) in a network with a fixed topology.

Related work

The last few years have seen a flourishing of research on quantitative models of information leakage. In the context of language-based security, Clark et al. (Clark *et al.* 2001) first motivated the use of mutual information to quantify information leakage in a setting of imperative programs. Boreale (Boreale 2009) extended this study to the setting of process calculi, and introduced a notion of rate of leakage. In both cases, the considered systems do not exhibit probabilistic behaviour. Closely related to ours is the work by Chatzikokolakis, Palamidessi and their collaborators. (Chatzikokolakis *et al.* 2008) examines information leakage mainly from the point of view of Shannon entropy and capacity, but also contains results on asymptotic error probability, showing that, independently from the input distribution, the ML rule approximates the MAP rule. (Chatzikokolakis *et al.* 2008b) studies error probability mainly relative to one observation ($n = 1$), but also offers a lower-bound in the case of repeated observations (Chatzikokolakis *et al.* 2008b, Proposition 7.4). This lower-bound is generalized by our results. Compositional methods based on process algebras are discussed in (Braun *et al.* 2008); there, the average ML error probability is characterized in terms of MAP error probability under a uniform distribution of inputs. (Braun *et al.* 2009) introduces the notion of additive leakage and compares it to the min-entropy based leakage considered by Smith (Smith 2009), but again in the case of a single observation.

A model of "unknown-message" attacks is considered by Backes and Köpf in (Backes and Köpf 2008). This model is basically equivalent to the information hiding systems considered in (Chatzikokolakis *et al.* 2008; Chatzikokolakis *et al.* 2008b; Braun *et al.* 2009) and in the present paper. Backes e Köpf too consider a scenario of repeated independent observations, but from the point of view of Shannon entropy, rather than of error probability. They rely on the information-theoretic method of types to determine the asymptotic behaviour of the considered quantities, as we do in the present paper. An application of their setting to the modular exponentiation algorithm is the subject of (Köpf and Dürmuth 2009), where the effect of *bucketing* on security of RSA is examined (see Section 5). This study has recently been extended to the case of one-try attacks by Köpf and Smith in (Köpf and Smith 2010). Earlier, Köpf and Basin had considered a scenario of adaptive chosen-message attacks (Köpf and Basin 2007). They offer an algorithm to compute conditional Shannon entropy in this setting, but not a study of its asymptotic behaviour, which seems very difficult to characterize.

In the context of side-channel cryptanalysis, Standaert et al. propose a framework to reason on side-channel correlation attacks (Standaert *et al.* 2009). Both a Shannon entropy based metric and a security metric are considered. This model does not directly compare to ours, since, as we will discuss in Section 5, correlation attacks are inherently known-message – that is, they presuppose the explicit or implicit knowledge of the plaintext on the part of the attacker.

Hypothesis testing is at the basis of the analysis considered in the present paper. The classical, binary case is covered in (Cover and Thomas 2006, Ch.11). Baignères and Vaudenay (Baignères and Vaudenay 2008) refine these results and characterize the optimal asymptotic rate of convergence in a number of variations of the basic setting, including the case where one of the two hypotheses is "composite" – that is, consisting of several sub-hypotheses chosen according to a prior probability distribution. They apply these results to study the *advantage* an attacker may have in distinguishing the output of a given cipher from a random output.

Structure of the paper

The rest of the paper is organized as follows. Section 2 establishes some notations and terminology. Section 3 introduces the model and the quantities that are the object of our study. Section 4 discusses the main results about error probability and leakage. Section 5 illustrates these results with a few examples. Section 6 presents the extension to hidden Markov models. Section 7 illustrates the new setting with an example. Section 8 contains some concluding remarks.

2. Notations and preliminary notions

Let $\mathcal{A}$ be a finite nonempty set. A probability distribution on a $\mathcal{A}$ is a function $p : \mathcal{A} \to [0,1]$ such that $\sum_{a\in\mathcal{A}} p(a) = 1$. For any $A \subseteq \mathcal{A}$ we let $p(A)$ denote $\sum_{a\in A} p(a)$. Given $n \geq 0$, we let $p^n : \mathcal{A}^n \to [0,1]$ be the n-th extension of p, defined as $p^n((a_1,\ldots,a_n)) \stackrel{\triangle}{=} \Pi_{i=1}^n p(a_i)$; this is in turn a probability distribution on $\mathcal{A}^n$. For $n = 0$, we set $p^0(\epsilon) = 1$, where ϵ denotes here the empty string. Given two distributions p and q on $\mathcal{A}$, the *Kullback-Leibler (KL) divergence* of p and q is defined as (all the log's are taken with base 2)

$$D(p||q) \stackrel{\triangle}{=} \sum_{a\in\mathcal{A}} p(a) \cdot \log \frac{p(a)}{q(a)}$$

with the proviso that $0 \cdot \log \frac{0}{q(a)} = 0$ and that $p(a) \cdot \log \frac{p(a)}{0} = +\infty$ if $p(a) > 0$. It can be shown that $D(p||q) \geq 0$, with equality if and only if $p = q$ (*Gibbs inequality*). KL-divergence can be thought of as a sort of distance between p and q, although strictly speaking it is not – it is not symmetric, nor satisfies the triangle inequality.

$\Pr(\cdot)$ will generally denote a probability measure. Given a random variable X taking values in $\mathcal{A}$, we write $X \sim p$ if X is distributed according to p, that is for each $a \in \mathcal{A}$, $\Pr(X = a) = p(a)$.

3. Probability of error, leakage, indistinguishability

Definition 1. An *information hiding system* is a quadruple $\mathcal{H} = (\mathcal{S}, \mathcal{O}, p(\cdot), p(\cdot|\cdot))$, composed by a finite set of *states* $\mathcal{S} = \{s_1, ..., s_m\}$ representing the secret information, a finite set of *observables* $\mathcal{O} = \{o_1, ..., o_l\}$, an a priori probability distribution on $\mathcal{S}$, p, and a *conditional probability matrix*, $p(\cdot|\cdot) \in [0, 1]^{\mathcal{S}\times\mathcal{O}}$, where each row sums up to 1.

The entry of row s and column o of this matrix will be written as $p(o|s)$, and represents the probability of observing o given that s is the (secret) input of the system. For each s, the s-th row of the matrix is identified with the probability distribution $o \mapsto p(o|s)$ on $\mathcal{O}$, denoted by p_s. The probability distribution p on $\mathcal{S}$ and the conditional probability matrix $p(o|s)$ together induce a probability distribution r on $\mathcal{S} \times \mathcal{O}$ defined as $r(s, o) \triangleq p(s) \cdot p(o|s)$, hence a pair of random variables $(S, O) \sim r$, with S taking values in $\mathcal{S}$ and O taking values in $\mathcal{O}$. Note that $S \sim p$ and, for each s and o s.t. $p(s) > 0$, $\Pr(O = o|S = s) = p(o|s)$.

Let us now discuss the attack scenario. Given any $n \geq 0$, we assume the adversary is a passive eavesdropper that gets to know the observations corresponding to n independent executions of the system, $o^n = (o_1, ..., o_n) \in \mathcal{O}^n$, throughout which the secret state s is kept fixed. Formally, the adversary knows a random vector of observations $O^n = (O_1, ..., O_n)$ such that, for each $i = 1, ..., n$, O_i is distributed like O and the individual O_i are *conditionally independent* given S, that is, the following equality holds true for each $o^n \in \mathcal{O}^n$ and $s \in \mathcal{S}$ s.t. $p(s) > 0$

$$\Pr\left(O^n = (o_1, \ldots, o_n) \,|\, S = s\right) = \Pi_{i=1}^n p(o_i|s) \,.$$

We will often abbreviate the right-hand side of the above equation as $p(o^n|s)$. For any n, the attacker strategy is modeled by a function $g : \mathcal{O}^n \to \mathcal{S}$, called *guessing function*: this represents the single guess the attacker is allowed to make about the secret state s, after observing o^n.

Definition 2 (error probability). Let $g : \mathcal{O}^n \to \mathcal{S}$ be a guessing function. The *probability of error after n observations* (relative to g) is given by

$$P_e^{(g)}(n) \triangleq 1 - P_{succ}(n), \quad \text{where } P_{succ}^{(g)}(n) \triangleq \Pr(g(O^n) = S)\,.$$

It is well-known (see e.g. (Cover and Thomas 2006)) that the optimal strategy for the adversary, that is the one that minimizes the error probability, is the Maximum A Posteriori (MAP) rule, defined below.

Definition 3 (Maximum A Posteriori rule, MAP). A function $g : \mathcal{O}^n \to \mathcal{S}$ satisfies the *Maximum A Posteriori (*MAP*) criterion* if for each o^n and s

$$g(o^n) = s \text{ implies } p(o^n|s)p(s) \geq p(o^n|s')p(s') \text{ for each } s'\,.$$

In the above definition, for $n = 0$ one has $o^n = \epsilon$, and it is convenient to stipulate that $p(\epsilon|s) = 1$: that is, with no observations at all, g selects some s maximizing the prior distribution. With this choice, $P_e^{(g)}(0)$ denotes $1 - \max_s p(s)$. It worthwhile to note that, once n and $p(s)$ are fixed, the MAP guessing function is not in general unique. It is readily checked, though, that $P_e(n)$ does *not* depend on the specific MAP function g that is chosen. Hence, throughout the paper we assume w.l.o.g. a fixed guessing function g for each given n and probability distribution $p(s)$. We shall omit the superscript $^{(g)}$, except where this might cause confusion.

Another widely used criterion is *Maximum Likelihood (*ML*)*, which given o^n selects a state

s maximizing the likelihood $p(o^n|s)$ among all the states. ML coincides with MAP if the uniform distribution on the states is assumed. ML is practically important because it requires no knowledge of the prior distribution, which is often unknown in security applications. Our main results will also apply to the ML rule (see Remark 2 in the next section).

We now come to information leakage: this is a measure of the information leaked by the system, obtained by comparing the prior and the posterior (to the observations) success probabilities. Indeed, two flavours of this concept naturally arise, depending on how the comparison between the two probabilities is expressed. If one uses subtraction, one gets the additive form of (Braun *et al.* 2009), while if one uses the ratio between them, one gets a multiplicative form. In the latter case, one could equivalently consider the difference of the log's, obtaining the *min-entropy* based definition considered by Smith (Smith 2009)[†].

Definition 4 (Additive and Multiplicative Leakage (Braun *et al.* 2009; Smith 2009)). The *additive* and *multiplicative leakage* after n observations are defined respectively as

$$L_+(n) \triangleq P_{succ}(n) - \max_s p(s) \quad \text{and} \quad L_\times(n) \triangleq \frac{P_{succ}(n)}{\max_s p(s)} .$$

In an information hiding system, it may happen that two secret states induce the same distribution on the observables. A common example is that of a degenerate channel matrix modelling a deterministic function $\mathcal{S} \to \mathcal{O}$ with $|\mathcal{O}| < |\mathcal{S}|$. An important role in determining the fundamental security parameters of the system will be played by an indistinguishability equivalence relation over states, which is defined in the following. Recall that, for each $s \in \mathcal{S}$, we let p_s denote the probability distribution $p(\cdot|s)$ on $\mathcal{O}$.

Definition 5 (Indistinguishability). Given $s, s' \in \mathcal{S}$, we let $s \equiv s'$ iff $p_s = p_{s'}$.

Concretely, two states are indistinguishable iff the corresponding rows in the conditional probability matrix are the same. This intuitively says that there is no way for the adversary to tell them apart, no matter how many observations he performs. We stress that this definition does not depend on the prior distribution on states, nor on the number n of observations.

4. Bounds and asymptotic behaviour

We introduce some notation that will be used throughout the section. Let $\mathcal{S}/\equiv$ be $\{C_1, ..., C_K\}$, the set of equivalence classes of $\equiv$. For each $i = 1, ..., K$, let

$$s_i^* \triangleq \text{argmax}_{s \in C_i} p(s) \quad \text{and} \quad p_i^* \triangleq p(s_i^*) . \tag{1}$$

We assume wlog that $p_i^* > 0$ for each $i = 1, ..., K$ (otherwise all the states in class C_i can be just discarded from the system).

[†] Smith (Smith 2009) defines the leakage as $\log \frac{V_{post}}{V_{pr}}$, where, using our notation, $V_{pr} \triangleq \max_s p(s)$ is the *prior vulnerability* and $V_{post} \triangleq \sum_{o^n} \Pr(O^n = o^n) \cdot \max_s \Pr(S = s|O^n = o^n)$ is the *posterior vulnerability* (after n observations; Smith only defines the case $n = 1$). To see that $V_{post} = P_{succ}(n)$, just note that $P_{succ}(n) = \sum_{o^n} \Pr(O^n = o^n) \cdot \Pr(g(o^n) = S|O^n = o^n) = \sum_{o^n} \Pr(O^n = o^n) \cdot \max_s \Pr(S = s|O^n = o^n)$, where the last equality follows because g is MAP.

4.1. *Main results*

We shall prove the following bounds and asymptotic behaviour for $P_e(n)$.

Theorem 1. $P_e(n)$ converges exponentially fast to $1 - \sum_{i=1}^K p_i^*$. More precisely, there is $\epsilon > 0$ s.t.

$$1 - \sum_{i=1}^K p_i^* \;\leq\; P_e(n) \;\leq\; 1 - (\sum_{i=1}^K p_i^*) \cdot r(n)$$

where $r(n) = 1 - (n+1)^{|O|} \cdot 2^{-n\epsilon}$. Here, the lower-bound holds true for any n, while the upper-bound holds true for any $n \geq n_0 \triangleq \epsilon^{-1} \cdot \max_{i,j} \log(\frac{p_i^*}{p_j^*})$. Moreover, ϵ only depends on the rows $p_{s_i^*}$ $(i = 1, ..., K)$ of the conditional probability matrix $p(\cdot|\cdot)$.

Note that in the practically important case of the uniform distribution on states, we have $n_0 = 0$, that is the upper-bound as well holds true for any n. The theorem has a simple interpretation in terms of the attacker's strategy: after infinitely many observations, he can determine the indistinguishability class of the secret, say C_i, and then guess the most likely state in that class, s_i^*.

In order to discuss this result, we recall some terminology and a couple of preliminary results from the information-theoretic method of types (Cover and Thomas 2006, Ch.11). Given $n > 0$, a sequence $o^n \in O^n$ and a symbol $o \in O$, let us denote by $\mathrm{n}(o, o^n)$ the number of occurrences of o inside o^n. The *type* (or empirical distribution) of o^n is the probability distribution t_{o^n} on O defined as: $t_{o^n}(o) \triangleq \frac{\mathrm{n}(o,o^n)}{n}$. Let q any probability distribution on O. A *neighborhood* of q is a subset of n-sequences of O^n whose empirical distribution is close to q. Formally, for each $n \geq 1$ and $\epsilon > 0$

$$U_q^{(n)}(\epsilon) \triangleq \{o^n \in O^n \,|\, D(t_{o^n}||q) \leq \epsilon\}\,.$$

The essence of the method of types is that (i) there is only a polynomial number of types in n, and that (ii) the probability under q of the set of n-sequences of a given type decreases exponentially with n, at a rate determined by the KL-divergence between q and that type. These considerations are made precise and exploited in the proof of the following lemma, which can be found in (Cover and Thomas 2006, Ch.11). The lemma basically says that the probability that a sequence falls in a neighborhood of q of radius ϵ approaches 1 exponentially fast with n.

Lemma 1. Let q be a probability distribution on O. Then $q^n(U_q^{(n)}(\epsilon)) \geq 1 - (n+1)^{|O|} \cdot 2^{-n\epsilon}$.

We shall also rely upon the well-known fact that the MAP test can be expressed in terms of KL-divergence. Basically, the distribution that is most likely to have generated a given sequence is the one that is closest to the type of the sequence in the sense of KL-divergence. In the statement, α and β represent the prior probabilities of the states corresponding to the two distributions. The proof follows from easy manipulations of log's and summations (see (Cover and Thomas 2006, Ch.11)).

Lemma 2. Let p and q be two distributions on O, $\alpha,\beta > 0$ and $o^n \in O^n$. Then $p^n(o^n)\alpha > q^n(o^n)\beta$ is equivalent to $D(t_{o^n}||p) < D(t_{o^n}||q) + \frac{1}{n}\log\frac{\alpha}{\beta}$.

Let us now come back to the proof of the main result. For any $s \in S$, we let $A_s^{(n)} \triangleq g^{-1}(s) \subseteq O^n$ be the *acceptance region* for state s. We note that it is not restrictive to assume that g maps each o^n in one of the K representative elements $s_1^*, ..., s_K^*$ that maximize the prior: indeed, if this were not the case, it would be immediate to build out of g a new MAP function that fulfills this requirement. Thus, from now on we will assume w.l.o.g. that $A_s^{(n)} = \emptyset$ for $s \neq s_1^*, ..., s_K^*$. For the sake of notation, from now on we will denote $U^{(n)}_{p_{s_i^*}}$ as $U_i^{(n)}$ and $A^{(n)}_{s_i^*}$ as $A_i^{(n)}$, for $i = 1, ..., K$. The sets $U_i^{(n)}$ and $A_i^{(n)}$ are related by the following lemma.

Lemma 3. There is $\epsilon > 0$, not depending on the prior probability on states, such that for each $n \geq n_0$ as defined in Theorem 1 and for each $i = 1, ..., K$, it holds that $U_i^{(n)}(\epsilon) \subseteq A_i^{(n)}$.

Proof. We choose $\epsilon > 0$ s.t. for any $i \neq j$, $U_i^{(n)}(2\epsilon) \cap U_j^{(n)}(2\epsilon) = \emptyset$: the existence of such an ϵ is guaranteed by (Cover and Thomas 2006, Lemma 11.6.1) and only depends on $p_{s_1^*}, ..., p_{s_K^*}$ (see also Proposition 1 later in this section for an estimation of the permissible ϵ's). Fix $i \in \{1, ..., K\}$ and $o^n \in U_i^{(n)}(\epsilon)$, we will show that, for n large enough, $o^n \in A_i^{(n)}$. By the conditions on ϵ, we have that for any $j \neq i$:

$$D(t_{o^n} \| p_{s_i^*}) \ \leq \epsilon \ \text{ and } \ D(t_{o^n} \| p_{s_j^*}) \ > \ 2\epsilon\,.$$

After some algebra, one gets that, for any $n \geq n_0$

$$D(t_{o^n} \| p_{s_i^*}) \ < \ D(t_{o^n} \| p_{s_j^*}) + \frac{1}{n} \log \frac{p_i^*}{p_j^*}\,. \tag{2}$$

Now, by Lemma 2, inequality (2) is equivalent to

$$p(o^n|s_i^*)p(s_i^*) \ = \ p^n_{s_i^*}(o^n)p_i^* \ > \ p^n_{s_j^*}(o^n)p_j^* \ = \ p(o^n|s_j^*)p(s_j^*)\,. \tag{3}$$

Since this inequality holds true for each $j \neq i$, by definition of MAP rule we deduce that g maps o^n to s_i^*, that is $o^n \in A_i^{(n)}$. □

We now come to the proof of the main theorem above.

Proof. (of Theorem 1). We focus equivalently on the probability of success, $P_{succ}(n)$. Under the assumptions on g explained above, we compute as follows

$$\begin{array}{rclcl} P_{succ}(n) & = & \sum_{s \in \mathcal{S}} \Pr(g(O^n) = S | S = s) p(s) & = & \sum_{s \in \mathcal{S}} p_s^n(A_s^{(n)}) p(s) \\ & = & \sum_{i=1}^K \underbrace{p^n_{s_i^*}(A_i^{(n)})}_{\leq 1} p_i^* & \leq & \sum_{i=1}^K p_i^* \end{array}$$

which implies the lower-bound in the statement. Choose now ϵ as given by Lemma 3. Let $n \geq n_0$. Note that for $n = 0$ the upper-bound holds trivially, as $P_e(0) = 1 - \max_s p(s)$, so assume $n \geq 1$. For each $i = 1, ..., K$ we have

$$p^n_{s_i^*}(A_i^{(n)}) \ \geq \ p^n_{s_i^*}(U_i^{(n)}(\epsilon)) \ \geq \ 1 - (n+1)^{|\mathcal{O}|} \cdot 2^{-n\epsilon}$$

where the first inequality comes from Lemma 3 and second one from Lemma 1. In the end, from $P_{succ}(n) = \sum_{i=1}^K p^n_{s_i^*}(A_i^{(n)}) p_i^*$, we obtain that for $n \geq n_0$

$$P_{succ}(n) \ \geq \ (\sum_{i=1}^K p_i^*) \cdot (1 - (n+1)^{|\mathcal{O}|} \cdot 2^{-n\epsilon})$$

which implies the upper-bound in the statement. □

Remark 1. *In the expression for $r(n)$, the term $(n+1)^{|\mathcal{O}|}$ is a rather crude upper bound on the number of types of n-sequences. It is possible to replace this term with the expression*

$$\binom{n + |\mathcal{O}| - 1}{|\mathcal{O}| - 1}$$

which is less easy to manipulate analytically, but gives the exact number of types, hence a more accurate upper bound on $P_e(n)$.

The following results show that, asymptotically, the security of the systems is tightly connected to the number of its indistinguishability classes – and in the case of uniform prior distribution *only* depends on this number

Corollary 1. If the a priori distribution on $\mathcal{S}$ is uniform, then $P_e(n)$ converges exponentially fast to $1 - \frac{K}{|\mathcal{S}|}$.

Remark 2 (on the ML rule). *(Braun* et al. *2008) shows that the probability of error under the* ML *rule,* averaged *on all distributions, coincides with the probability of error under the* MAP *rule and the uniform distribution. From Corollary 1 we therefore deduce that the average* ML *error converges exponentially fast to the value $1 - \frac{K}{|\mathcal{S}|}$ as $n \to \infty$.*

We discuss now some consequences of the above results on information leakage. Recall that for $i = 1, ..., K$, we call s_i^* a representative of the indistinguishability class C_i that maximizes the prior distribution $p(s)$ in the class C_i, and let $p_i^* = p(s_i^*)$. Assume w.l.o.g. that $p_1^* = \max_s p(s)$. In what follows, we denote by p_{max} the distribution on $\mathcal{S}$ defined by: $p_{max}(s) = \frac{1}{K}$ if $s \in \{s_1^*, ..., s_K^*\}$ and $p_{max}(s) = 0$ otherwise.

Corollary 2.

1 $L_+(n)$ converges exponentially fast to $\sum_{i=2}^K p_i^*$. This value is maximized by the prior distribution p_{max}, which yields the limit value $1 - \frac{1}{K}$.

2 $L_\times(n)$ converges exponentially fast to $\frac{\sum_{i=1}^K p_i^*}{p_1^*}$. This value is maximized by the prior distribution p_{max}, which yields the limit value K.

Proof.

1 The value of the limit follows directly from the definition of $L_+(n)$ and Theorem 1. Concerning the second part, for any fixed $p(s)$, it is easily checked that $\sum_{i=2}^K p_i^* \leq 1 - \frac{1}{K}$ (this is done by separately considering the cases $\max_s p(s) \geq \frac{1}{K}$ and $\max_s p(s) < \frac{1}{K}$). But the value $1 - \frac{1}{K}$ is obtained asymptotically with the distribution p_{max}.

2 Again, the value of the limit follows directly from the definition of $L_\times(n)$ and Theorem 1. Concerning the second part, for any fixed $p(s)$, of course we have $\sum_{i=1}^K \frac{p_i^*}{p_1^*} \leq \sum_{i=1}^K 1 = K$. But the value K is obtained asymptotically with the distribution p_{max}.

□

Remark 3. *A consequence of Corollary 2(2) is that, in the case of uniform distribution on states, the multiplicative leakage coincides with the number of equivalence classes K. This generalizes a result of (Smith 2009) for deterministic systems.*

In (Braun *et al.* 2009) additive and multiplicative leakages are compared in the case of a single observation ($n = 1$). It turns out that, when comparing two systems, the two forms of leakage are in agreement, in the sense that they individuate the same maximum-leaking system w.r.t. a fixed prior distribution on inputs. However, (Braun *et al.* 2009) also shows that the two forms disagree as to the distribution on inputs that maximizes leakage w.r.t. a fixed system. This is shown to be the uniform distribution in the case of multiplicative leakage, and a function that uniformly distributes the probability on the set of "corner points" in the case of additive leakage (see (Braun *et al.* 2009) for details). Here, we have shown that, despite this difference, additive and multiplicative leakage do agree on the maximizing distribution asymptotically.

4.2. *Rate of convergence*

The quantity ϵ in the statement of Theorem 1 determines how fast the error probability approaches its limit value. Let us call *achievable* any $\epsilon > 0$ for which the upper bound in Theorem 1 holds true for any $n \geq n_0$. The following result gives sufficient and practical conditions for achievability. Let us stress that the achievable rates given by this proposition do not depend on the prior distribution, but only on the relation $\equiv$, and specifically on the minimum norm 1 distance between equivalence classes: the larger this distance, the higher the achievable rates. This result is essentially a re-elaboration on (Cover and Thomas 2006, Lemma 11.6.1).

Proposition 1. Let $\delta \triangleq \min_{s_i \not\equiv s_j} ||p_{s_i} - p_{s_j}||_1$. Then any rate ϵ satisfying $0 < \epsilon < \frac{\delta^2}{16 \ln 2}$ is achievable.

Proof. Using the notation introduced immediately above Lemma 3, we show that, for any ϵ satisfying the hypotheses in the present lemma, one has $U_i^{(n)}(2\epsilon) \cap U_j^{(n)}(2\epsilon) = \emptyset$ for $i \neq j$, which guarantees that ϵ can be used in the proof of Lemma 3, hence in the statement of Theorem 1. According to (Cover and Thomas 2006, Lemma 11.6.1), for any two distributions p and q on the same set O, it holds true that $D(p||q) \geq \frac{1}{2 \ln 2}||p - q||_1^2$. Take any $o^n \in U_i^{(n)}(2\epsilon)$. Assume by contradiction that $o^n \in U_j^{(n)}(2\epsilon)$ for any $j \neq i$. Then we would get

$$\begin{aligned}\delta &\leq ||p_{s_i^*} - p_{s_j^*}||_1 \leq ||p_{s_i^*} - t_{o^n}||_1 + ||t_{o^n} - p_{s_j^*}||_1 \\ &\leq \sqrt{2 \ln 2 D(t_{o^n}||p_{s_i^*})} + \sqrt{2 \ln 2 D(t_{o^n}||p_{s_j^*})} \\ &\leq 2\sqrt{4\epsilon \ln 2} \\ &< \delta\end{aligned}$$

where: the first inequality above exploits the triangle inequality for $|| \cdot ||_1$, the second one exploits the inequality $D(p||q) \geq \frac{1}{2 \ln 2}||p - q||_1^2$ mentioned above, the third one follows by definition of the balls $U_j^{(n)}(2\epsilon)$ and $U_i^{(n)}(2\epsilon)$, and the last one follows from the hypothesis on ϵ and some algebra. □

In the practically important case where the p_i^*'s are all equal, the above proposition can be strengthened as follows.

Proposition 2. Suppose $p_1^* = p_2^* = \cdots = p_K^*$. Let δ be like in Proposition 1. Then any rate ϵ satisfying $0 < \epsilon < \frac{\delta^2}{8 \ln 2}$ is achievable.

Proof. In the proof of Lemma 3, the condition (3) is equivalent to (2). Under the hypothesis the p_i^*'s are all equal, the term $\frac{1}{n} \log \frac{p_j^*}{p_i^*}$ vanishes, so (2) reduces to $D(t_{o^n}||p_{s_i^*}) < D(t_{o^n}||p_{s_j^*})$. For this to be the case, it is sufficient that $U_i^{(n)}(\epsilon) \cap U_j^{(n)}(\epsilon) = \emptyset$. Proceeding like in the proof of Proposition 1, we can show that the conditions on ϵ imposed in the present proposition are sufficient to guarantee this disjointness. □

The above results prompt the following question. Suppose one somehow ignores the rows of $p(\cdot|\cdot)$ that are close together with each other, and only consider rows that are far from each other: is it then possible to achieve a higher rate of convergence ϵ? The answer is expected to be *yes*, although ignoring some rows might lead to a possibly higher asymptotic error probability. In other word, it should be possible to trade off accuracy in guessing with rate of convergence. This is the content of the next proposition.

Proposition 3. Let $\emptyset \neq \mathcal{S}_0 \subseteq \{s_1^*, ..., s_K^*\}$. Then there is $\epsilon > 0$ only depending on the rows p_s, $s \in \mathcal{S}_0$, of $p(\cdot|\cdot)$, such that for each $n \geq n_0 \triangleq \epsilon^{-1} \max_{s_i^*, s_j^* \in \mathcal{S}_0} \log(\frac{p_i^*}{p_j^*})$, it holds true that

$$P_e(n) \leq 1 - (\sum_{s_j^* \in \mathcal{S}_0} p_j^*) \cdot r(n) \quad \text{with} \quad r(n) = 1 - (n+1)^{|O|} \cdot 2^{-n\epsilon} .$$

Proof. For any n, let $A_s^{(n)}$ ($s \in \mathcal{S}$) be the acceptance regions determined by any MAP guessing function. Choose any $s^* \in \mathcal{S}_0$. Define the new acceptance regions $B_s^{(n)}$ as follows: $B_s^{(n)} \triangleq \emptyset$ if $s \notin \mathcal{S}_0$; $B_s^{(n)} \triangleq A_s^{(n)}$ if $s \in \mathcal{S}_0 \setminus \{s^*\}$; $B_{s^*}^{(n)} \triangleq A_{s^*}^{(n)} \cup \cup_{s \notin \mathcal{S}_0} A_s^{(n)}$. For each n, the regions $B_s^{(n)}$ determine a new guessing function, say g', which will in general *not* be MAP. Now, repeating the computation in the proof of Theorem 1 with the regions $B_s^{(n)}$ instead of $A_s^{(n)}$, one finds

$$P_e^{(g')}(n) \leq 1 - (\sum_{s_j^* \in \mathcal{S}_0} p_j^*) \cdot (1 - (n+1)^{|O|} \cdot 2^{-n\epsilon}) .$$

The wanted result follows from the optimality of the MAP rule, which implies $P_e(n) \leq P_e^{(g')}(n)$. □

These concepts are demonstrated in the following example.

Example 1. *Consider the small imperative procedure* `p()` *described below. There,* `h` *and* `l` *are two-bits integer global variables, while* `rnd()` *is a procedure returning a random real value in the interval* $[0,1]$. *Boolean values* `true` *and* `false` *are identified with integers 1 and 0, respectively.*

```
proc p();
{
 l=rnd();
 if not(h mod 2) then l=(l >= .5) else l=1+(l >= (.5 + (h div 2)*10^-5) );
 return l
}
```

Now we consider the case that `h` *is a sensitive variable, whose initial vale is initially chosen in the range 0..3, and then never modified. We assume that* `p()` *can be invoked several times. One is interested in analysing the asymptotic information leakage relative to* `h` *caused by* `p()`. *We can model the procedure* `p()` *as an information hiding system, as follows.*

Let $\mathcal{S} = \{0,1,2,3\}$ *be the set of possible values of* `h`, *and* $O = \{0,1,2\}$ *the set of possible values returned by* `p()`. *The prior probability distribution on* $\mathcal{S}$ *is non-uniform and given by:* $p(0) = p(1) = \frac{1}{2} - 10^{-9}$ *and* $p(2) = p(3) = 10^{-9}$. *The behaviour of* `p()` *can be described by the conditional probability matrix displayed on the next page.*

	0	1	2
0	$\frac{1}{2}$	$\frac{1}{2}$	*0*
1	*0*	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{2}$	$\frac{1}{2}$	*0*
3	*0*	$\frac{1}{2}+10^{-5}$	$\frac{1}{2}-10^{-5}$

Note that $0 \equiv 2$*. Applying Theorem 1, we find that, for n sufficiently large,* $1 - E \leq P_e(n) \leq 1 - E \cdot r(n)$*, where* $E = 1 - 10^{-9}$ *and* $r(n) = 1 - (n+1)^3 \cdot 2^{-n\epsilon}$*. Applying Proposition 1, we find that any rate* $\epsilon < 3.6067 \times 10^{-11}$ *is achievable. Thus the convergence to the value* $1 - E = 10^{-9}$ *is very slow. One wonders if there is some value* $1 - E'$ *that is only slightly higher than* $1 - E$*, but that can be reached much faster. This is indeed the case. Observe that rows* 1 *and* 3 *are very close with each other in norm-1 distance:* $||p_1 - p_3||_1 = 2 \times 10^{-5}$*. We can discard row* 3*, which has a very small probability, and then apply Proposition 3 with* $\mathcal{S}_0 = \{0, 1\}$ *to get*

$$P_e(n) \leq 1 - E' \cdot r'(n)$$

where $E' = \frac{1}{2} - 10^{-9} + \frac{1}{2} - 10^{-9} = 1 - 2 \times 10^{-9}$ *and* $r'(n) = 1 - (n+1)^3 \cdot 2^{-n\epsilon'}$*. The rate* ϵ' *can be computed by applying Proposition 2, as* $p(0) = p(1)$*. By doing so, we get that any* $\epsilon' < 0.18034$ *is achievable. This implies that the value* $1 - E'$ *is approached much faster as n grows. For instance, already after* $n = 350$ *invocations we get that* $(1 - E')/P_e(n) > 0.99$*.*

Remark 4 (on optimal achievable rates). *Proposition 1 does not give indications as to the best achievable rate. Now, the case* $|\mathcal{S}| = 2$*, in which the attacker has to decide between* s_1 *and* s_2*, corresponds to classical* Bayesian hypothesis testing. *In this case, provided the distributions* p_{s_1} *and* p_{s_2} *have both full support (are everywhere positive on* $\mathcal{O}$*), it is well-known that the optimal rate of convergence* ϵ *is given by the* Chernoff Information *between* p_{s_1} *and* p_{s_2} *(see (Cover and Thomas 2006, Ch.11) for details). It is possible to extend this result to the general case of multiple hypotesis testing, hence to our setting, again with the proviso that all the distributions* p_{s_i} *have full support: in this case, the rate of convergence will be given by the* least *Chernoff Information between any two distinct distributions (Leang and Johnson 1997). In practice, we have found that most of the times Proposition 1 and 2 provide very good approximations of the optimal achievable rates.*

5. Examples

5.1. *Timing leaks and blinding in modular exponentiation*

In the '90's, P. Kocher showed that RSA and other public-key crypto-systems are subject to side-channel attacks exploiting information leaked by implementations of the modular exponentiation algorithm, such as execution time (Kocher 1996) and/or power consumption (Kocher *et al.* 1999). Many of these attacks are based on the assumption that the attacker can observe repeated independent execution of the system, throughout which the exponent – the secret key – is kept fixed. Here, we concentrate on timing attacks. *Blinding* (Kocher 1996) was early proposed as a countermeasure to thwart such attacks. The essence of blinding is that exponentiation is performed on a random message unknown to the attacker, rather than on the original message (to be decrypted or digitally signed) known to the attacker. This appears to be sufficient to thwart Kocher's attack, which is of chosen-ciphertext type.

Köpf and Dürmuth (Köpf and Dürmuth 2009) have recently argued that, however effective in practice, the exact degree of protection provided by blinding remains unclear. They therefore propose to enhance blinding with *bucketing*, a technique by which the algorithm's execution times are adjusted so as to always fall in one of few predefined values. This allows them to prove information-theoretic bounds, based on Shannon's entropy, on the quantity of information leaked by the algorithm. This quantity is in fact very small for typical RSA parameters, which lets them conclude that bucketing in the presence of blinding provides a formal guarantee of security, at the cost of a very moderate overhead. In another recent work (Köpf and Smith 2010), Köpf and Smith have extended this result to the case of one-try attacks, where the leaked information is measured in terms of min-entropy, which is precisely the log of the multiplicative leakage $L_\times$ we have considered.

Below, we refine Köpf and Smith's analysis, under a reasonable assumption on the functioning of the algorithm that will be described shortly. We consider an implementation of the modular exponentiation algorithm with blinding, but *no* bucketing. To such an implementation, there corresponds an information hiding system where: $\mathcal{S} = \mathcal{K} = \{0,1\}^N$ is the set of private keys, i.e. the possible exponents of the algorithm, over which we assume a uniform distribution[‡]; $O = \{t_1, t_2, ...\}$ is the set of possible execution times; $p(t|k)$ is the probability that, depending on the deciphered message, the execution of the algorithm takes times t given that the private key is k. To be more specific about the last point, we assume an underlying set of messages $\mathcal{M}$, with a known prior distribution $p_M(m)$, and a function[§] $\text{time} : \mathcal{M} \times \mathcal{S} \to O$ that yields the duration of the execution of the algorithm when its argument is a given pair (m, k). Then the entries of the probability matrix $p(t|k)$ can be defined thus

$$p(t|k) \overset{\triangle}{=} \textstyle\sum_{m \in \mathcal{M}:\text{time}(m,k)=t} p_M(m)\,.$$

Now, modular exponentiation functions in such a way that at the i-th iteration ($0 \leq i < N$), either a squaring or *both* a squaring and a multiply are performed, depending on whether the i-th bit of the exponent is 0 or 1. Given this functioning, it seems reasonable to assume that, for each m, *the execution time only depends on the number of '1' digits in k*. In other words, we assume that whenever k and k' have the same Hamming weight, $\text{time}(m, k) = \text{time}(m, k')$, for any m. From this assumption and the definition of $p(t|k)$, it follows that whenever k and k' have the same Hamming weight then $p(\cdot|k) = p(\cdot|k')$. So, in the system there are *at most* as many $\equiv$-classes as Hamming weights, that is $N + 1$. The results in Section 4 then allow us to conclude that for any n

$$P_e(n) \geq 1 - \tfrac{N+1}{2^N}\,.$$

For any practical size of the key, say $N = 1024$, this value is ≈ 1. Accordingly, additive and multiplicative leakage satisfy, asymptotically,

[‡] In the case of RSA, a negligible fraction of the exponents is ruled out by virtue of number theoretic requirements, so the resulting distribution is not exactly uniform on $\mathcal{S}$. This fact does not substantially affect the significance of our analysis.

[§] A more realistic modeling would make $\text{time}(m, k)$ a joint probability distribution. Again, this modification would not affect the final result.

$$L_+ \leq \frac{N}{2^N} \quad \text{and} \quad L_\times \leq N + 1\,.$$

For any practical size of the key, say $N = 1024$, these upper bounds yield negligible values: $L_+ \approx 0$ and $L_\times \leq 1025$. In the latter case, taking the log we obtain that no more than $\log(1025) = 10.001$ bits of min-entropy are leaked, out of 1024. In conclusion, under the further assumption on the behaviour of modular exponentiation we made above, blinding alone appears to provide satisfactory guarantees of security against one-try attacks.

5.2. *Protocol re-execution in Crowds*

The Crowds protocol (Reiter and Rubin 1998) is designed for protecting the identity of the senders of messages in a network where some of the nodes may be corrupted, that is, under the control of an attacker. Omitting a few details, the functioning of the protocol can be described quite simply: the sender first forwards the message to a node of the network chosen at random; at any time, any node holding the message can decide whether to (a) forward in turn the message to another node chosen at random, or (b) submit it to the final destination. The choice between (a) and (b) is made randomly, with alternative (a) being assigned probability p_f (forwarding probability) and alternative (b) probability $1 - p_f$. The rationale here is that, even if a corrupted node C receives the message from a node N (in the Crowds terminology, *C detects N*), C, hence the attacker, cannot decide whether N is the original sender or just a forwarder. In fact, given that N is detected, the probability of N being the true sender is only slightly higher than that of any other node being the true sender. So the attacker is left with a good deal of uncertainty as to the sender's identity. Reiter and Rubin have showed that, depending on p_f, on the fraction of corrupted nodes in the network and on a few other conditions, Crowds offers very good guarantees of anonymity (see (Reiter and Rubin 1998)).

Chatzikokolakis et al. have recently analyzed Crowds from the point of view of information hiding systems and one-try attacks (Chatzikokolakis *et al.* 2008; Chatzikokolakis *et al.* 2008b). In their modelling, $\mathcal{S} = \{s_1, ..., s_m\}$ is the set of possible senders (honest nodes), while $\mathcal{O} = \{d_1, ..., d_m\}$ is the set of observables. Here each d_i has the meaning that node s_i has been detected by some corrupted node. The conditional probability matrix is given by

$$p(d_j|s_i) \stackrel{\triangle}{=} \Pr\left(s_j \text{ is detected} \mid s_i \text{ is the true sender and some honest node has been detected}\right)$$

(see (Reiter and Rubin 1998) for details of the actual computation of these quantities). An example of such a system with $m = 20$ users, borrowed from (Chatzikokolakis *et al.* 2008b), is given in Figure 1.

The interesting case for us is that of re-execution, in which the protocol is executed several times, either forced by the attacker himself (e.g. by having corrupted nodes suppress messages) or by some external factor, and the sender is kept fixed through the various executions. This implies the attacker collects a sequence of observations $o^n = (o_1, ..., o_n) \in \mathcal{O}^n$, for some n. The repeated executions are assumed to be independent, hence we are precisely in the setting considered in this paper. This case is also considered in (Chatzikokolakis *et al.* 2008b), which gives lower bounds for the error probability holding for any n. Our results in Section 4 generalize those in (Chatzikokolakis *et al.* 2008b) by providing both lower- and upper- bounds converging exponentially fast to the asymptotic error probability. As an example, for the system in the table above, we have $P_e(n) \to 0$, independently of the prior distribution on the senders. An achievable

$$\begin{array}{c|cccc} & d_1 & d_2 & \cdots & d_{20} \\ \hline s_1 & 0.468 & 0.028 & \cdots & 0.028 \\ s_2 & 0.028 & 0.468 & \cdots & 0.028 \\ \vdots & & & \vdots & \\ s_{20} & 0.028 & 0.028 & \cdots & 0.468 \end{array}$$

Figure 1. The conditional probability matrix of Crowds for 20 honest nodes, 5 corrupted nodes and $p_f = 0.7$.

convergence rate, estimated with the method of Proposition 1, is $\epsilon \approx 0.13965$. This implies that already after observing $n = 1000$ re-executions the probability of error is, using the refined bound given in Remark 1, < 0.01.

It is worth to stress that protocol re-execution is normally prevented in Crowds for the very reason that it decreases anonymity, although it may be necessary in some cases. See the discussion on static vs. dynamic paths in (Reiter and Rubin 1998).

5.3. *Hamming weight attacks against S-boxes*

Timing (Kocher 1996) and power analysis (Kocher *et al.* 1999) are two flavours of side-channel *correlation attacks* against cryptographic devices (Brier *et al.* 2004; Standaert *et al.* 2009). These attacks presuppose, explicitly or implicitly, that attacker knows the inputs (messages) processed by the target device[¶]. Basically, the attack is carried out by simulating the device's computations under the different candidate keys, each time using as inputs the same messages processed by the device. This way, the attacker obtains different samplings of the leakage from the side-channel, one for each candidate key. He will then choose the key that generates the sampling that is most correlated with the one obtained from the device.

Here, we wonder to what extent knowledge of the messages is necessary to extract significant amount of information from the side-channel. Differently from correlation attacks, we will therefore assume that input messages have a nonzero, moderate redundancy, but not that they are known to the attacker. We analyze the case of DES S-boxes. Similar analyses could be conducted against different types of symmetric keys devices. Our analysis applies to any round, in fact, to any context where an adversary may get to observe the Hamming weight of the S-box output. A DES S-box can be described as a function that takes as an input a pair of a message and a key and yields as an output a block of ciphertext, $SB : \mathcal{K} \times \mathcal{M} \to C$, where: $\mathcal{K} = \{0, 1\}^6$ is the set of keys, $\mathcal{M} = \{0, 1\}^6$ is the set of messages and $C = \{0, 1\}^4$ is the set of ciphertexts. The internal details of the device are unimportant for the purpose of this illustration. We assume a uniform prior distribution on $\mathcal{K}$ and some known prior distribution on $\mathcal{M}$, say p_M. Similarly to (Kelsey *et al.* 2000), we assume the attacker can create a side-channel delivering him the Hamming weight of the target S-box' output. To the S-box thus described there corresponds an information hiding

[¶] In some circumstances, this knowledge is granted by the application. For example, in an attack against the final round of any Feistel cipher, the left half of the output is also the input of the target round function (see (Kelsey *et al.* 2000)).

system where: $\mathcal{S} = \mathcal{K}$, $\mathcal{O} = \{0, 1, 2, 3, 4\}$ is the set of observables, i.e. the Hamming weights, and $p(o|k)$ is defined as

$$p(o|k) \stackrel{\triangle}{=} \textstyle\sum_{m \in \mathcal{M}: W(SB(m,k))=o} p_M(m)$$

where $W(\cdot)$ is the Hamming weight function.

We report here on our results about the first of the eighth S-boxes of DES. Analysis of other S-boxes leads to similar conclusions. The distribution of the plaintext, p_M, plays a crucial role here: the lower the redundancy, the less information is expected to be extracted from the side-channel. For example, if p_M is the uniform distribution (0% redundancy), then it is easy to see that all the rows of the matrix $p(o|k)$ are the same, hence $P_e(n) = 1 - 1/64$ for each n: the adversary cannot do any better than random guessing. For our analysis, we have chosen a plaintext with a redundancy of about 27% ($H(p_M) = 4.39$ bits), obtained by sampling ASCII text from some web pages. In the resulting matrix, $p(o|k)$, all the rows are different, which implies that $P_e(n) \to 0$. Concerning the rate of convergence, the method of norm-1 difference (Proposition 1) yields $\epsilon \approx 4.0822 \times 10^{-4}$. This means that with $n \geq 1.7 \times 10^5$ observations the error probability is < 0.045 (here and in what follows, we use the refined bound given in Remark 1). Discarding the keys corresponding to the seven shortest norm-1 distances, one would get $\epsilon \approx 1.2179 \times 10^{-3}$. Applying Proposition 3, one gets an error probability ≤ 0.11 already with $n = 6 \times 10^4$ observations.

In a more realistic scenario, the attacker could not directly measure the Hamming weight of the target S-box, but rather the global weight of the eight S-boxes composing the round function of DES. This scenario can be modeled as a noisy version of the previous one. The Hamming weight of the target S-box, O, is now disturbed by the noise N, the sum of the Hamming weights of the remaining seven S-boxes, say $W_2, ..., W_8$. Assuming that the variables W_i are independent from each other and from O and identically distributed – this is not strictly true, but seems a reasonable approximation – the central limit theorem would tell us that their sum $N = \sum_{i=2}^{8} W_i$ has approximately a normal distribution. Here, for simplicity we model N as a random variable having binomial distribution $B(n, p)$ with $n = 28$ and $p = \frac{1}{2}$. What is observed by the attacker now is $O' \stackrel{\triangle}{=} O + N$. Hence the new set of observables is $\mathcal{O}' = \{0, ..., 32\}$. Explicitly, for each $i \in \mathcal{O}'$ and $k \in \mathcal{K}$, the entries of the new conditional probability matrix $p'(\cdot|\cdot)$ are given by

$$p'(i|k) \stackrel{\triangle}{=} \Pr(O + N = i \,|\, K = k) \;=\; \sum_{j=0}^{\min\{i,4\}} p(j|k) \cdot \binom{28}{i-j} \cdot 2^{-28}\,.$$

Proposition 1 applied to the matrix $p'(\cdot|\cdot)$ yields a rate of $\epsilon \approx 1.9275 \times 10^{-6}$. Theorem 1 gives $P_e(n) < 0.0007$ for $n \geq 4.2 \times 10^8$. As expected, the convergence rate is lower than in the noiseless case. However, the effort needed to break the system is certainly in the reach of a well determined attacker.

Our simple analysis confirms that unprotected implementations of DES S-boxes are quite vulnerable to attacks based on Hamming weights. Software simulations have reinforced this conclusion, showing that, in practice, a good success probability for the adversary is achieved with a relatively small n. For instance, in the noiseless case, already with $n = 10^3$, we have obtained an experimental success rate of 99%.

6. Sequential observations and hidden Markov models

The attack model discussed in the preceding sections presupposes that the computation involving the secret information takes place in a single step. Or, more accurately, that the intermediate states of the computation are not accessible to the attacker. Here, we consider a more refined scenario, where computations may take several steps to terminate, or even not terminate at all. In any case, to each state of the computation there corresponds one observation on the part of the attacker. Hence, to each computation there corresponds a sequential *trace* of observations. The attacker may collect multiple such traces, corresponding to multiple independent executions of the system. Throughout these executions, the secret information is kept fixed. This set up is well suited to describe situations where the attacker collects information from different sources at different times, like in a coalition of different local eavesdroppers. An instance of this situation in the context of an anonymous routing application will be examined later on.

Discrete-time Hidden Markov Models (Rabiner 1989) provide a convenient setting to formally model such systems, which we may designate as *sequential* information hiding system.

6.1. *Definitions*

Let $\mathcal{S}$ and $\mathcal{O}$ be finite sets of states and observations, respectively. A (discrete-time, homogeneous) *Hidden Markov Model (*HMM*)* with states in $\mathcal{S}$ and observations in $\mathcal{O}$ is a a pair of random processes $\langle (S_i)_{i\geq 1}, (O_i)_{i\geq 1} \rangle$, such that, for each $t \geq 1$

— S_t and O_t are random variables taking values in $\mathcal{S}$ and $\mathcal{O}$, respectively; and,

— the following equalities hold true (whenever the involved conditional probabilities are defined):

$$\Pr(S_{t+1} = s_{t+1} | S_t = s_t, O_t = o_t, ..., S_1 = s_1, O_1 = o_1) = \Pr(S_{t+1} = s_{t+1} | S_t = s_t) \quad (4)$$

$$\Pr(O_t = o_t | S_t = s_t, S_{t-1} = s_{t-1}, O_{t-1} = o_{t-1}, ..., S_1 = s_1, O_1 = o_1) = \Pr(O_t = o_t | S_t = s_t) \quad (5)$$

Moreover, the value of the above probabilities does not depend on the index t, but only on s_t, s_{t+1} and o_t.

Equation (4) says that the state at time $t + 1$ only depends on the state at time t, that is $(S_i)_{i\geq 1}$ forms a Markov chain. Equation (5) says that the observation at time t only depends on the state at time t. A consequence of this equation is that the state at time $t + 1$ is independent from the observation at time t, given the state at time t, that is

$$\Pr(O_t = o_t, S_{t+1} = s_{t+1} | S_t = s_t) = \Pr(O_t = o_t | S_t = s_t) \cdot \Pr(S_{t+1} = s_{t+1} | S_t = s_t). \quad (6)$$

Graphically, a HMM can be represented by a diagram like the one below, where the nodes are random variables and the presence of a pair of arrows $X \leftarrow Y \rightarrow Z$ or $X \rightarrow Y \rightarrow Z$ signifies conditional independence of X and Z given Y.

$$\begin{tikzcd}
S_1 \arrow[r] \arrow[d] & S_2 \arrow[r] \arrow[d] & S_3 \arrow[r] \arrow[d] & \cdots \\
O_1 & O_2 & O_3 & \cdots
\end{tikzcd}$$

Assume now $\mathcal{S} = \{s_1, ..., s_m\}$ and $\mathcal{O} = \{o_1, ..., o_l\}$. A finite-state HMM on $\mathcal{S}$ and $\mathcal{O}$ is completely specified by, hence can be identified with, a triple (π, F, G) such that:

— $\pi \in \mathbb{R}^{1\times m}$ is a row-vector representing the prior distribution on $\mathcal{S}$, that is $\pi(i) = p(S_1 = s_i)$ for each $1 \leq i \leq m$;

— $F \in \mathbb{R}^{m\times m}$ is a matrix such that $F(i, j)$ is the probability of transition from s_i to s_j, for $1 \leq i, j \leq m$;

— $G \in \mathbb{R}^{m\times l}$ is a matrix such that $G(i, j)$ is the probability of observing o_j at state s_i, for $1 \leq i \leq m$ and $1 \leq j \leq l$.

In our scenario, a Bayesian attacker targets the first state of the computation, that is the value of S_1. We are interested in analyzing the attacker's probability of error after observing n traces of length t, corresponding to n conditionally independent executions of the system up to and including time t, as both n and t go to $+\infty$. This we define in the following. Let σ range over the set of observation traces, that is $\mathcal{O}^*$. For any $\sigma = o_1 \cdots o_t$ ($t \geq 0$) and $s \in S$, define[||]

$$p(\sigma \,|\, s) \overset{\triangle}{=} \Pr(O_1 = o_1, O_2 = o_2, ..., O_t = o_t \,|\, S_1 = s)$$

with the proviso that $p(\epsilon \,|\, s) \overset{\triangle}{=} 1$. We note that for any fixed $t \geq 0$ and $s \in \mathcal{S}$, $p(\sigma|s)$ defines a probability distribution as σ ranges over $\mathcal{O}^t$, the set of traces of length t, or *t-traces*. In other words, for *any fixed t*, we have an information hiding system in the sense of Section 3, with $\mathcal{S}$ as a set of states, $\mathcal{O}^t$ as a set of observables, a conditional probability matrix $p(\sigma|s)$ ($s \in \mathcal{S}, \sigma \in \mathcal{O}^t$) and π as a prior distribution on states. Call $\mathcal{H}^{(t)}$ this system. We have the following error probabilities of interest ($t \geq 0$):

$$\begin{aligned} P_e^{(t)}(n) &\overset{\triangle}{=} \text{probability of error after } n \text{ observations (of } t\text{-traces) in } \mathcal{H}^{(t)} && (7)\\ P_e^{(t)} &\overset{\triangle}{=} \lim_{n\to\infty} P_e^{(t)}(n) && (8)\\ P_e &\overset{\triangle}{=} \lim_{t\to\infty} P_e^{(t)}. && (9) \end{aligned}$$

We will show in the next subsection that the above two limits exist and are easy to compute. Correspondingly, we have the information leakage quantities of interest (here $P_{succ} = 1 - P_e$):

$$L_+^{(t)}(n) \overset{\triangle}{=} P_{succ}^{(t)}(n) - \max_s \pi(s) \quad L_+^{(t)} \overset{\triangle}{=} P_{succ}^{(t)} - \max_s \pi(s) \quad L_+ \overset{\triangle}{=} P_{succ} - \max_s \pi(s)\,.$$

Multiplicative leakages are defined similarly.

6.2. *Results*

That the limit (8) exists is an immediate consequence of Theorem 1 applied to $\mathcal{H}^{(t)}$. Indeed, let us denote by $\equiv^{(t)}$ the indistinguishability relation on states for $\mathcal{H}^{(t)}$, that is, explicitly

$$s \equiv^{(t)} s' \text{ iff for each } \sigma \in \mathcal{O}^t : \; p(\sigma|s) = p(\sigma|s')\,.$$

Let $C_1^{(t)}, ..., C_{K_t}^{(t)}$ be the equivalence classes of $\equiv^{(t)}$ and let $p_i^{*(t)} \overset{\triangle}{=} \max_{s\in C_i^{(t)}} \pi(s)$. Then we have by Theorem 1 that

$$P_e^{(t)} = 1 - \sum_{i=1}^{K_t} p_i^{*(t)}\,. \tag{10}$$

[||] Or, more formally, $p(\sigma \,|\, s) \overset{\triangle}{=} \Pr(O_h = o_1, O_{h+1} = o_2, ..., O_{h+t-1} = o_t \,|\, S_h = s)$, for any index h s.t. $\Pr(S_h = s) > 0$. Note that this definition does not depend on the chosen index h, given that the chain is homogeneous. Also, we are assuming w.l.o.g. here that for each s there is an index h s.t. $\Pr(S_h = s) > 0$.

Note that, for any fixed t, Corollary 2 carries over to $\mathcal{H}^{(t)}$. We now consider the case $t \to \infty$. We introduce the following fundamental relation.

Definition 6 (Indistinguishability for HMM). The *indistinguishability relation* on a HMM is defined as

$$\equiv \ \stackrel{\triangle}{=} \ \bigcap_{t\geq 0} \equiv^{(t)} \ .$$

Equivalently, $s \equiv s'$ iff for every $\sigma \in \mathcal{O}^*$, $p(\sigma|s) = p(\sigma|s')$.

It is immediate to check that $\equiv$ is an equivalence relation. Let $C_1, ..., C_K$ be its equivalence classes and let $p_i^* \stackrel{\triangle}{=} \max_{s\in C_i} \pi(s)$, for $i = 1, ..., K$.

Proposition 4. The limit (9) is given by $P_e \ = \ 1 - \sum_{i=1}^{K} p_i^*$.

Proof. First, we note that $\{\equiv^{(t)}\}_{t\geq 0}$ forms a monotonically non-increasing chain of relations: $\equiv^{(0)} \supseteq \equiv^{(1)} \supseteq \equiv^{(2)} \ \cdots$. To prove this fact, note that, for each t, $\sigma \in \mathcal{O}^t$ and $s \in \mathcal{S}$, $p(\sigma|s) = \sum_{o\in\mathcal{O}} p(\sigma \cdot o|s)$. Therefore, $s \equiv^{(t+1)} s'$ implies $s \equiv^{(t)} s'$.

The above fact implies that the sequence $\{P_e^{(t)}\}_{t\geq 0}$ is monotonically non-increasing: indeed, the finer the equivalence classes of $\equiv^{(t)}$, the greater the value of the sum in (10). Therefore, the limit (9) exists. In order to determine the value of this limit, we reason as follows. Since $\mathcal{S}$ is finite and the chain of sets $\{\equiv^{(t)}\}_{t\geq 0}$ is monotonically non-increasing, there must exist t_0 such that

$$\equiv^{(t_0)} \ = \ \equiv^{(t_0+1)} \ = \ \cdots \ = \equiv \ .$$

According to (10) then, from t_0 onward the sequence $\{P_e^{(t)}\}_{t\geq 0}$ stabilizes to the value $P_e \ = 1 - \sum_{i=1}^{K} p_i^*$. □

The actual computation of P_e, and of the corresponding information leakage quantities, is therefore reduced to the computation of $\equiv$. Below, we show that this computation can indeed be performed quite efficiently. We do so by using some elementary linear algebra. Let us introduce some additional notation. We define the transition matrices $M_{o_k} \in \mathbb{R}^{m\times m}$, for any $o_k \in \mathcal{O}$, as follows[††]:

$$\begin{aligned} M_{o_k}(i,j) &\stackrel{\triangle}{=} \Pr(S_{t+1} = s_j, O_t = o_k | S_t = s_i) \\ &= F(i,j) \cdot G(i,k) \end{aligned}$$

where the last equality is justified by equation (6). For any $\sigma = o_1 \cdots o_t$, we let M_σ denote $M_{o_1} \times \cdots \times M_{o_t}$. Finally, we let $e_i \in \mathbb{R}^{1\times m}$ denote the row vector with 1 in the i-th position and 0 elsewhere and let $e \stackrel{\triangle}{=} \sum_{i=1}^{m} e_i$ denote the everywhere 1 vector. The following lemma provides an alternative characterization of $\equiv$; the lemma is easily proven by induction on the length of σ.

Lemma 4. For each σ and $s_i \in \mathcal{S}$, $p(\sigma \,|\, s_i) = e_i M_\sigma e^T$. Hence $s_i \equiv s_j$ iff for each $\sigma \in \mathcal{O}^*$, $e_i M_\sigma e^T = e_j M_\sigma e^T$.

We say a row vector v is *orthogonal* to a set of column vectors U, written $v \perp U$, if $vu = 0$ for each $u \in U$. Also, for any set of vectors U, $U^\perp$ denotes the orthogonal complement of U given by $U^\perp = \{v \,|\, v \perp U\}$. It is easily seen that $U^\perp$ is a sub-space of the space of column vectors.

[††] Again, due to homogeneity, in the definition we can choose any index t such that $\Pr(S_t = s_i) > 0$.

Moreover, $U \subseteq V$ implies $V^{\perp} \subseteq U^{\perp}$. Of course, the above definitions extend as expected when exchanging the roles of "row" and "column". We finally note that if U is a vector space, then $(\cdot)^{\perp}$ is an involution, that is $(U^{\perp})^{\perp} = U$.

Theorem 2. Let B be a basis of the (finite-dimensional) sub-space of $\mathbb{R}^{m\times 1}$ spanned by $\bigcup_{\sigma\in\mathcal{O}^*}\{M_\sigma e^T\}$. For $s_i, s_j \in \mathcal{S}$, $s_i \equiv s_j$ iff $(e_i - e_j) \perp B$.

Proof. The condition of Lemma 4 can be expressed as

$$\text{for each } \sigma \in \mathcal{O}^* : \ (e_i - e_j)M_\sigma e^T = 0$$

$$\text{iff}$$

$$(e_i - e_j) \in \cap_\sigma\{M_\sigma e^T\}^{\perp} = (\cup_\sigma\{M_\sigma e^T\})^{\perp}$$

$$\text{iff}$$

$$(e_i - e_j) \perp B\,.$$

□

A basis B of span($\bigcup_\sigma\{M_\sigma e^T\}$) can be expressed as

$$B = \{M_\sigma e^T \,|\, \sigma \in \mathcal{F}\} \tag{11}$$

for a suitable finite, prefix-closed $\mathcal{F} \subseteq \mathcal{O}^*$. More precisely, B can be computed by a procedure that starts with the set $B := \{e^T\}$ and iteratively updates B by joining in the vectors $M_{o\cdot\sigma}e^T = M_o \cdot (M_\sigma e^T)$, with $M_\sigma e^T \in B$ and $o \in \mathcal{O}$, that are linearly independent from the vectors already present in B, until no other vector can be joined in. This procedure must terminate in a number of steps $\leq m$. The set of strings $\mathcal{F}$ can be computed alongside with B.

We now briefly discuss the rate of convergence to P_e. We have already seen that $P_e^{(t_0)} = P_e$. Therefore, there is no advantage, for an attacker wanting to determine $\equiv$, in considering traces of length greater than t_0. The convergence rate for the attacker is hence determined by the matrix of the system $\mathcal{H}^{(t_0)}$. For this reason, it is of practical importance to be able to compute t_0. This is in fact quite an easy task, as stated by the following proposition.

Proposition 5. Let B be a basis of the space spanned by $\cup_\sigma\{M_\sigma e^T\}$ and $\mathcal{F}$ the corresponding set of strings, as specified by (11). Assume B and $\mathcal{F}$ have been obtained by the algorithm described above. Then $t_0 = \max\{|\sigma| \,:\, \sigma \in \mathcal{F}\}$.

Proof. For any equivalence relation R over $\mathcal{S}$, let the *kernel* of R be the subspace of $\mathbb{R}^{1\times m}$ defined thus

$$\ker(R) \triangleq \text{span}(\{e_i - e_j \,|\, s_i\, R\, s_j\})\,.$$

Now, by a reasoning similar to that in the proof of Theorem 2, for any t we have

$$\ker(\equiv^{(t)}) \ = \ (\text{span}(\cup_{\sigma\in\mathcal{O}^t} M_\sigma e^T))^{\perp} \tag{12}$$

while, by definition of B and $\mathcal{F}$

$$\ker(\equiv) \ = \ (\text{span}(\cup_{\sigma\in\mathcal{F}} M_\sigma e^T))^{\perp}\,. \tag{13}$$

Let R, R' be two equivalence relations of the form $\equiv$ or $\equiv^{(t)}$. The above equations imply that $s_i R s_j$ iff $e_i - e_j \in \ker(R)$. Moreover, $R \subseteq R'$ iff $\ker(R) \subseteq \ker(R')$. Thus, the equivalence relations of interest are completely characterized by their kernels. By Lemma 4, we deduce that for each t, $\ker(\equiv^{(t)}) \supseteq \ker(\equiv^{(t+1)})$. From this fact, and using the fact that $U \subseteq V$ implies $V^\perp \subseteq U^\perp$, and that $(U^\perp)^\perp = U$, we obtain that for each t, $\mathrm{span}(\cup_{\sigma \in O^t} M_\sigma e^T) \subseteq \mathrm{span}(\cup_{\sigma \in O^{t+1}} M_\sigma e^T)$, hence

$$\ker(\equiv^{(t)})^\perp \;=\; \mathrm{span}(\cup_{\sigma \in O^t} M_\sigma e^T) \;=\; \mathrm{span}(\cup_{0 \le i \le t} \cup_{\sigma \in O^i} M_\sigma e^T)\,.$$

Take now $t = \max\{|\sigma| \;:\; \sigma \in \mathcal{F}\}$ in the equation above: we obtain

$$\ker(\equiv^{(t)})^\perp \;=\; \mathrm{span}(\cup_{0 \le i \le t} \cup_{\sigma \in O^i} M_\sigma e^T) \supseteq \mathrm{span}(\cup_{\sigma \in \mathcal{F}} M_\sigma e^T) \;=\; \ker(\equiv)^\perp$$

hence $\ker(\equiv^{(t)}) \subseteq \ker(\equiv)$, which implies $\ker(\equiv^{(t)}) \;=\; \ker(\equiv)$, that is $\equiv^{(t)} = \equiv$.

On the other hand, take any $t < \max\{|\sigma| \;:\; \sigma \in \mathcal{F}\}$. Assume by contradiction that $\equiv^{(t)} = \equiv$, that is $\ker(\equiv^{(t)}) = \ker(\equiv)$. By (12) and (13), and using $(U^\perp)^\perp = U$, we obtain that $\mathrm{span}(\cup_{\sigma \in O^t} M_\sigma e^T) = \mathrm{span}(\cup_{\sigma \in \mathcal{F}} M_\sigma e^T)$. This implies that there is a string of maximal length in $\mathcal{F}$, say σ_0, s.t. $M_{\sigma_0} e^T$ can be obtained as a linear combination of vectors $M_\sigma e^T$, for σ of length $t < |\sigma_0|$. But, by construction of B and $\mathcal{F}$, this cannot be the case. □

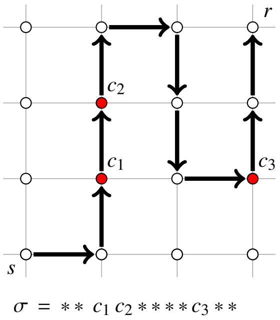
The practical computation of the rate relative to P_e can be carried out applying Proposition 1 to the system $\mathcal{H}^{(t_0)}$, which requires one has at hand the conditional probability matrix of the system. The entries of this matrix are of the form $p(\sigma|s)$ with $\sigma \in O^{t_0}$. The computation of individual entries $p(\sigma|s)$ can be performed quite efficiently, running the so-called *Forward-Backward algorithm* on the underlying HMM (see (Rabiner 1989)). Unfortunately, the number of columns in the matrix, i.e. of traces of length t_0, is $|O|^{t_0}$. Most likely, this makes the exact computation of the rate impractical for significant systems (say, systems with thousands of states). Forms of approximations are conceivable to tackle this problem, such as "lumping" the matrix by aggregating sets of columns, which leads to tractable dimensions, while underestimating the rate. We will not discuss this issue here, though.

7. An example: hiding routing information

We discuss a scenario where messages are routed from a sender to a receiver in a network with a fixed topology, as can be found for instance in a structured peer-to-peer overlay. Anonymity protocols such as *onion routing* (Goldschlag *et al.* 1998) are designed to protect the identity of the sender and/or of the receiver in the presence of corrupted nodes. Initially, the routing path from the sender to the receiver is established randomly. In each exchanged message, nested layers of encryptions ensure that any intermediate on the path node only gets to know the preceding and the next node in the path, but not the identity of the original sender and of the final receiver.

We present and analyze a model of this protocol below. We should warn the reader that, for the sake of presentation, we have chosen to analyze an over-simplified version of the protocol. For example, we assume that, upon receiving a message, a corrupted node can tell whether the message pertains to the target sender-receiver conversation, but cannot identify the predecessor node in the path followed by the message. More powerful forms of eavesdropping can be easily accommodated. Again, we are interested in the case of re-execution, where, for some reason, the initiator is forced to establish new paths with the responder several times. We will concentrate on the asymptotic error probability and leakage, ignoring issues related to the rate of convergence.

We assume the topology of the network is specified by a nonempty graph $\mathcal{G} = (V, E)$. For each node $v \in V$, we let $N(v)$ denote the set of neighbours of v, that is the set of nodes u for which an arc $\{v, u\}$ in E exists; $N(v)$ is always assumed nonempty. Let $C \subseteq V$ represent the subset of corrupted nodes. We let $\mathcal{S} \triangleq V \times V$ be the set of states of the system and $\mathcal{O} \triangleq C \cup \{*\}$ be the set of observables; here $c \in C$ means that the message is presently hold by the node c, while $*$ means no observation other than the elapse of a discrete time interval. What the attacker can observe are therefore traces like σ in the picture on the right.



We assume the sender and the receiver are chosen at random independently from each other, and that the sender is always a honest node (there is no point for the attacker in eavesdropping on corrupted nodes). This formally means that the first state of the Markov chain is a random vector $S_1 = (S, R)$, where S and R are independent random variables taking values uniformly in $V \setminus C$ and V, respectively. The transitions and the observations of the hidden Markov model are defined by the following equations, where $u, v, r \in V$, $c \in C$ and $s \in V \setminus C$:

$$p(\,(u,r)\,|\,(v,r)\,) \;\triangleq\; \begin{cases} \frac{1}{|N(v)|} & \text{if } u \in N(v) \text{ and } v \neq r \\ 0 & \text{if } u \notin N(v) \text{ and } v \neq r \end{cases} \qquad p(\,(r,r)\,|\,(r,r)\,) \;\triangleq\; 1$$

$$p(\,c\,|\,(c,r)\,) \;\triangleq\; 1 \qquad\qquad p(\,*\,|\,(s,r)\,) \;\triangleq\; 1\,.$$

The above equations define a hidden Markov model, say $\mathcal{M}$. For any specific topology $\mathcal{G}$, it is easy to compute the corresponding probability P_e defined by (9), as indicated by Proposition 4. Recall that P_e is the probability that, after observing n independent executions of the system up to time t, for $n, t \to \infty$, the attacker fails to correctly guess the *pair* (s, r) of the true sender and receiver.

In fact, in order to asses the degree of anonymity provided by the system, it is more convenient to have at hand the error probabilities for the senders and for the receivers separately. To see how these probabilities are defined and computed, we examine in detail the case of the sender; the receiver case is basically the same. Formally, for each $\sigma \in \mathcal{O}^t$ and sender $s \in V \setminus C$, let

$$p_{send}(\sigma|s) \triangleq \Pr(O^t = \sigma | S = s)\,.$$

For any fixed t, this equation defines a conditional probability matrix, hence an information hiding system where the states are the senders, i.e. $V \backslash C$: let us denote by $P^{(t)}_{e,send}$ the corresponding asymptotic error probability. The probability we are after is obtained by letting t tend to ∞:

$$P_{e,send} \triangleq \lim_{t\to\infty} P^{(t)}_{e,send}\,.$$

Reasoning as we did for Proposition 4, one checks that $P_{e,send}$ can be computed from the limit indistinguishability relation as $t \to \infty$, say $\equiv_{send}$. Explicitly, this relation can be defined as $s \equiv_{send} s'$ iff for each $\sigma \in \mathcal{O}^*$, $p_{send}(\sigma|s) = p_{send}(\sigma|s')$. The next lemma says how $\equiv_{send}$ can be computed starting from the hidden Markov model $\mathcal{M}$ defined above, by a suitable aggregation of the rows of the basis matrix B. The proof consists of easy manipulations of the transition matrices M_σ and is omitted. Recall that the states of $\mathcal{M}$ are pairs (u, v), thus $e_{(u,v)}$ denotes the row vector in $\mathbb{R}^{1\times|V|^2}$

whose entry corresponding to the element (u, v) is 1, while the others are 0. For each s, we let f_s denote the row vector $\sum_{(s,v)\in\mathcal{S}} e_{(s,v)}$.

Lemma 5. Let B a basis like in the hypotheses of Theorem 2 for the hidden Markov model $\mathcal{M}$ defined above. For any two senders s and s', $s \equiv_{send} s'$ iff $(f_s - f_{s'}) \perp B$.

We have applied this setting to a few instances of a grid network, like the one displayed above, relative to different sizes d of the grid and different sets C of corrupted nodes. Table 1summarizes the outcomes of these experiments. The nodes in the grid are numbered from 1 to d^2, starting from the top left corner and proceeding row-wise from left to right. To avoid end effects, we make the grid wrap up, i.e. the top and bottom rows are connected together, as well as the rightmost and leftmost columns. The sets C are chosen so as to give rise to configurations where no two corrupted nodes are directly connected: we have checked experimentally that these are the most advantageous for the attacker; otherwise, the relative distance of the corrupted nodes seems unimportant. K_{send} and K_{rec} denote the number of classes of $\equiv_{send}$ and of $\equiv_{rec}$, respectively. Moreover, from Corollary 2(2) in Section 4, we know that the asymptotic multiplicative leakage coincides with the number of classes in the case of uniform distribution. The probability $P_{e,send}$ is computed as $1 - \frac{K_{send}}{|V|-|C|}$, while $P_{e,rec}$ is computed as $1 - \frac{K_{rec}}{|V|}$. Finally, additive leakages are computed as indicated by Corollary 2(2).

Although a systematic study of anonymous routing protocols is outside the scope of the present paper, some qualitative considerations can be drawn from these data. If one keeps d fixed and lets $|C|$ grow, the data are simple to interpret: the error probability goes to 0 and the leakage gets larger. On the other hand, if one keeps $|C|$ fixed and compares configurations of different size d, the interpretation becomes less obvious. The leakage tends to increase when moving from smaller to larger values of d, which is particularly evident from the columns of multiplicative leakage. This increase occurs barely because, as the number of nodes grows, the number of indistinguishability classes tends to grow as well: all this means is just that a large system tends to leak more information than a small one. Concerning error probability, which is supposed to measure the "absolute" resistance of a system, the data seem to partially contradict the intuition that the more nodes in a network, the stronger the guarantee of anonymity. Indeed, it may happen that the error probability *decreases* when moving from smaller to larger values of d. Also, the receiver seems more vulnerable than the sender from the point of view of anonymity. At the moment we have no explanation to offer for these phenomena[‡‡]. Their full understanding certainly requires a more systematic study of the anonymous routing protocols.

8. Conclusion and further work

We have characterized the asymptotic behaviour of error probability, and information leakage in terms of indistinguishability in a scenario of one-try attacks after repeated independent, noisy observations. We have first examined the case in which each execution gives rise to a single

[‡‡] Heuristically, the first phenomenon seems to be connected with the fact that, as d grows, the number of indistinguishability classes may grow faster than the number of nodes, because more and more observables (traces) become available.

d	C	K_{send} $= L_{\times,send}$	K_{rec} $= L_{\times,rec}$	$P_{e,send}$	$P_{e,rec}$	$L_{+,send}$	$L_{+,rec}$
3	$\{1\}$	2	4	0.75	0.56	0.12	0.33
3	$\{1, 5\}$	4	9	0.43	0	0.43	0.89
3	$\{2, 4, 6, 8\}$	5	9	0	0	0.8	0.89
4	$\{1\}$	4	9	0.73	0.44	0.2	0.5
4	$\{1, 6\}$	7	12	0.5	0.25	0.43	0.69
4	$\{2, 5, 7, 10\}$	12	16	0	0	0.92	0.94
5	$\{1\}$	5	15	0.79	0.4	0.17	0.56
5	$\{1, 7\}$	13	25	0.43	0	0.52	0.96
5	$\{2, 6, 8, 12\}$	21	25	0	0	0.95	0.96
6	$\{1\}$	10	10	0.71	0.72	0.26	0.25
6	$\{1, 8\}$	19	36	0.44	0	0.53	0.97
6	$\{2, 7, 9, 14\}$	32	36	0	0	0.97	0.97

Table 1. *Sender and receiver anonymity for several instances of a grid network.*

observation, then extended our results to the case where each state traversed during an execution induces one observation.

In the future, we would like to systematically characterize achievable rates of convergence given an error probability threshold, thus generalizing Proposition 3. It would also be natural to generalize the present one-try scenario to the case of k-tries attack, for $k \geq 2$. Experiments and simulations with realistic anonymity protocols may be useful to asses at a practical level the theoretical results of our study. For example, we believe that HMM's are relevant to security in peer-to-peer overlays. We would also like to clarify the relationship of our model with the notion of probabilistic *opacity* (Bérard *et al.* 2010), and with the huge amount of work existing on *covert channels* (see e.g. (Mantel and Sudbrock 2009) and references therein).

Another interesting research direction concerns the nature of the guarantees provided by error probability related metrics. These quantities provide a synthetic way to express the security of a system under a specific attack scenario. However, they are tightly connected mainly with the number of indistinguishability classes, which may be inadequate in some cases. For instance, in an anonymity protocol characterized by a high error probability (or small leakage), it might well be the case that an individual user belongs to a singleton class, hence being totally exposed to eavesdropping. For this reason, one would like to devise a framework where the analysis can be conducted both at a quantitative level (*how much* is leaked) and at a qualitative one (*what* is leaked).

Acknowledgments

The authors should like to thank Alessandro Celestini for a careful reading of the manuscript. Three anonymous FOSSACS 2011 reviewers provided valuable comments.

References

M. Backes, B. Köpf. Formally Bounding the Side-Channel Leakage in Unknown-Message Attacks. *ESORICS 2008* , *LNCS* 5283: 517-532, 2008.

T. Baignères, S. Vaudenay. The Complexity of Distinguishing Distributions (Invited Talk). *ICITS 2008 LNCS* 5155: 210-222, 2008.

B. Bérard, J. Mullins, M. Sassolas. Quantifying Opacity. *Proc. of QEST 2010*: 263-272, IEEE Society, 2010.

M. Boreale. Quantifying information leakage in process calculi. *Information and Computation* 207(6): 699-725, 2009.

M. Boreale, F. Pampaloni, M. Paolini. Asymptotic information leakage under one-try attacks. *Proc. of FoSSaCS 2011*, LNCS. To appear.

C. Braun, K. Chatzikokolakis, C. Palamidessi. Compositional Methods for Information-Hiding. *Proc. of FoSSaCS 2008*, *LNCS* 4962: 443-457, 2008.

C. Braun, K. Chatzikokolakis, C. Palamidessi. Quantitative Notions of Leakage for One-try Attacks. *Proc. of MFPS 2009*, *Electr. Notes Theor. Comput. Sci.* 249: 75-91, 2009.

E. Brier, C. Clavier, F. Olivier. Correlation Power Analysis with a Leakage Model. *Proc. of CHES 2004*, *LNCS* 3156: 16-29, 2004.

K. Chatzikokolakis, C. Palamidessi, P. Panangaden. Anonymity protocols as noisy channels. *Information and Computation* 206(2-4): 378-401, 2008.

K. Chatzikokolakis, C. Palamidessi, P. Panangaden. On the Bayes risk in information-hiding protocols. *Journal of Computer Security* 16(5): 531-571, 2008.

D. Clark, S. Hunt, P. Malacaria. Quantitative Analysis of the Leakage of Confidential Data. *Electr. Notes Theor. Comput. Sci.* 59(3), 2001.

T. M. Cover, J. A. Thomas. *Elements of Information Theory,2/e*. John Wiley & Sons, 2006.

D. M. Goldschlag, M. G. Reed, P. F. Syverson. Anonymous Connections and Onion Routing. *IEEE Journal on Selected Areas in Communication*, *Special Issue on Copyright and Privacy Protection*, 1998.

J. Kelsey, B. Schneier, D. Wagner, C. Hall. Side Channel Cryptanalysis of Product Ciphers. *Journal of Computer Security* 8(2/3): 2000.

P. C. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. *CRYPTO 1996, LNCS* 1109:104-113, 1996.

P. C. Kocher, J. Jaffe, B. Jun. Differential Power Analysis. *CRYPTO 1999, LNCS* 1666:388-397, 1999.

B. Köpf, D. A. Basin. An information-theoretic model for adaptive side-channel attacks. *ACM Conference on Computer and Communications Security* 2007: 286-296, 2007.

B. Köpf, M, Dürmuth. A Provably Secure and Efficient Countermeasure against Timing Attacks. *CSF 2009*: 324-335, 2009.

B. Köpf, G. Smith. Vulnerability Bounds and Leakage Resilience of Blinded Cryptography under Timing Attacks. *CSF 2010*: 44-56, 2010.

C.C. Leang, D.H. Johnson. On the asymptotics of M-hypothesis Bayesian detection. *IEEE Transactions on Information Theory* 43: 280-282, 1997.

H. Mantel, H. Sudbrock. Information-Theoretic Modeling and Analysis of Interrupt-Related Covert Channels. In *Proc. of the Workshop on Formal Aspects in Security and Trust, FAST 2008*. LNCS 5491, 2009.

J. L. Massey. Guessing and Entropy. In *Proc. 1994 IEEE Symposium on Information Theory (ISIT 1994)*: 204, 1994.

L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proc. of the IEEE* 77 (2): 257-286, 1989.

M. K. Reiter, A. D. Rubin. Crowds: Anonymity for Web Transactions. *ACM Trans. Inf. Syst. Secur.* 1(1): 66-92, 1998.

G. Smith. On the Foundations of Quantitative Information Flow. *FoSSaCS 2009*, *LNCS* 5504: 288-302, 2009.

F.-X. Standaert, T.G. Malkin, M. Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. *Proc. of Eurocrypt 2009*, *LNCS* 5479: 443-461, 2009.