Quantitative Information Flow, with a View*

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

[1] Università di Firenze, Italy
[2] IMT - Institute for Advanced Studies, Lucca, Italy

Abstract. We put forward a general model intended for assessment of system security against passive eavesdroppers, both quantitatively (*how much* information is leaked) and qualitatively (*what* properties are leaked). To this purpose, we extend information hiding systems (IHS), a model where the secret-observable relation is represented as a noisy channel, with *views*: basically, partitions of the state-space. Given a view W and n independent observations of the system, one is interested in the probability that a Bayesian adversary wrongly predicts the class of W the underlying secret belongs to. We offer results that allow one to easily characterise the behaviour of this error probability as a function of the number of observations, in terms of the channel matrices defining the IHS and the view W. In particular, we provide expressions for the limit value as $n \to \infty$, show by tight bounds that convergence is exponential, and also characterise the rate of convergence to predefined error thresholds. We then show a few instances of statistical attacks that can be assessed by a direct application of our model: attacks against modular exponentiation that exploit timing leaks, against anonymity in mix-nets and against privacy in sparse datasets.

Keywords: quantitative information flow, statistical attacks, anonymity, privacy, information theory.

1 Introduction

Statistical attacks against secrecy, anonymity, privacy and other confidentiality properties in systems that handle sensitive data abound in the literature. In these attacks, the adversary gets to know a sample of observations of a target system – such as timing or power consumption traces of a smart-card [14], attribute values in a dataset [19], etc. – and, exploiting some form of correlation existing between the secret and the observables, tries to infer the secret – the private key, the identity of an individual, etc. Many of these attacks seem to exploit very specific features of the target system. This fact makes assessing the security of a system against this form of threat a difficult task in general. A major motivation of the present paper is to put forward a general Bayesian model where this kind of assessment can be conducted rigorously. One of our objectives is to characterise the *information leakage* of a system, both quantitatively and qualitatively, as the number of observations of the attacker increases.

* Work partially supported by the EU funded project ASCENS. 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

V. Atluri and C. Diaz (Eds.): ESORICS 2011, LNCS 6879, pp. 588–606, 2011.

It has been recently argued [8] that, for the purpose of quantifying the amount of sensitive information that is leaked by a system, it is useful to model the system itself as a *channel* in the sense of Information Theory: inputs to the channel represent the secret information, outputs represent the observable information, and the two sets are related by a conditional probability matrix. We collectively designate systems amenable to this kind of analysis as *information hiding systems (*IHS*)*. Initial works on IHS's concentrated on Shannon entropy and capacity as measures of information leakage [8,9]. More recently, it has been argued [21] that min-entropy based metrics, taking into account the success probability of an optimal attacker, provide a more operational and sensible formalization of leakage. Analysis of IHS's in the case of min-entropy and repeated independent observations, which encompasses several forms of statistical attacks, has been carried out in our previous paper [5].

A drawback of the IHS approach so far is that it focuses exclusively on the quantitative aspect of the analysis (*how much* is leaked), while ignoring the qualitative aspect (*what* is leaked) at all. In [5] it is shown that, when a uniform distribution on the secrets is assumed, the asymptotic information leakage of a system corresponds to the log of the number of indistinguishability classes in the system – where two states are indistinguishable if they induce the same probability distribution on the observables. For instance, an anonymity protocol in which users are grouped into a small number of classes is considered as globally secure. However, it might well be the case that, while the vast majority of users belong indeed to large classes, few individual users belong to a singleton classes, hence being totally exposed to eavesdropping. To make another, extreme example, consider the two small imperative procedures `P1` and `P2` below. Both of them receive as an argument a confidential variable `h` that can take on a value in the set $\mathcal{S} = \{0, ..., 15\}$, possibly corresponding to user identifiers or other sensitive information. Part of the information about `h` is disclosed by the procedures through the public variable `l`.

```
P1(h): l=-1; if (h==0) then l=0;          P2(h): l=h mod 4;
```

In the case of `P1`, there are two possible observables, -1 and 0, hence $\mathcal{S}$ is partitioned into two indistinguishability classes: thus, assuming `h` is uniformly distributed, `P1` leaks 1 bit of information about `h`. In the case of `P2` there are four classes, hence `P2` leaks two bits. From a global point of view, `P1` is therefore more secure than `P2`. Needless to say, though, *from the point of view of user 0*, `P2` is preferable over `P1`. One would like to conduct the analysis both at a quantitative and at a qualitative level, revealing not only how much is leaked, but also what. This is particularly relevant in relation to the privacy of individuals or groups.

In this paper, we propose a framework to deal with this issue by extending the IHS's considered in [5] and elsewhere with *views*. A view is, in short, a partition of the states, representing perhaps a subdivision in "buckets" of a large population (in fact, we are more general and also admit probabilistic partitions). In the example above, the view of interest to user 0 is the partition of $\mathcal{S}$ into $(\{0\}, \mathcal{S}\setminus\{0\})$. Given a view W, one is interested in the adversary's probability of wrongly predicting the class of W the secret belongs to, after observing n independent executions of the system, throughout which the secret state is kept fixed: call this quantity $P_e^W(n)$. In the example above, the involved systems are deterministic, hence a single observation is all the attacker needs. One easily finds

that $P_e^W(1)$ equals 0 in the case of P1, and $\frac{1}{16}$ in the case of P2. In the general case of probabilistic systems, computation of the limit value of $P_e^W(n)$ is not as obvious. Nevertheless, we offer results that allow one to easily characterise the behaviour of $P_e^W(n)$ from the channel matrices defining the IHS and the view W. In particular, we show how to determine the limit value of $P_e^W(n)$ and its *rate*. In fact, the security of a system (w.r.t. W) depends not only on the limit in question, but also on the shape of $P_e^W(n)$ as a function of n. We show that the convergence is exponential, and provide bounds for the rate of convergence. More generally, we give bounds on the rate at which a chosen probability threshold can be reached[1].

We then give a few examples of statistical attacks that can be assessed as a direct application of our results: timing attacks against exponentiation with blinding [14,17], attacks against anonymity in mix-nets [13] and attacks against privacy in sparse datasets [19]. In the last case, we show that the condition of (ϵ, δ)-sparsity directly translates into a rate $-\log \epsilon$ for the threshold δ in our framework. In all cases, we highlight the role played by views.

In summary, we offer a unifying model for assessing a variety of statistical attacks, both at the global level and at the level of specific partitions of the secrets. We believe that this model can gain us a qualitative insight about the security of IHS's. Whenever the system is found to be insecure, an attack can be explicitly described, usually with little effort, as an instantiation of the Bayesian attacker underlying our framework. We are *not* claiming, of course, that our bounds always match the performance of existing attacks, tailored against specific, real-world systems.

The rest of the paper is organized as follows. In Section 2 some terminology and notation are introduced. Section 3 introduces the formal set up. Section 4 discusses the main results on asymptotic error probability. Section 5 presents an application to mix-nets, while Section 6 discusses sparse datasets. Some concluding remarks and discussion of related work are found in Section 7. Some technical material has been confined to a separate Appendix.

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 $\sum_{a\in\mathcal{A}} p(a) = 1$. We let $\mathrm{supp}(p)$ denote $\{a \in \mathcal{A} : p(a) > 0\}$. 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 $A \subseteq \mathcal{A}^n$, we will often write $p^n(A)$ as just $p(A)$, if n is clear from the context.

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)

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

[1] Indeed, it may well be the case that, even if the asymptotic rate of convergence to the limit value is extremely slow, convergence to the chosen threshold is very fast, leading to consider the system insecure.

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 Formal Set Up

3.1 Basic Definitions

We recall from [5] that an *information hiding system* (IHS for short) 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 $\mathcal{S}$, $p(\cdot)$, 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 $p(\cdot|s)$.

Definition 1 (views). *Let $\mathcal{H} = (\mathcal{S}, \mathcal{O}, p(\cdot), p(\cdot|\cdot))$ be a* IHS. *A* view *of $\mathcal{H}$ is a pair $(\mathcal{W}, q(\cdot|\cdot))$, where $\mathcal{W}$ is a finite alphabet and $q(\cdot|\cdot) \in [0,1]^{\mathcal{S}\times\mathcal{W}}$ is a matrix where all rows sum to 1.*

Informally, $q(w|s)$ is the probability that the property w holds when in state s. The probability distribution p on $\mathcal{S}$ and the conditional probability matrices $p(o|s)$ and $q(w|s)$ induce a probability distribution r on $\mathcal{W}\times\mathcal{S}\times\mathcal{O}$, defined as $r(w, s, o) \stackrel{\triangle}{=} p(s)\cdot p(o|s)\cdot q(w|s)$. This distribution induce a triple of discrete random variables $(W, S, O) \sim r$, taking values in $\mathcal{W} \times \mathcal{S} \times \mathcal{O}$. We shall denote the marginal probability distributions of this triple for S, W and O by p_S, p_W and p_O, respectively. Of course, $p_S(\cdot)$ coincides with the prior $p(\cdot)$ given in the IHS, while the marginal distributions p_W and p_O can be computed from the given data, $p(\cdot)$, $p(\cdot|\cdot)$ and $q(\cdot|\cdot)$.

Let us now discuss the observation 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 both the secret state s and the corresponding view w are 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. Moreover, the individual O_i and the view W are *conditionally independent* given S. This means that the following equality holds true for each $o^n \in \mathcal{O}^n$, $w \in \mathcal{W}$ and $s \in \mathcal{S}$ s.t. $p(s) > 0$

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

Note that the right-hand side of the above equality can be equivalently written as $\Pi_{i=1}^n p(o_i|s)q(w|s)$. Concerning the goals of the attacker, there are two cases, which we examine in the following subsections.

Notation: We shall drop the subscripts from the above defined (conditional) probability distributions when no ambiguity can arise. We will often abbreviate $\Pi_{i=1}^n p(o_i|s)$

as $p(o^n|s)$. Moreover, by slightly abusing notation, we will freely identify a view $(\mathcal{W}, q(\cdot|\cdot))$ of $\mathcal{H}$ with the induced random variable W.

3.2 Attacker Targets *S*

We first discuss the case when the attacker targets the states, like in [5]. In this case, his strategy, for any fixed length n of observations, is modeled by a *guessing function* $g : \mathcal{O}^n \to \mathcal{S}$, which represents the single guess the attacker is allowed to make about the secret state s, after observing o^n. In this case, one is interested in the *probability of error after n observations* (relative to g), given by

$$P_e^{(g)}(n) \triangleq \Pr(g(O^n) \neq S)\,.$$

It is well-known (see e.g. [12]) that the optimal strategy for the adversary, that is the one that minimizes the error probability, is the Maximum A Posteriori (MAP) rule. A function $g : \mathcal{O}^n \to \mathcal{S}$ satisfies the *Maximum A Posteriori (MAP) criterion*[2] if for each $o^n \in \mathcal{O}^n$ and $s \in \mathcal{S}$

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

In the above definition, for the case $n = 0$ it is convenient to stipulate that $p(\epsilon|s) = 1$: that is, with no observations at all, it is selected some s maximizing the prior distribution. With this choice, $P_e^{(g)}(0)$ denotes $1 - \max_s p(s)$. Once n and $p(s)$ are fixed, $P_e^{(g)}(n)$ does *not* depend on the specific MAP function g that is chosen. Unless otherwise stated, throughout the paper we assume the underlying guessing function is MAP and shall normally omit the superscript $^{(g)}$.

In [5], it is proven that $P_e(n)$ converges exponentially fast to a quantity that depends on an *indistinguishability* relation on states. This relation is defined as follows: $s \equiv s'$ if $p(\cdot|s) = p(\cdot|s')$. Concretely, two states are indistinguishable if the corresponding rows in the conditional probability matrix $p(\cdot|\cdot)$ are equal. This intuitively says that there is no way for the adversary to tell s and s' apart, no matter how many observations he performs. Let us stress that this definition does not depend on the prior distribution on states, nor on the number n of observations. Assume $\equiv$ partitions $\mathcal{S}$ into K equivalence classes $C_1, ..., C_K$. For each i, let $s_i^* \in C_i$ be a state that $p_S(s_i^*) = \max_{s \in C_i} p_S(s)$. Let

$$\pi_i \triangleq p_S(s_i^*) \quad \text{and} \quad p_i(\cdot) \triangleq p(\cdot|s_i^*)\,. \tag{1}$$

We can assume w.l.o.g. that $\pi_i > 0$ for each i. In [5], it is shown that as $n \to \infty$, then exponentially fast

$$P_e(n) \to 1 - \sum_{i=1}^{K} \pi_i\,. \tag{2}$$

Note that the case $|\mathcal{S}| = 2$ with a nontrivial indistinguishability corresponds to the Bayesian version of the classical binary Hypothesis Testing; in this case, the Chernoff information is known to be the optimal exponent (see [12, Ch.11] and Section 4).

[2] Another widely used criterion for guessing functions is *Maximum Likelihood (ML)*, which requires no knowledge of the prior distribution. Our main results can be extended to the ML rule, although we will not discuss this issue in the present paper. See [5, Remark 2].

3.3 Attacker Targets W

We discuss now the case when the attacker targets a property of states represented by a view W. Similarly to the previous case, the attacker's strategy corresponds to a guessing function, which this time is of the form $g : O^n \to \mathcal{W}$. The corresponding error probability (after n observations, relative to g) is

$$P_e^{g,W}(n) \triangleq \Pr\left(g(O^n) \neq W \right). \tag{3}$$

A function g minimizes this quantity if it is W-MAP, that is if satisfies the following condition. For each $o^n \in O^n$ and $w \in \mathcal{W}$

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

Unless otherwise stated, given a view of $\mathcal{H}$, we shall assume an underlying guessing function that is W-map. Consequently, we shall normally omit the indication of g from $P_e^{g,W}(n)$.

In many systems, the practically important views are those that partition the state-space into equivalence classes. A view W is called a *partition* of $\mathcal{H}$ if W is a function of S, that is $W = f(S)$ for some function $f : \mathcal{S} \to \mathcal{W}$. Equivalently, the matrix $q(\cdot|\cdot)$ has a single entry '1' for each row. Let $\mathcal{W} = \{w_1, ..., w_L\}$, and let $E_i \triangleq f^{-1}(w_i)$ for $1 \leq i \leq L$. Of course $E_1, ..., E_L$ forms a partition of $\mathcal{S}$, in the set-theoretic sense.

3.4 Information Leakage

Information leakage aims at measuring, typically in bits, the information leaked by a system, by comparing the prior to the posterior (to the observations) adversary's *success* probability. Below, we follow Smith [21] and define information leakage as the difference between the min-entropies of the prior and posterior probability distributions. In what follows, we pose $P_{succ}(n) \triangleq 1 - P_e(n)$; similarly for P_{succ}^W. The intuition underlying this definition is that gaining 1 bit of information corresponds to doubling the success probability.

Definition 2 (Information leakage [21]). *The* information leakage *of $\mathcal{H}$ after n observations is defined as*

$$\mathcal{L}(n) \triangleq \log\left(\frac{P_{succ}(n)}{\max_s p_S(s)} \right).$$

Similarly, information leakage after n observations relative to a view W *is defined as* $\mathcal{L}^W(n) \triangleq \log(\frac{P_{succ}^W(n)}{\max_w p_W(w)})$.

4 Asymptotic Error Probability

Throughout the section $\mathcal{H}$ denotes a generic IHS $(\mathcal{S}, O, p(\cdot), p(\cdot|\cdot))$. We begin with a few preliminary definitions concerning the rate of convergence. Then prove a result giving strong bounds for $P_e(n)$ and its rate of convergence, Theorem 1. This result greatly improves on the bounds in [5] and is the key to the results for $P_e^W(n)$.

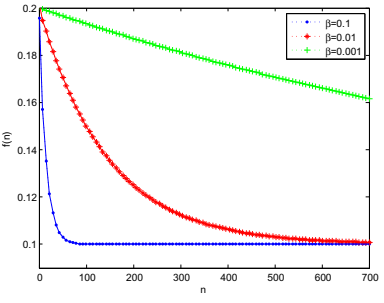
Definition 3 (rate). *Let $f : \mathbb{N} \to \mathbb{R}^+$ be a nonnegative, monotonically non-increasing function. Let $\gamma = \lim_{n\to\infty} f(n)$. The* rate *of f is defined as the nonnegative quantity*

$$\rho(f) \stackrel{\triangle}{=} -\lim_{n\to\infty} \frac{1}{n} \log(f(n) - \gamma)\,. \tag{4}$$

We further say that f reaches *δ at rate ϵ if there is a nonnegative, monotonically non-increasing function h s.t. $\lim_{n\to\infty} h(n) \le \delta$, $\rho(h) \ge \epsilon$ and $f(n) \le h(n)$ for each n large enough.*

Note that we admit rates of 0, as well as of $+\infty$.

Example 1. Consider $f(n) = \alpha + \beta 2^{-n\lambda_1} + \gamma 2^{-n\lambda_2}$, for some nonnegative α, β and γ, and $0 < \lambda_1 < \lambda_2$. Then $f(n) \to \alpha$ and $\rho(f) = \lambda_1$. On the other hand, since $f(n) \le h(n) = \alpha + \beta + \gamma 2^{-n\lambda_2}$, one has that f reaches $\alpha + \beta$ at a rate of λ_2. The picture on the right displays a plot of three functions, characterised by identical values of $\alpha = 0.1$, $\gamma = 0.01$, $\lambda_1 = 0.01$, and $\lambda_2 = 2$, and by three different values of β: $\beta = 0.1$ (top curve), 0.01 (middle curve) and 0.001 (bottom curve).



One can see that although the convergence to the limit value, 0.1 for all of them, is extremely slow, convergence to the value 0.11, which is only slightly higher, in the third case is very fast. A system with an error probability function of this shape would not be considered as secure.

Recall from [12] that given two probability distributions p and q on O, the *Chernoff Information* between p and q is the nonnegative quantity

$$C(p,q) \stackrel{\triangle}{=} -\min_{0\le\lambda\le 1} \log \sum_{o\in \mathrm{supp}(p)\cap \mathrm{supp}(q)} p^{\lambda}(o) q^{1-\lambda}(o) \tag{5}$$

with the convention that $C(p,q) = +\infty$ if $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \emptyset$. Recall that, in our notation, $p_1(\cdot), ..., p_K(\cdot)$ are the representative probability distributions of $\mathcal{H}$, defined in (1). By adapting the proof for the case $|\mathcal{S}| = 2$ that is given in [12] (see also [18]), it is not difficult to prove the following result, which gives the exact rate of convergence for $P_e(n)$, in the case where the distributions $p_1(\cdot), ..., p_K(\cdot)$ *all have the same support*[3].

Proposition 1. *Suppose that* $\mathrm{supp}(p_1) = \cdots = \mathrm{supp}(p_K)$. *Then* $\rho(P_e) = \min_{i\neq j} C(p_i, p_j)$.

The next result provides tight bounds on the error probability $P_e(n)$ and its rate in the general case, although in general not the *exact* rate. More generally, the result below provides a means to tradeoff bounds on error probability with bounds on the rate of convergence. We make use of the following notations. For all $i, j = 1, ..., K$, define

$$c_{ij} \stackrel{\triangle}{=} C(p_i, p_j)\,.$$

[3] In the case where the distributions have different supports, the argument of [12] does not apply. The ultimate reason is that that $D(p\|q)$ is not continuous in the first argument if q has not full support; see also [2] for a discussion on this issue.

We also stipulate that $2^{-\infty} = 0$. The next theorem has the following interpretation. The attacker focuses on a subset of the representative states, $\{s_i^* | i \in I\}$, and tries to identify one of them as S. This strategy can fail for two reasons: either S is not in the target subset (first term in the error expression), or it is, but the attacker mistakes one state in the subset for another (second term in the error expression). The latter probability decreases exponentially fast with n, at a rate that is at least as big as the minimum "distance" ρ_I between the distributions $p_i(\cdot)$, for $i \in I$. The proof can be found in the Appendix.

Theorem 1. *Let I be a nonempty subset of $\{1, ..., K\}$. Let $\rho_I \triangleq \min_{i,j\in I, i\neq j} c_{ij}$. Let $\pi_{\max} = \max_{i\in I} \pi_i$. Then, for all $n \geq 1$*

$$P_e(n) \;\leq\; (1 - \sum_{i\in I} \pi_i) \;+\; \frac{|I|^2}{2} \pi_{\max} 2^{-n\rho_I} \,. \tag{6}$$

As a consequence, $P_e(n)$ reaches $(1 - \sum_{i\in I} \pi_i)$ at a rate of ρ_I. In particular, by taking $I = \{1, ..., K\}$, we obtain that $\rho(P_e) \geq \rho_I$.

Remark 1. (a) In the practically important case where the prior p_S on $\mathcal{S}$ is uniform, the term $\frac{|I|^2}{2}\pi_{\max} 2^{-n\rho_I}$ is bounded above by $\frac{K}{2} 2^{-n\rho_I}$.

(b) Computation of the Chernoff Information (5) is an optimization problem that may be difficult to solve exactly. In practice, setting $\lambda = \frac{1}{2}$ in the argument of the min often yields a good lower bound of $C(p, q)$, known as *Bhattacharyya distance*. Another lower bound that we will find useful in the case of distributions with sparse support (see Section 6), is obtained by taking the min limited to the cases $\lambda = 0$ and $\lambda = 1$. Letting $\sigma = \text{supp}(p) \cap \text{supp}(q)$, this quantity amounts to $-\min\{\log p(\sigma)\,,\, \log q(\sigma)\}$.

We analyse now the case of P_e^W, where W is a generic view of an ihs $\mathcal{H}$. We follow the notation and terminology established in the previous section. It would be tempting to proceed as follows: build a new ihs, say $\mathcal{H}^W$, where the states are $\mathcal{W}$ and the channel matrix is $p_{O|W}$. The error probability function for $\mathcal{H}^W$ would then coincide with $P_e^W(n)$. It would then be enough to apply Theorem 1 to $\mathcal{H}^W$. This approach however is doomed to failure. In fact, the assumption that the observations O_i are conditionally independent given W is in general false:

$$p(o_1 \cdots o_n | w) \neq p(o_1|w) \cdots p(o_n|w) \,.$$

As a consequence, the ihs $\mathcal{H}^W$ is meaningless for what concerns our purposes. However, conditional independence of the O_i's given W is guaranteed, and the approach outlined above *does* work, in the special case where W is a partition finer than $\equiv$. This intuition leads us to develop to the method illustrated below for P_e^W in the general case.

Some more notation first. For notational simplicity, assume $\mathcal{W}$ is a set of integers $\{1, ..., |\mathcal{W}|\}$. Let $q(\cdot|\cdot)$ be the matrix defining the view W. We denote by $\sim_W$ the equivalence relation on $\mathcal{S}$ induced by $q(\cdot|\cdot)$, that is

$$s \sim_W s' \text{ iff } \text{ for each } o \in \mathcal{O} : \; q(o|s) = q(o|s') \,. \tag{7}$$

In other words, two states are $\sim_W$-equivalent if the corresponding rows of $q(\cdot|\cdot)$ are equal. Let $\mathcal{S}/\sim_W$ be $\{E_1, ..., E_L\}$, the equivalence classes of $\sim_W$. The intersection $\equiv \cap \sim_W$

is still an equivalence relation on $\mathcal{S}$, that is finer than both $\equiv$ and $\sim_W$. Recall that $\mathcal{S}/\equiv$ is $\{C_1, ..., C_K\}$. For $1 \leq i \leq K$ and $1 \leq j \leq L$, we let the equivalence classes of $\equiv \cap \sim_W$ be denoted as

$$F_{ij} \triangleq C_i \cap E_j \tag{8}$$

and furthermore

$$F_i^* \triangleq \max_j p_S(F_{ij}) \quad \text{and} \quad q_j^* \triangleq \max_w q(w|s), \text{ for an arbitrary } s \in E_j\,. \tag{9}$$

The next theorem has the following interpretation. The attacker focuses on a subset of the representative states, $\{s_i^* | i \in I\}$. He tries to identify first the class C_i of S, then guesses the class F_{ij} – this is given by the j that maximizes $p_S(F_{ij})$. Finally he guesses the view w that is most likely in E_j. This strategy can fail for two reasons: either w is wrong (first term in the expression), or F_{ij} is wrong (second + third term). We report a proof of this result in the Appendix.

Theorem 2. *Let I and ρ_I be chosen as in Theorem 1. Let W be a view of $\mathcal{H}$. Let $\Pi_{\max} = \max_{i \in I} F_i^*$. Then*

$$P_e^W(n) \;\leq\; \sum_{j=1}^{L} (1 - q_j^*) \;+\; (1 - \sum_{i \in I} F_i^*) \;+\; \frac{|I|^2}{2} \Pi_{\max} 2^{-n\rho_I}\,. \tag{10}$$

Note that the determination of the upper-bound in (10) is computationally practical: the partitions induced by $\equiv \cap \sim_W$ can be directly computed by inspection of the matrices $p(\cdot|\cdot)$ and $q(\cdot|\cdot)$. Their intersection (8), and the probability mass of the corresponding classes $p_S(F_{ij})$, are then straightforward to compute. Theorem 2 only provides an (exponential) upper bound to $P_e^W(n)$. The following theorem provides the exact limit of $P_e^W(n)$ in the special, but important case when W is a partition.

We introduce quickly a few concepts of the *method of types* from Information Theory [12, Ch11] that will be used in the proof. Fix $n \geq 1$. Given a sequence $o^n \in \mathcal{O}^n$ and $o \in \mathcal{O}$, denote by $\mathrm{n}(o, o^n)$ the number of occurrences of o inside o^n. The empirical distribution or *type* of o^n is the distribution on $\mathcal{O}$ defined as $t_{o^n}(o) \triangleq \mathrm{n}(o, o^n)/n$, for each $o \in \mathcal{O}$. The "balls" of center $p_i(\cdot)$ and radius $\epsilon > 0$ in $\mathcal{O}^n$ are defined as $U_i^n(\epsilon) \triangleq \{o^n : D(t_{o^n} \| p_i) \leq \epsilon\}$. It is a result from the method of types that, as $n \to +\infty$, $p_i(U_i^n(\epsilon)) \to 1$, while, for any $p \neq p_i$ there is $\epsilon > 0$ small enough s.t. $p(U_i^n(\epsilon)) \to 0$. Moreover, the convergence is exponential in both cases.

Theorem 3. *Let W be a partition of $\mathcal{H}$. Then $P_e^W(n)$ converges exponentially fast to $1 - \sum_{i=1}^K F_i^*$. More precisely, with the same notation of Theorem 2, for each $n \geq 1$, $1 - \sum_{i=1}^K F_i^* \leq P_e^W(n) \leq (1 - \sum_{i=1}^K F_i^*) + \frac{K^2}{2} \Pi_{\max} 2^{-n\rho_I}$, where $I = \{1, ..., K\}$.*

Proof. (Outline) First, note that for W a partition, the first term in (10) vanishes, as each q_j^* equals 1. The upper bound is then a consequence of Theorem 2 with $I = \{1, ..., K\}$. We now seek for a lower bound of $P_e^W(n)$. We equivalently focus on an upper bound of $P_{succ}^W(n)$. Assume without loss of generality that $\mathcal{W} = \{1, ..., L\}$. For any $n \geq 1$, let

$g : \mathcal{O}^n \to \{1, ..., L\}$ be a W-MAP guessing function, and let $A_j = g^{-1}(j)$, for $j \in \{1, ..., L\}$, be the acceptance region in $\mathcal{O}^n$ for j. It is a routine task to check that

$$P^W_{succ}(n) = \sum_{i=1}^{K} \sum_{j=1}^{L} p_i(A_j) p_S(F_{ij}) \,. \tag{11}$$

Now, fix any $i \in \{1, ..., K\}$, and let $j_i = \text{argmax}_{j=1,...,L} p_S(F_{ij})$, that is $p_S(F_{ij_i}) = F^*_i$. We claim that $p_i(A_{j_i}) \to 1$ as $n \to +\infty$. In fact, fixed $\epsilon > 0$ small enough, for any n large enough A_{j_i} contains the "ball" $U^n_i(\epsilon)$ of center $p_i(\cdot)$ and radius ϵ in $\mathcal{O}^n$. To see that this is true, note that a sufficient condition for $o^n \in A_{j_i}$ is that for each $j \neq j_i$

$$p_{O^n|W}(o^n|j_i) p_W(j_i) = \sum_{l=1}^{K} p_l(o^n) p_S(F_{lj_i}) \;>\; \sum_{l=1}^{K} p_l(o^n) p_S(F_{lj}) = p_{O^n|W}(o^n|j) p_W(j) \,. \tag{12}$$

Now from results of the method of types it follows that, for $o^n \in U^n_i(\epsilon)$, we have that all the $p_l(o^n)$ with $l \neq i$ go exponentially fast to 0 as n grows. Thus the condition (12) reduces, for n large enough, to $F^*_i = p_S(F_{ij_i}) > p_S(F_{ij})$: this is satisfied by definition of j_i[4]. Now $A_{j_i} \supseteq U^n_i(\epsilon)$ implies that $p_i(A_{j_i})$ goes to 1 exponentially fast as n grows; for the same reason, $p_i(A_j)$ goes to 0 for each $j \neq j_i$ as n grows (recall that the A_j's form a partition of $\mathcal{O}^n$). This way, and taking (11) into account, we have proved that

$$\lim_{n\to\infty} P^W_{succ}(n) \;=\; \sum_{i=1}^{K} F^*_i \,.$$

Since $P^W_{succ}(n)$ is monotonically non-decreasing, we have proved that $P^W_{succ}(n) \leq \sum_{i=1}^{K} F^*_i$ holds true for each $n \geq 1$. This implies in turn the wanted statement.

Example 2 (modular exponentiation). We consider timing attacks against implementations of the modular exponentiation algorithm with blinding, used in public-key cryptography – see e.g. [14,16,17,5] and references therein. A typical implementation of modular exponentiation works as follows. The bits of the secret exponent are scanned from right to left, or vice-versa. When the ith bit is considered ($0 \leq i < N$), either one or two modular multiplications are performed, depending on whether the i-th bit is 0 or 1. In timing attacks, the attacker tries to reconstruct the secret key by sampling the duration of several independent executions of the algorithm. To an implementation as described above there corresponds an IHS where: $\mathcal{S} = \{0, 1\}^N$ is the set of secret keys, i.e. the possible exponents of the algorithm, over which we assume a uniform distribution can be assumed; $\mathcal{O} = \{t_1, t_2, ...\}$ is the finite set of possible execution times; $p(t|s)$ is the probability that, depending on the deciphered message, the execution of the algorithm takes times t given that the secret key is k. As argued in [5], it is sensible to assume that any two keys having the same Hamming weights are indistinguishable in $\mathcal{H}$. Therefore,

[4] If there is more than one index j maximizing $p_i(F_{ij})$, then the choice of j_i gets more involved: among those j's that maximize $p_S(F_{ij})$, one chooses the one that maximizes $p_S(F_{i'j})$, where $p_{i'}(\cdot)$ is the distribution closest to $p_i(\cdot)$ in terms of KL-distance, if this j is unique; otherwise one must look at the second closest distribution $p_{i''}(\cdot)$, and so on. We omit the details here.

we have $N+1$ indistinguishability classes. From each of them we choose a representative s_i^* of probability $\pi_i = \frac{1}{2^N}$. Applying Theorem 1, we find $P_e(n) \to 1 - \frac{N+1}{2^N}$, which for realistic values of N, is very close to 1. E.g., for $N = 1024$, the attacker gets on the limit $\log(1025) \approx 10.01$ bits of information leakage out of 1024.

One would then like to prove that this small leakage is not concentrated in few individual bits of the exponent, which would make them potentially vulnerable. For instance, let us examine the error probability of guessing the least two significant bits of the exponent. Let W be the partition of $\mathcal{S}$ s.t. $s \sim_W s'$ iff $s \bmod 4 = s' \bmod 4$. We apply Theorem 3 to P_e^W. We have four $\sim_W$-classes $E_0, ..., E_3$, that intersect with the $N+1$ classes C_i to form $4(N+1)$ classes F_{ij}. Assume N even. For all $i = 0, ..., \frac{N-2}{2}$, the class F_{ij} that has more elements, hence determines the probability F_i^*, is F_{i0}; by symmetry, for $i = \frac{N-2}{2} + 1, ..., N$ the class with more elements is F_{i3}. For $i = \frac{N}{2}$, instead, we can choose between F_{i1} and F_{i2}. According to Theorem 3 then

$$P_{succ}^W \;\to\; \sum_{i=0}^{N} F_i^* \;\approx\; \frac{1}{2^N}\Big(\sum_{i=0}^{N-2} \binom{N-2}{i}\Big) \;=\; \frac{1}{4}\,.$$

Thus, asymptotically the observations do not increase the prior probability of success, which is already $\frac{1}{4}$. In terms of information leakage, one gets $\mathcal{L}^W(n) \to\approx 0$. One can generalize this reasoning to the case where W represent the least m significant bits, and arrive at similar conclusions.

5 Example 1: Unlinkability in Threshold Mix-Nets

Statistical attacks against anonymity protocols may take advantage of sender-receiver relationships that remain fixed through repeated rounds of the protocol. In this section, we consider the case of a *mix network*, a concept due to Chaum [10]. In a mix-network, messages are relayed through a sequence of trusted intermediary nodes, called *mixes*, in order to hide sender-receiver relationships (*unlinkability*). In the scenario we consider, a single mix is used by a number of senders and receivers. The *threshold* of the mix is $b+1$: at each round, the mix waits for $b+1$ messages from the senders and then distributes the messages to the corresponding receivers. We consider the situation where one of the senders is always Alice, with her receiver being always a node Bob, initially unknown to the attacker. The recipients of the remaining b messages are assumed be chosen at random in a set of nodes $R_1, ..., R_N$. A similar scenario is at the basis of the statistical disclosure attack by Danezis [13]. We analyse the situation of a local eavesdropper that observes one fixed receiver, say R_j, and after each round is able to tell whether at least one message has reached R_j. More sophisticated forms of eavesdropping could be easily accommodated (e.g. attacker observing all the nodes), but would not change significantly the outcome of the analysis. The task of the attacker is to discover which node is Bob; or at least, to tell if Bob is or not the observed node, R_j.

We can model the scenario described above by an IHS $\mathcal{H}$ where: the set of states is given by all possible nodes (potential receivers of Alice's messages), that is $\mathcal{S} = \{R_1, \ldots, R_N\}$, with $p_S(R_i) = \frac{1}{N}$ for each $i = 1, ..., N$; the set of observations is $\mathcal{O} = \{0, 1\}$, where $o = 1$ iff R_j has received at least one message at the end of the round.

The conditional probability matrix $p(\cdot|\cdot)$ is then given by the following equalities:

$$\begin{aligned} p(0|R_j) &= 0 & p(1|R_j) &= 1 \\ p(0|R_i) &= (1-\tfrac{1}{N})^b & p(1|R_i) &= 1-(1-\tfrac{1}{N})^b \quad \text{for all } i \neq j. \end{aligned}$$

Here, the first row means that, if Bob=R_j, then the attacker will observe at least one message with certainty. The second row means that, in case Bob is any node different from R_j, then the attacker will observe 0 messages only if all the b messages – other than the one sent to Bob – are not sent to R_j (Alice surely does not send to R_j). In other words, except for a permutation of the rows, we have the matrix below. Here the last row refers to R_j. This means that there are only two classes of indistinguishability: $\mathcal{S}/{\equiv}$ is $\{C_1, C_2\}$, with $C_1 = \{R_j\}$ and $C_2 = \mathcal{S} \setminus \{R_j\}$.

$$\begin{bmatrix} (1-\frac{1}{N})^b & 1-(1-\frac{1}{N})^b \\ \vdots & \vdots \\ (1-\frac{1}{N})^b & 1-(1-\frac{1}{N})^b \\ 0 & 1 \end{bmatrix}$$

We first apply apply Theorem 1 to $\mathcal{H}$, which will tell us what is the error probability in case the attacker wishes to know exactly who is Bob. We can set $I = \{i, j\}$, for any $i \neq j$, and get the following bound:

$$P_e(n) \leq \left(1-\frac{2}{N}\right) + \frac{2}{N}\left(1-\left(1-\frac{1}{N}\right)^b\right)^n.$$

As expected, the limit value $1-\frac{2}{N}$ is > 0, and the security of the system increases as N increases. The corresponding asymptotic information leakage is $\log(N \cdot \frac{2}{N}) = 1$, that is, the attacker gains 1 bit of min-entropy on the limit about the identity of Bob.

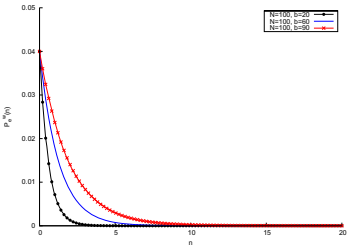
To see qualitatively *what* the single bit gained by the attacker corresponds to, we analyse the error probability with respect to the view $W \in \{0, 1\}$ given by:

$$W = 1 \text{ iff } \; S = R_j\,.$$

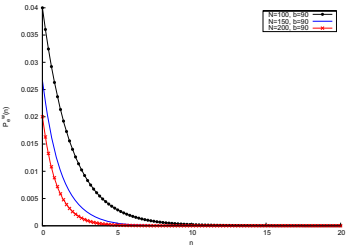
That is, W yields 1 iff Bob is R_j. The partition induced on $\mathcal{S}$ by W coincides with $\equiv$, hence its classes are C_1, C_2. Concerning the sets F_{ij}, we note that: $F_{11} = \{R_j\}$, $F_{12} = F_{21} = \emptyset$ and $F_{22} = \mathcal{S} \setminus \{R_j\}$. Since the distribution on the states is uniform, we have: $F_1^* = \frac{1}{N}$ and $F_2^* = 1-\frac{1}{N}$. Take $I = \{i, j\}$ as defined as above. According to Theorem 2, the limit of $P_e^W(n)$ vanishes, moreover

$$P_e^W(n) \leq \frac{2}{N}\left(1-\left(1-\frac{1}{N}\right)^b\right)^n.$$

The attacker's success probability of guessing whether R_j =Bob or not approaches very fast 1. It is also interesting to study the behaviour of the rate $\rho_I = -\log\left(1-(1-\frac{1}{N})^b\right)$ depending on b and N. It is easy to see that as b increases, ρ_I decreases; on the contrary, as N increases and b is kept fixed, ρ_I increases. The shape of $P_e^W(n)$ is illustrated qualitatively by the plots in the figures below: very few rounds of the protocols ($n < 10$) are sufficient to achieve $P_e^W \approx 0$.



(a) Plots of $P_e^W(n)$ depending on parameter b



(b) Plots of $P_e^W(n)$ depending on parameter N

As mentioned above, it is easy to repeat this kind of analysis with more sophisticated observations on the part of the attacker: we do not do so here for lack of space. On the other hand, note that just repeating this simple attacks for each of the potential Alice's receivers (that is, setting $R_j = R_1, R_2, ..., R_{N-1}$ in turn), would lead the attacker to uncover the identity of Bob after a low number of rounds. This is sufficient to show that the single threshold mix system is totally insecure.

6 Example 2: Privacy in Sparse Datasets

We consider datasets collecting *micro-data* – preferences, recommendations, transaction records, health histories and so on – about a large number of individuals. Datasets of this kind are sometimes published for commercial or research purposes. Making micro-data public poses serious threats to the privacy of individuals, even when the data are released in anonymized form – that is with personal identifiers, such as SSN's, removed. The risk is that an attacker, using a little of background information about a given individual and cross-correlation of attributes, might *re-identify* the individual within the dataset, leading to the disclosure of the whole set of her/his attributes. An example of this technique is the spectacular de-anonymization attack of Narayanan and Shmatikov against the Netflix Prize dataset [19][5].

In this section, we show that (sparse) datasets naturally arise as instances of IHS, and that assessment of statistical attacks against dataset privacy is easily accomplished using the general results of Section 4.

We view a dataset as a table $\mathcal{D}$, with rows and columns corresponding to individuals (or more generally, records) and attributes, respectively. Formally, $\mathcal{D} \in \mathcal{V}^{R \times A}$, where $\mathcal{V}$, R and A are finite nonempty sets of values, records and attributes, respectively. One can view any dataset $\mathcal{D}$ as an IHS $\mathcal{H}_{\mathcal{D}}$, as follows. Records are equiprobable states, that is we set $\mathcal{S} = R$ and let $p_S(\cdot)$ be the uniform distribution on R. Concerning observables, there is a variety of sensible choices, depending on the observation power one wishes to grant the attacker with. For instance, a sensible choice is $O \triangleq A \times \mathcal{V}$. Another choice,

[5] The Netflix Prize dataset collects anonymous movie ratings of 500,000 subscribers. Using background information publicly available from the Internet Movie Database, Narayanan and Shmatikov successfully re-identified known users within the Netflix dataset.

if $\mathcal{V}$ is a totally ordered, is to observe attributes and *ranges* of values. The last choice is more robust than the former in case the dataset is published in a perturbed form. In fact, even setting $O \triangleq A$ is sensible, as just knowledge of non-null attributes of a record provides a great deal of information[6]. In any case, the technical development presented below does not depend on the specific choice of O. Finally, the conditional probability matrix models the process of acquiring background information about the individuals in the dataset. Depending on its exact nature, this information might come from various sources, e.g. personal blogs, Google searches, or even a water-cooler conversation with a colleague (see [19]). For example, if $O = A$, then it is sensible to assume that the background knowledge consists of randomly chosen attributes and set, for each record r and attribute a

$$p(a|r) \triangleq \begin{cases} \frac{1}{n_r} & \text{if } a \text{ is a non-null attribute of } r \\ 0 & \text{otherwise} \end{cases}$$

where n_r is the number of non-null attributes in the row of the dataset corresponding to r. Of course, non-uniform distributions can be equally accommodated, e.g. if it is felt that certain attributes are more likely to be publicly released than others.

Having shown how to model a dataset as a IHS, we have to point out that, in the formal development below, there is no need to restrict to IHS's of the form $\mathcal{H}_{\mathcal{D}}$. To work in full generality, we will just assume a dataset is simply an IHS.

In a sparse dataset, most of the entries in the table are null. Specifically, we consider a dataset sparse if, except possibly for a small fraction of records, for no record there is another "similar" record in the dataset. To make the notion of sparsity precise, we have first to make precise the notion of similarity between records. We will work with a similarity function $\mathrm{Sim} : \mathcal{S} \times \mathcal{S} \to [0, 1]$. The intuition underlying the following definition, which is different from that proposed in [19], is that the similarity of s' to s is related to the fraction of non-null attributes they share. More precisely, it is the fraction of non-null attributes that can be inferred on any of the two by looking at the other.

Definition 4 (similarity). *Given an* IHS $\mathcal{H}$*, for any* $s, s' \in \mathcal{S}$*, let* $\sigma_{ss'} = \mathrm{supp}(p(\cdot|s)) \cap \mathrm{supp}(p(\cdot|s'))$*. We set*

$$\mathrm{Sim}(s, s') \triangleq \min\{\, p(\sigma_{ss'} \mid s)\,,\ p(\sigma_{ss'} \mid s')\,\}\,.$$

Note that $\mathrm{Sim}(s, s') = 1$ iff $\mathrm{supp}(p(\cdot|s)) = \mathrm{supp}(p(\cdot|s'))$. The following notion of sparsity is not related to - not weaker nor stronger than - the one considered in [19]. It seems to be satisfied by typical sparse datasets, like the Neflix Prize [19,20]. In fact, our results extend, although in a different form, to the notion of sparsity of [19], but we shall not give any detail here for lack of space.

Definition 5 (sparsity). *Let* $\mathcal{H}$ *be a* IHS *with* $p_S(\cdot)$ *the uniform distribution. Let* $\epsilon > 0$ *and* $\delta > 0$*. We say* $\mathcal{H}$ *is* (ϵ, δ)-sparse *if*

$$\Pr\big(\max_{s:s\neq S} \mathrm{Sim}(S, s) \geq \epsilon\big) < \delta\,. \tag{13}$$

[6] See [19] for further considerations on the structure of sparse datasets.

Our results apply to a situation where the attacker gets to know an entire copy of the dataset.We begin with a result on error probability.

Theorem 4. *Let $\mathcal{H}$ be (ϵ, δ)-sparse, with $|\mathcal{S}| = N$. Then $P_e(n)$ reaches δ at a rate $-\log \epsilon$. More precisely, $P_e(n) \leq \delta + \frac{1}{2}[(1-\delta)^2 N + 2(1-\delta) + \frac{1}{N}]\epsilon^n$.*

Proof. By definition of sparsity, it is possible to find a subset of the records, say $\mathcal{R} = \{s_i^* | i \in I\}$, s.t. for each $s \in \mathcal{R}$, there is no other record in $\mathcal{S}$ which is ϵ-similar to s, and such that $p_S(\mathcal{R}) \geq 1 - \delta$. Moreover, by uniform distribution of the probability mass on records, we can choose the size of I satisfying $\frac{|I|-1}{N} < (1-\delta) \leq \frac{|I|}{N}$, which means $(1-\delta)N \leq |I| < (1-\delta)N + 1$. Next note that, with the notation introduced in Section 4 and by virtue of Remark 1(b), for any $i, j \in I$ with $i \neq j$, the Chernoff information c_{ij} satisfies: $c_{ij} \geq -\log \mathrm{Sim}(s_i^*, s_j^*) \geq -\log \epsilon$. Applying Theorem 1 we get the thesis.

In some cases, all the adversary needs to determine about a record its "similarity class". In fact, knowledge of this class already provides him with almost all the information about the record. If this class is disclosed then a privacy breach has occurred. The next definition formalizes this intuition. Recall from (7) that $\sim_W$ is the equivalence relation induced on $\mathcal{S}$ by W.

Definition 6 **((ϵ, δ, ρ)-breach).** *Let $\mathcal{H}$ be a* IHS*. Consider a partition W of $\mathcal{H}$ such that whenever $s \sim_W s'$ then* $\mathrm{Sim}(s, s') \geq \epsilon$*. We say W is an* (ϵ, δ, ρ)-breach *if $P_e^W(n)$ reaches δ at rate ρ.*

The following result establishes strong upper bounds on the resistance to privacy breaches in sparse datasets (the proof is reported in the Appendix).

Theorem 5. *Any (ϵ, δ)-sparse* IHS *has an $(\epsilon, \delta, -\log \epsilon)$-breach W. In particular, to $P_e^W(n)$ the same bound applies as given for $P_e(n)$ in Theorem 4.*

Example 3. Real-world datasets tend to be extremely sparse. For instance, $(0.15, 0.2)$-sparsity in a dataset containing $N = 5 \times 10^5$ records should not be considered as exceptional (cf. [19, Fig.1], referring to the Netflix Prize dataset). Applying the bound of Theorem 4 to these figures, we see that already after coming across $n = 10$ randomly chosen attribute values of a target individual, the probability of uncorrect re-identification in the dataset is < 0.201. This may still seem quite high in absolute terms. Consider, however, that the success probability prior to the observations was $\frac{1}{5\times 10^5}$. In terms of information leakage, this means that the attacker has obtained $\mathcal{L}(10) \approx 18.6$ bits of min-entropy, out of $\log N \approx 18.9$. The privacy breach is therefore absolutely relevant. Note that attacks against real-world datasets can exploit specific features of the target and get more impressive success probabilities [19].

7 Conclusion

We have put forward a model to analyse a variety of statistical attacks in a uniform fashion. This permits the assessment of systems security against passive eavesdroppers both at the global level and at the level of specific partitions of the secrets. In particular,

we give precise bounds for the probability of misclassification on the part of the attacker, characterising both the limit value and the rate of convergence of the error probability as a function of the number independent observations.

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. [11] first motivated the use of mutual information to quantify information leakage in a setting of imperative programs. Boreale [4] extended this study to the setting of process calculi, and introduced a notion of rate of leakage, albeit with a different technical meaning than that considered in the present paper. Chatzikokolakis, Palamidessi and their collaborators have studied IHS's from the point of view of both capacity and error probability, but mainly confining to the case of a single observation [8,9,6,7]. The min-entropy based information leakage has been proposed by Smith [21], originally in the case of a single observation.

Backes and Köpf in [1] too consider a scenario of repeated independent observations, but from the point of view of Shannon entropy, rather than of error probability. An application of their setting to the modular exponentiation algorithm is the subject of [16], where the effect of *bucketing* on security of RSA is examined. This study has recently been extended to the case of min-entropy by Köpf and Smith in [17]. Earlier, Köpf and Basin had considered a scenario of adaptive chosen-message attacks [15]. Our previous paper [5] studies the asymptotic behaviour of information leakage. The bounds obtained there for the asymptotic rates are much looser than those we obtain here, though. Moreover, considerations on views are absent.

Our work is also related, at least conceptually, to the notion of probabilistic *opacity* as studied by Bérard, Mullins and Sassolas [3]. Indeed, although their setting is different – they work with finite-state machines – our partitions could be viewed as a generalization of the binary predicates they consider. Note however that [3] is based on Shannon entropy, and considers observations consisting of a single run of the system, rather than repeated observations, hence not statistical attacks. The Bayesian traffic analysis framework of Troncoso and Danezis [22] is tailored to the analysis of mix-networks, but mostly focuses on simulation rather than on formal models and analytical results.

As for future work, it would be natural to generalize the present scenario to the case where the attacker is given k tries for guessing the secret, with $k \geq 2$, rather than just one. Finally, the application to sparse datasets prompts a connection to databases privacy issues that deserves further attention.

Acknowledgments. The first author wishes to thank V. Shmatikov for a stimulating discussion on the notion of sparsity in datasets. Three anonymous ESORICS 2011 referees provided valuable comments.

References

1. Backes, M., Köpf, B.: Formally Bounding the Side-Channel Leakage in Unknown-Message Attacks. In: Jajodia, S., Lopez, J. (eds.) ESORICS 2008. LNCS, vol. 5283, pp. 517–532. Springer, Heidelberg (2008)
2. Baignères, T., Vaudenay, S.: The Complexity of Distinguishing Distributions (Invited Talk). In: Safavi-Naini, R. (ed.) ICITS 2008. LNCS, vol. 5155, pp. 210–222. Springer, Heidelberg (2008)

3. Bérard, B., Mullins, J., Sassolas, M.: Quantifying Opacity. In: Proc. of QEST 2010, pp. 263–272. IEEE Society, Los Alamitos (2010)
4. Boreale, M.: Quantifying information leakage in process calculi. Information and Computation 207(6), 699–725 (2009)
5. Boreale, M., Pampaloni, F., Paolini, M.: Asymptotic information leakage under one-try attacks. In: Hofmann, M. (ed.) FOSSACS 2011. LNCS, vol. 6604, pp. 396–410. Springer, Heidelberg (2011)
6. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Compositional Methods for Information-Hiding. In: Amadio, R.M. (ed.) FOSSACS 2008. LNCS, vol. 4962, pp. 443–457. Springer, Heidelberg (2008)
7. Braun, C., Chatzikokolakis, K., Palamidessi, C.: Quantitative Notions of Leakage for One-try Attacks. In: Proc. of MFPS 2009. Electr. Notes Theor. Comput. Sci, vol. 249, pp. 75–91 (2009)
8. Chatzikokolakis, K., Palamidessi, C., Panangaden, P.: Anonymity protocols as noisy channels. Information and Computation 206(2-4), 378–401 (2008)
9. Chatzikokolakis, K., Palamidessi, C., Panangaden, P.: On the Bayes risk in information-hiding protocols. Journal of Computer Security 16(5), 531–571 (2008)
10. Chaum, D.: Untraceable electronic mail, return address, and digital pseudonyms. Communications of the ACM 24(2) (1981)
11. Clark, D., Hunt, S., Malacaria, P.: Quantitative Analysis of the Leakage of Confidential Data. Electr. Notes Theor. Comput. Sci. 59(3) (2001)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2/e edn. John Wiley & Sons, Chichester (2006)
13. Danezis, G.: Statistical Disclosure Attacks. In: SEC 2003. IFIP Conference Proceedings, vol. 250, pp. 421–426 (2003)
14. Kocher, P.C.: Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996)
15. Köpf, B., Basin, D.A.: An information-theoretic model for adaptive side-channel attacks. In: ACM Conference on Computer and Communications Security, pp. 286–296 (2007)
16. Köpf, B., Dürmuth, M.: A Provably Secure and Efficient Countermeasure against Timing Attacks. In: CSF 2009, pp. 324–335 (2009)
17. Köpf, B., Smith, G.: Vulnerability Bounds and Leakage Resilience of Blinded Cryptography under Timing Attacks. In: CSF 2010, pp. 44–56 (2010)
18. Leang, C.C., Johnson, D.H.: On the asymptotics of M-hypothesis Bayesian detection. IEEE Transactions on Information Theory 43, 280–282 (1997)
19. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: IEEE Symposium on Security and Privacy 2008, pp. 111–125. IEEE Computer Society, Los Alamitos (2008)
20. Shmatikov, V.: Personal communication (2011)
21. Smith, G.: On the Foundations of Quantitative Information Flow. In: de Alfaro, L. (ed.) FOSSACS 2009. LNCS, vol. 5504, pp. 288–302. Springer, Heidelberg (2009)
22. Troncoso, C., Danezis, G.: The bayesian traffic analysis of mix networks. In: ACM Conference on Computer and Communications Security, pp. 369–379 (2009)

A Appendix

Unless otherwise stated, we use the notation and conventions introduced in Section 4.

Theorem 6 (Theorem 1). *Let I be a nonempty subset of $\{1, ..., K\}$. Let $\rho_I \triangleq \min_{i,j\in I, i\neq j} c_{ij}$. Let $\pi_{\max} = \max_{i\in I} \pi_i$. Then, for all $n \geq 1$*

$$P_e(n) \;\leq\; (1 - \sum_{i\in I} \pi_i) \;+\; \frac{|I|^2}{2}\pi_{\max} 2^{-n\rho_I}\,. \tag{14}$$

As a consequence, $P_e(n)$ reaches $(1 - \sum_{i\in I} \pi_i)$ at a rate of ρ_I. In particular, by taking $I = \{1, ..., K\}$, we obtain that $\rho(P_e) \geq \rho_I$.

Proof. Fix $n \geq 1$. Let $\mathcal{R} = \{s_i^* | i \in I\}$ and $g : \mathcal{O}^n \to \mathcal{R}$ be a function satisfying: $g(o^n) = s_i^*$ implies $p(o^n|s_i^*)\pi_i \geq p(o^n|s_j^*)\pi_j$ for each $j \in I$. Note that g need not be MAP for $\mathcal{H}$, and that $g^{-1}(s) = \emptyset$ for $s \notin \mathcal{R}$. For each $i \in I$, let $A_i = g^{-1}(s_i^*)$ be the acceptance region for s_i^*. Then we have (the sums below run over s's s.t. $p_S(s) > 0$)

$$\begin{aligned}
P_e^g(n) &= \sum_{s\in\mathcal{S}} \Pr(g(O^n) \neq s | S = s) p_S(s) \\
&= \sum_{s\notin\mathcal{R}} \Pr(g(O^n) \neq s | S = s) p_S(s) + \sum_{i\in I} \Pr(g(O^n) \neq s_i^* | S = s_i^*)\pi_i \\
&= (1 - \sum_{i\in I} \pi_i) + \sum_{i\in I} p_i(A_i^c)\pi_i \\
&\leq (1 - \sum_{i\in I} \pi_i) + \sum_{i\in I} \sum_{j\in I, j\neq i} p_i(A_j)\pi_i \\
&= (1 - \sum_{i\in I} \pi_i) \;+\; \sum_{i\in I} \sum_{j\in I, j>i} p_i(A_j)\pi_i + p_j(A_i)\pi_j
\end{aligned} \tag{15}$$

where the inequality follows from $A_i^c = \cup_{j\in I\setminus\{i\}} A_j$ and a simple union bound, while the last equality is simply a rearrangement of summands. Now, we evaluate $p_i(A_j)\pi_i + p_j(A_i)\pi_j$ for each $i, j \in I$ and $i \neq j$.

Essentially by the same derivation given in [12, eqn.(11.239)–(11.251)], one finds that $p_i(A_j)\pi_i + p_j(A_i)\pi_j \leq \pi_i^\lambda \pi_j^{1-\lambda} 2^{-nc_{ij}}$, for a suitable $\lambda \in [0, 1]$. Since $\pi_i^\lambda \pi_j^{1-\lambda} \leq \pi_{\max}^\lambda \pi_{\max}^{1-\lambda} = \pi_{\max}$ and $c_{ij} \geq \rho_I$, we obtain

$$p_i(A_j)\pi_i + p_j(A_i)\pi_j \;\leq\; \pi_{\max} 2^{-n\rho_I} \tag{16}$$

Now, if we plug the bound (16) in (15), and then factor out $\pi_{\max} 2^{-n\rho_I}$ and reorder the summands, we get

$$P_e^g(n) \;\leq\; (1 - \sum_{i\in I} \pi_i) \;+\; \Big(\sum_{i\in I} \sum_{j\in I, j>i} 1\Big)\, \pi_{\max} 2^{-n\rho_I}\,.$$

Now, use the fact that $\big(\sum_{i\in I} \sum_{j\in I, j>i} 1\big) = \frac{|I|(|I|-1)}{2} \leq \frac{|I|^2}{2}$, which shows that the wanted inequality holds for $P_e^g(n)$. But, from optimality of MAP, $P_e(n) \leq P_e^g(n)$, which completes the proof.

Theorem 7 (Theorem 2). *Let I and ρ_I be chosen as in Theorem 1. Let W be a view of $\mathcal{H}$. Let $\Pi_{\max} = \max_{i\in I} F_i^*$. Then*

$$P_e^W(n) \;\le\; \sum_{j=1}^{L}(1-q_j^*) \;+\; (1-\sum_{i\in I} F_i^*) \;+\; \frac{|I|^2}{2}\Pi_{\max}2^{-n\rho_I}\,. \tag{17}$$

Proof. Denote a pair of indices $(i,j) \in \{1,...,K\}\times\{1,...,L\}$ as ij. For each $s \in \mathcal{S}$, define $\mathrm{ind}(s) = ij$ iff $s \in F_{ij}$. Fix $n \ge 1$ and any function $g' : \mathcal{O}^n \to \{1,...,K\}\times\{1,...,L\}$, and let $Succ'$ be the event $(g'(O^n) = \mathrm{ind}(S))$. That is, $Succ'$ is the event that g' correctly classifies the index (of the equivalence class F_{ij}) of S. Now define a guessing function for $\mathcal{H}$, $g : \mathcal{O}^n \to \mathcal{W}$, as $g(o^n) \triangleq w$, where $g'(o^n) = ij$ and $w = \mathrm{argmax}_w q(w|s)$ for any $s \in E_j$ (note that the information about i provided by g' is ignored by g). Let Err be the event $(g(O^n) \neq W)$. We have

$$\begin{aligned} P_e^W(n) &= \Pr(Err, Succ') + \Pr(Err|\neg Succ')\Pr(\neg Succ') &(18)\\ &\le \Pr(Err, Succ') + \Pr(\neg Succ')\,. &(19)\end{aligned}$$

Let us estimate $\Pr(Err, Succ')$ and $\Pr(\neg Succ')$ separately. It is an easy matter to prove that

$$\begin{aligned}\Pr(Err, Succ') &= \sum_{j=1}^{L}(1-q_j^*)\Pr(S \in E_j, Succ')\\ &\le \sum_{j=1}^{L}(1-q_j^*)\,. &(20)\end{aligned}$$

We now estimate $\Pr(\neg Succ')$. Consider the new IHS $\mathcal{H}' \triangleq (\{1,...,K\}\times\{1,...,L\}, \mathcal{O}, p'(\cdot), p'(\cdot|\cdot))$, where $p'(ij) \triangleq p_S(F_{ij})$ and $p'(o|ij) \triangleq p_i(o)$. Note that $ij \equiv i'j'$ iff $i = i'$. Hence we have K distinct classes in this system, whose representatives are elements $s'_1 = 1j_1, ..., s'_K = Kj_K$ such that $j_i = \mathrm{argmax}_j p_S(F_{ij})$, hence $p'(s'_i) = F_i^*$, for $i = 1,...,K$. The corresponding representative distributions (rows of the matrix $p'(\cdot|\cdot)$) are $p'_1(\cdot) = p_1(\cdot), ..., p'_K(\cdot) = p_K(\cdot)$.

Now take the function g' above to be a MAP guessing function for $\mathcal{H}'$. Call $P'_e(n)$ the error probability of $\mathcal{H}'$: clearly, $\Pr(\neg Succ') = P'_e(n)$. Take $I \subseteq \{1,...,K\}$ and apply Theorem 1 to $\mathcal{H}'$ and I to get

$$\Pr(\neg Succ') \le 1 - \sum_{i\in I} F_i^* \;+\; \frac{|I|^2}{2}\Pi_{\max}2^{-n\rho_I}\,. \tag{21}$$

When we plug the bounds (20) and (21) into (19), we get the wanted result.

Theorem 8 (Theorem 5). *Any (ϵ,δ)-sparse* IHS *has an $(\epsilon,\delta,-\log\epsilon)$-breach W. In particular, to P_e^W the same bound applies as given for P_e in Theorem 4.*

Proof. The proof is similar to that of Theorem 4. Consider the set $\mathcal{R} = \{s_i^* \mid i \in I\}$. Build the partition W as follows: take as blocks the singletons $\{s_i^*\}$, for $i \in I$, plus the blocks obtained by breaking $\mathcal{S}\setminus\mathcal{R}$ in such a way that any two records in the same block are ϵ-similar. Then apply Theorem 2 with W and I, taking into account the bounds for $|I|$ given in the proof of Theorem 4.