An Information-Theoretic Approach to Inference Attacks on Random Data Perturbation and a Related Privacy Measure

Poorvi L. Vora, *Member, IEEE*

***Abstract*—Random data perturbation (RDP) has been in use for several years in statistical databases and public surveys as a means of providing privacy to individuals while collecting information on groups, and has recently gained popularity as a privacy technique in data mining. This correspondence provides an information-theoretic framework for all inference attacks on RDP. The framework is used to demonstrate the existence of a tight asymptotic lower bound on the number of queries required per bit of entropy for *all* inference attacks with zero asymptotic error and bounded average power in the query sequence. A privacy measure based on security against inference attacks is proposed.**

***Index Terms*—Data mining, data perturbation, information-theoretic security, noisy channel, privacy, statistical database security.**

I. Introduction

In several instances, it is necessary to provide information on groups while protecting the privacy of individuals. The most common example is that of the statistical database, whose purpose is to provide statistics (in the form of mean, median, mode, variance, etc., on health, and demographics, for example) to researchers, while keeping private sensitive information regarding specific individuals. Other examples include censuses and public surveys. While direct disclosure of sensitive individual values can be restricted through appropriate access control mechanisms, it is well known that these values can often be indirectly determined through knowledge of well-chosen statistics, see Example 1.

Example 1: Consider a statistical database that holds the salaries, departmental associations and ethnicities of all faculty at a university. The access control rules of the database prevent it from revealing statistics over groups of size smaller than, say, five. Suppose that there is only one African-American faculty member in the Department of Computer Science. While the database may not be directly queried for the salary of the single African-American faculty member in Computer Science, this value can be determined indirectly as follows. The database can be queried for C_1, the average salary of all African-American faculty, and for C_2, the average salary of all African-American faculty who are not in the Department of Computer Science, assuming $C_1 > 5$ and $C_2 > 5$. $C_1 - C_2$ is then the salary of the single African-American faculty member in Computer Science.

Information on sensitive variables may also be revealed when nonsensitive individual database variables (not statistics), are revealed, see Example 2.

Example 2: Consider a database which stores sensitive bits S_1 and S_2 on individuals, and also nonsensitive bits Q_1, Q_2, Q_3, Q_4.

S_1	"Gender."
S_2	"Over forty years of age."
Q_1	"Losing Calcium."
Q_2	"Balding."
Q_3	"Greying."
Q_4	"Gaining weight."

Suppose an adversary wishes to determine bits S_1 and S_2, but is only allowed to query bits Q_1, Q_2, Q_3, and Q_4. Because the bits S_1 and S_2 are not independent of bits Q_1, Q_2, Q_3, and Q_4, the adversary will be able to reduce the uncertainty of bits S_1 and S_2. For example, women over forty are more likely to be losing Calcium than any of the three other categories. Hence, Q_1 reveals information about S_1 and S_2. Similarly, men over forty are almost the only category balding, hence Q_2 also reveals information on both S_1 and S_2.

Attempts—such as those in Examples 1 and 2—to indirectly determine private information have been extensively studied in the database literature [1], [13], [7]. They are termed *inference attacks* (a more precise definition may be found in Section III-A). As a determined data miner can obtain several related values through several distinct sources, this correspondence's model of the problem of inference attacks does not limit the number of queries or the relationship among them, except to require the queries to be discrete-valued and of bounded power to model the finite precision of digital systems. The model can be used to represent as precise a discrete representation as desired.

The general technical problem may be modeled as follows (the term *variable* is used to denote a database variable and *random variable* or *r.v.* to denote a random variable; upper case letters represent an r.v.; lower case letters a specific instance of the r.v.).

- Database $\mathcal{A}$ contains a set of (binary, discrete-valued or continuous-valued) variables: $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_i, \ldots\}$. Each of the Q_i may also be considered an r.v.
- Let f be *any* discrete-valued function of database variables.[1] Data collector $\mathcal{B}$ queries the value of

$$X = f(C_1, C_2, \ldots, C_j, \ldots)_{C_j \in \mathcal{Q}}$$

  X is an r.v. but need not be a database variable, and is computed by the database upon request.
- The number of queries made by $\mathcal{B}$ is not limited, and the ith query is denoted X_i. The query sequences have bounded average power, denoted σ^2—i.e., those for which $E[X_i^2] \le \sigma^2$.
- The individual variables Q_i are to be "protected" from $\mathcal{B}$.

A. Random Data Perturbation (RDP)

One technique for reducing the impact of inference attacks is the addition of independent noise to the queried value X before it is revealed.[2] The larger the probabilistic perturbation, the more privacy provided to the database variables C_i, and the less accurate the response to the query. This technique is known as random data perturbation (RDP) and has been in use for about twenty years in statistical database security [1], [13]. It has recently been proposed as a means of personal privacy protection in data mining applications [2], [3], and has elicited considerable interest [4], [8], [10], [9], [12], [14], [17], [18].

Manuscript received January 4, 2005; revised September 14, 2006. The material in this correspondence was presented in part at INDOCRYPT (2004), Chennai, India, December 2004.

The author is with the Department of Computer Science, George Washington University, Washington, DC 20052 USA (e-mail: poorvi@gwu.edu).

Communicated by T. Johansson, Associate Editor for Complexity and Cryptography.

Digital Object Identifier 10.1109/TIT.2007.901183

[1] For example, f can be: "the most common gender among records 1, 2, and 3", or "the average salary of records 1, 2, and 3, quantized to a hundredth of a cent.

[2] While the added noise need not be independent, this correspondence focuses on independent noise, the subject of several papers—see, for example, surveys [1], [7], [13]. An advantage of independent noise is the speed and ease of implementation, though it is known, in general, to be less secure [10].

RDP proceeds as follows ($P_Z(.)$ represents the probability mass function of discrete-valued r.v. Z, and $p_Z(.)$ the probability density function (pdf) of continuous-valued r.v. Z):

1) $\mathcal{B}$ requests random variable X from $\mathcal{A}$, where $X = f(C_1, C_2, \ldots, C_j, \ldots)_{C_j \in \mathcal{Q}}$.
2) $\mathcal{A}$ responds with the random variable $\phi(X) = Y$ generated as follows:
   a) Discrete-valued RDP: When the noise is discrete-valued (and X takes on n possible values)

   $$P_{Y\,|\,X}(y|x) = \begin{cases} \rho, & y = x \\ \frac{1-\rho}{n-1}, & y \neq x \end{cases}$$

   for some value ρ such that $0 < \rho < 1, \rho \neq \frac{1}{n}$.
   b) Continuous-valued Gaussian RDP: When the noise is continuous-valued Gaussian, and the average signal power bounded: $E[X_i^2] \leq \sigma^2$

   $$p_{Y\,|\,X}(y|x) = p_N(y - x)$$

   where $p_{Y\,|\,X}(.)$ is the conditional probability distribution of Y given X, and $p_N(.)$ is the probability distribution of the added noise, which is zero-mean, Gaussian with variance σ_n^2.

B. Inference Attacks on RDP

As described in [19], RDP provides a communication channel with input X, output Y, and transition probability $P_{Y\,|\,X}$ for the discrete case, or $p_{Y\,|\,X}$ for the continuous case. It does not provide perfect secrecy, and does not render inference attacks completely useless. To see this, consider the following examples.

Example 3: A simple inference attack on RDP consists of repeated queries of the same variable: $\mathcal{B}$ repeatedly asks for the same quantized value of continuous variable $C, X = \mathrm{quant}(C)$, where $\mathrm{quant}(.)$ represents a quantization function. (C can be, for example, the annual salary of an individual, and $\mathrm{quant}(C)$ its value rounded off to the nearest one hundred dollars).

$\mathcal{A}$ responds, each time, with $y_i = X + n_i$, where n_i is an instance of a zero-mean gaussian r.v. with variance σ_n^2.

$\mathcal{B}$ obtains the maximum likelihood (ML) estimate of X.

The probability of error of the ML estimate of (discrete-valued) X can be decreased without bound by increasing the number of queries without bound.

$\mathcal{A}$ can avoid this attack by keeping track of all queries and the corresponding responses, and by simply providing the same value of y_i whenever queried for $\mathrm{quant}(C)$. However, all inference attacks are not as easily avoided, see Example 4.

Example 4: Consider a survey conducted on individuals in the USA who are over forty years of age. Consider a query sequence that contains some of the variables of Example 2:

Q_1 "Female."

Q_2 "Losing Calcium."

Q_3 "Balding."

Affirmative responses to Q_1 imply an affirmative response to Q_2 (all women over forty are losing Calcium), and, with high probability, a negative response to Q_3 (a large fraction of those balding are men). Suppose the responses are perturbed with RDP, and $\mathcal{B}$ receives an affirmative response to Q_1, a negative response to Q_2, and an affirmative one to Q_3. This indicates to $\mathcal{B}$, with high probability, that at least one of the responses was flipped due to the perturbation. Thus the relationship among the queries has enabled $\mathcal{B}$ to detect an error. (Notice that inference attacks such as this one cannot be avoided by keeping track of past queries, because $\mathcal{B}$ need not repeat Q_1 to obtain information on its correct value.)

Inference attacks have been common in statistical databases in spite of data perturbation. If data mining becomes widespread, it is likely that inference attacks will provide a large privacy threat in data mining as well. Several recent papers provide measures of the privacy of perturbation [3], [2], [6] in data mining. Each paper studies the case when several instances of a *single r.v.* are perturbed. None of the papers examines the effect, on privacy, of the availability of the perturbed values of several related r.v.'s, such as in an inference attack. This correspondence examines the most general inference attacks, and their costs, in terms of the number of queries required per bit of entropy. The costs are important indicators of the privacy of RDP—the larger the cost to $\mathcal{B}$, the more difficult it is for $\mathcal{B}$ to obtain accurate values of the desired variables, and the more secure or private the perturbation.

C. Contributions

This correspondence makes more general the main ideas of [19], [20]. It provides a theoretical understanding of a) the extent to which inference attacks on RDP can improve estimation error, and b) the limits of inference attacks. Its main contributions are as follows.

- A framework for the study of the security of RDP, and the corresponding definitions and associations with information theory and coding. This includes a general characterization, in information-theoretic terms, of inference attacks.
- The use of the framework in deriving a very general efficiency result; in particular, showing that the following are true.
  — In all inference attacks that have zero asymptotic error and finite average power, the number of queries required per bit determined is asymptotically bounded below by a finite value. This value is denoted $\eta_{\min}$.
  — The bound is tight: inference attacks with zero asymptotic error, bounded average power and asymptotic cost $\eta_{\min}$ exist. The attacks are based on Shannon codes and are not expected to be realistic (models of realistic attacks do not exist).
- The definition of a privacy measure based on security with respect to inference attacks, and the relationship of this measure with that of [2].

The results are consistent with those of others who predict that the addition of independent noise is insecure [10], while the information-theoretic characterization, the consideration of inference attacks, the connections with coding theory, and the derivation of the tight, finite lower bound on asymptotic cost are, to the knowledge of the author, only found earlier in [19], [20], which are conference versions whose results are incorporated in this correspondence.

D. Organization of the Correspondence

Section II provides a review of existing work. Section III contains preliminaries, including the model and definitions. Section IV presents the results with proofs. The conclusions are presented in Section V.

II. Related Work

The database community has measures of the privacy of RDP [11], [3], [2]; these are, however, not motivated by a security analysis. The security analyses that do exist [13] focus on the variance of the estimation error.

The first use of randomization to provide privacy during data mining is found in [3], which uses perturbed training data to estimate a data distribution, and uses the estimated distribution to build decision tree classifiers. The classifiers thus built have accuracies close to those built from the true data itself. [3] also proposes a privacy measure based on how closely a true value may be estimated from its perturbed value. If X can be estimated to lie in the interval (x_l, x_u) with $c\%$ confidence, then the privacy is $x_u - x_l$ at $c\%$ confidence level. This privacy measure does not take into consideration, however, the fact that, in addition to the perturbed data points, the estimated data distribution is also available to estimate the real data values.

Reference [2] proposes the use of an expectation maximization (EM) algorithm to determine a data distribution from perturbed data values, and shows that it converges to the ML estimate. [2] is also the first data mining publication to point out that a tradeoff exists between privacy and accuracy, and proposes the following measure of *conditional privacy loss*:

$$\mathcal{P}(X|Y) = 1 - 2^{-I(X;Y)} \tag{1}$$

where $I(X;Y)$ is the mutual information between r.v.'s X and Y. Recall that mutual information may be expressed as the average loss of entropy in r.v. X due to knowledge of Y, over all values of Y

$$I(X;Y) = E[\mathcal{H}(X) - \mathcal{H}(X|Y = y)]$$

Thus, for a given r.v. $X, I(X;Y)$ may be seen as the average value of r.v. $Z(y) = \mathcal{H}(X) - \mathcal{H}(X|Y = y)$, termed *privacy loss* by [6]. A problem with the measure $\mathcal{P}$, pointed out in [6], is that it is not able to detect large values of $Z(y)$ that occur with small probability, $P_Y(y)$. This would be a natural limitation of a measure based on an average value of a r.v.—it cannot be expected to represent well the maximum value. However, large values of $Z(y)$ correspond to *privacy breaches* [6] (where certain properties of the true data are revealed with great accuracy) and are undesirable, even though low values of $P_Y(y)$ imply that the breaches arise only rarely.

The measure proposed in [6] is denoted *worst-case information*, and is defined as the maximum value of $Z(y)$

$$I_w(X;Y) = \max_y[\mathcal{H}(X) - \mathcal{H}(X|Y = y)] \tag{2}$$

A large value of this measure implies a large worst-case privacy violation. [6] also defines a methodology, called *amplification*, for limiting privacy breaches. This methodology limits the value

$$\max_{y,x_1,x_2} \frac{P(Y = y|X = x_1)}{P(Y = y|X = x_2)}$$

and is proven to limit privacy breaches.

I_w is able to identify large privacy violations that occur with very low probabilities, while $\mathcal{P}$ is not. On the other hand, $\mathcal{P}$, unlike I_w, is able to distinguish cases where large privacy violations are the norm from those where they rarely occur. (Note that the values of I_w and I are identical for RDP, because, in RDP, the r.v. $Z(y)$ is constant for all y. Hence measures I_w and $\mathcal{P}$ do not provide different evaluations of RDP). The following example, similar to one in [6], demonstrates the strengths and weaknesses of both $\mathcal{P}$ and I_w as privacy measures.

Example 5: Suppose X is binary, and $P_X[0] = P_X[1] = 0.5$. Suppose e represents the empty record. The randomization R_1 produces perturbed response Y_1 as follows:

$$P_{Y_1|X}(y_1|x) = \begin{cases} 0.5, & y_1 = e \\ 0.4, & y_1 = x \\ 0.1, & \text{else} \end{cases}$$

Compare this to randomization R_2, producing output Y_2:

$$P_{Y_2|X}(y_2|x) = \begin{cases} 0.9999, & y_2 = e \\ 8 \times 10^{-5}, & y_2 = x \\ 2 \times 10^{-5}, & \text{else} \end{cases}$$

It can be seen that $I_w(X;Y_1) = I_w(X;Y_2) \approx 0.278, I(X;Y_1) \approx 0.139 \neq I(X;Y_2) \approx 0$, and $\mathcal{P}_{R_1} \approx 0.9 \neq \mathcal{P}_{R_2} \approx 1$. Thus, while $\mathcal{P}$ clearly indicates that R_1 is worse than R_2, I_w is not able to do so. On the other hand, $\mathcal{P}$ implies that R_2 is very close to perfect, while I_w clearly shows that R_2 does violate privacy. It appears that a combination of the two measures would be most useful in assessing the privacy impact of single instances of data perturbation.

Neither [2] nor [6] examine how the uncertainty in X may be reduced further if the data collector has access to perturbed information about other r.v.'s that are not independent of X. That is, neither examine the privacy of data perturbation with respect to inference attacks.

III. Preliminaries

A. The Model

We assume that $\mathcal{B}$ wishes to determine the values of k discrete-valued random variables: $\mathbf{S} = (S_1, S_2, \ldots S_k)$. (Note that these may be continuous-valued variables that have been quantized as finely as desired.) For this purpose, $\mathcal{B}$ makes a set of m queries to $\mathcal{A} : \mathbf{X} = (X_1, X_2, \ldots X_m)$. $\mathcal{A}$ provides perturbed responses, $\mathbf{Y} = (\phi(X_1), \phi(X_2), \ldots \phi(X_m))$. $\mathcal{B}$ uses these to obtain a ML estimate of $\mathbf{S}$, denoted $\hat{\mathbf{S}} = (\hat{S_1}, \hat{S_2}, \ldots, \hat{S_k}) = g(\Phi(\mathbf{X}))$, where $g(.)$ represents the ML estimate.

We now provide a few definitions (a list of symbols is in the Appendix).

Definition 1: An *inference attack* is a set of queries $\mathbf{X}$ such that $\mathbf{X}$ and $\mathbf{S}$ are not independent, i.e., $I(\mathbf{S};\mathbf{X}) \neq 0$.

The definition of an inference attack is intentionally broad, as we wish to demonstrate a limit on the capability of inference attacks. The definition also assumes nothing about the relationship between queried value X_i and previously received responses: $\phi(X_1), \phi(X_2), \ldots \phi(X_{i-1})$, and hence includes adaptive inference attacks.

We now define two important measures of an attack, the probability of estimation error and the query complexity per bit.

Definition 2: Given a particular query sequence $\mathbf{X}$ of size m, ω_m denotes the maximum probability of error of estimate $\hat{\mathbf{S}}$, over all possible values of $\mathbf{x}$ of r.v. $\mathbf{X}$. That is, if $E_{\mathbf{x}} = Pr[\mathbf{S} \neq g(\phi(\mathbf{x}))]$, then $\omega_m = \max_{\mathbf{x}} E_{\mathbf{x}}$.

Before we proceed to define the query complexity per bit, we establish some more notation. Let the alphabet of the queries be denoted $\mathcal{M}$, and its size $|\mathcal{M}|$. The number of possible values of $\mathbf{X}$ need not be $|\mathcal{M}|^m$. This is because certain symbol combinations may not be possible, as the queries are not generally independent. We denote the size of the set of all possible values of $\mathbf{X}$ by M. As $\mathcal{B}$ would want to correct for the RDP, $\mathbf{X}$ would consist of more than $\log_{|\mathcal{M}|} M$ queries.

Definition 3: The *query complexity per bit*, of query sequence $\mathbf{X}$ of length m, when $\mathbf{X}$ takes on M possible values, is $\eta_m = \frac{m}{\log_2 M}$ queries per bit.

Clearly, ω_m and η_m are related, and a lower value of ω_m would require a higher value of η_m. Because several attackers can collude to get as many queries as desired, we do not bound m, and, instead, examine the relationship between $\lim_{m\to\infty} \omega_m$ and $\lim_{m\to\infty} \eta_m$. To simplify exposition, we define the inference attack with zero asymptotic error.

Definition 4: A *small error* inference attack is one in which $\lim_{m\to\infty} \omega_m = 0$.

The behavior of η_m is well understood for the repeated query attack of Example 3, which is a small error attack

$$\lim_{m\to\infty} \omega_m = 0 \Rightarrow \lim_{m\to\infty} \eta_m = \lim_{m\to\infty} m = \infty$$

However, the attacker is not limited to the repeated query attack. It is possible that small error inference attacks may have finite values of $\lim_{m\to\infty}\eta_m$. The lower bound, over all possible inference attacks, on $\lim_{m\to\infty}\eta_m$ represents the best efforts of $\mathcal{B}$: with the "best" designed queries, $\mathcal{B}$ is forced to expend more than a minimum cost, in queries per bit, to obtain information. The larger this cost, the more the privacy provided by the RDP. We, hence, propose the following.

Definition 5: The *privacy* of RDP is the tight asymptotic lower bound on η_m for a small error attack.

B. Our Approach and the Results of [19], [20]

[19] views RDP as a communication channel. Inference attacks consist of communication of $\mathbf{S}$ over the channel. The values input to the channel are $\mathbf{X}$, and the output values $\mathbf{Y}$. $P_{Y\mid X}$ and $p_{Y\mid X}$ represent the transition probabilities of the channel for discrete and continuous-valued noise respectively.

The relationships among the values of X_i, and between $\mathbf{X}$ and $\mathbf{S}$, are controlled by $\mathcal{B}$, through the choice of $\mathbf{X}$. In a rare inference attack, $\mathcal{B}$ can choose $\mathbf{X}$ to form a channel code in $\mathbf{S}$, though this is not typical. In these attacks, $\mathbf{X}$ is a function of $\mathbf{S}$, and η_m^{-1} are the rates of the codes. When such attacks are zero-error with constant $\eta_m = \eta$, they correspond to reliable codes, and the inverse of channel capacity is the minimum value of η, achieved by attacks that correspond to Shannon codes for binary X_i and S_i. All inference attacks do not correspond to channel codes, and [20] addresses the case when the inference attack is not a code (that is, $\mathbf{X}$ is not a deterministic function of $\mathbf{S}$), is not reliable (that is, η_m is not constant), and X_i and S_i are binary.

The results of this correspondence generalize those of [19], [20] to discrete-valued X_i and S_i and continuous-valued gaussian noise when the query sequence has finite average power. The methods of proof are almost identical to those used in [20], though the framework has been modified considerably to correctly incorporate discrete-valued and continuous-valued X_i and S_i and continuous-valued Gaussian noise. The results are not direct consequences of Shannon's theorems, which hold for a) queries $\mathbf{X}$ that are deterministic functions of the desired variables $\mathbf{S}$, and b) queries corresponding to reliable codes (small error attacks with constant η_m), for which there is no analogy in this correspondence. The bulk of this correspondence shows that the channel capacity is an asymptotic limit on the rate of any small error inference attack when the queries have bounded average power. Hence, the inverse of channel capacity provides a tight asymptotic bound on η_m for any small error attack. Similar results have been shown for binary X in [20]; in this correspondence these results are shown to hold for binary, discrete and continuous RDP, and binary and discrete-valued X_i and S_i.

C. Some Channel Capacities

In this section, we present the channel capacities of some specific types of RDP for illustrative purposes. We denote the channel corresponding to the randomization ϕ by Φ, and its channel capacity by $\mathcal{C}(\Phi)$. The channel corresponding to n-ary RDP defined in Section I-A is denoted $\Phi_n(\rho)$, and the corresponding channel capacity by $\mathcal{C}(\Phi_n(\rho))$. Its value in bits is

$$\mathcal{C}(\Phi_n(\rho)) = \log_2 n + \rho\log_2\rho + (1-\rho)\log_2\left(\frac{1-\rho}{n-1}\right).$$

When the perturbation has a small bias, i.e., $\rho = \frac{1}{n} + \beta$ for small β, its capacity is determined by the second-order term of the Taylor expansion (zeroth- and first-order terms are zero)

$$\mathcal{C}\left(\Phi_n\left(\frac{1}{n}+\beta\right)\right) \approx \frac{n^2}{2(n-1)ln2}\times\beta^2, \quad \beta \text{ small.} \tag{3}$$

The channel corresponding to the continuous RDP with small signal to noise ratio $\frac{\sigma^2}{\sigma_n^2} = \beta^2$ is denoted $\Phi_{p_N}(\beta)$ and the corresponding channel capacity $\mathcal{C}(\Phi_{p_N}(\beta))$. Its value in bits is the maximum value of the mutual information between X and Y. This is the difference between the maximum differential entropy of Y when its average power is bounded by $\sigma^2+\sigma_{\mathrm{noise}}^2$, and the differential entropy of the noise pdf p_N. The value of the first is $\frac{1}{2}\log 2\pi e(\sigma^2+\sigma_n^2)$, and that of the second $\frac{1}{2}\log 2\pi e(\sigma_n^2)$ [5], hence, the capacity is

$$\mathcal{C}(\Phi_{p_N}(\beta)) = \frac{1}{2}\log(1+\beta^2) \approx \frac{\beta^2}{2ln2} \text{ for small } \beta. \tag{4}$$

IV. Our Results

In this section, we demonstrate an asymptotic lower bound on η_m for a small error inference attack. We use the methods of [20], which addressed only binary RDP, while we address binary, discrete-valued, and continuous-valued RDP. [19] implies that the bound is tight for binary RDP; it is easily shown to be tight for binary, discrete-valued, and continuous-valued RDP. To illustrate the rare attack where a deterministic relationship exists among the queries, we first provide an example and then proceed to prove our results.

A. An Example of an Inference Attack That Corresponds to a Channel Code

Example 6: Consider a database of records of all residents of a county. From each record, consider the set of the following bits.

X_1. Tuberculin Skin Test: "1" represents "Positive," and "0" "Negative";

X_2. Chest X-Ray: "1" represents "Positive for TB" and "0" represents "Negative for TB";

X_3. Lab Test: "1" represents that a laboratory test was performed, and "0" that it was not.

A lab test follows only if the other two tests were both positive, and not otherwise. Hence

$$X_3 = X_1 \text{ AND } X_2 \tag{5}$$

for all records.

Suppose $\mathcal{B}$ chooses as desired bits $\mathbf{S} = (X_1, X_2)$ for all records, and designs an overdetermined query sequence by also requesting X_3. Without RDP, $\mathcal{B}$ would not need to do so; with RDP, X_3 serves as a check for the values of X_1 and X_2, or, in the communication channel framework, as an error-detecting symbol. The queries $\mathbf{X} = (X_1, X_2, X_3)$ may be thought of as the code bits. In general, one can have an overdetermined sequence of m queries whose values are completely determined by $\mathbf{S}$ - through a set of m equations known to be satisfied by $\mathbf{S}$ and $\mathbf{X}$. Equation (5) is one such equation.

B. Existence of Efficient Inference Attacks

Attacks that correspond to channel codes satisfy the channel coding theorem if the value of η_m is held constant.

Theorem 1: For an RDP $\Phi, \forall\Lambda > \frac{1}{\mathcal{C}(\Phi)}$, there exists a small error inference attack on Φ with $\eta_m = \Lambda, \forall m$.

Proof: Follows from the channel coding theorem [16].

Attacks that correspond to codes are those where the queries $\mathbf{X}$ are deterministic functions of the desired bits $\mathbf{S}$, as in Example 6. The inference attack does not, in general (see Examples 2 and 4, and Definition 1), however, consist of queries $\mathbf{X}$ that are functions of $\mathbf{S}$. Nor do inference attacks require constant η_m as m increases. We now show a result on the efficiency of all small error inference attacks, not just those that correspond to reliable channel codes.

C. An Asymptotic Bound on η_m for All Zero-Error Inference Attacks

[20] shows that the inverse of the channel capacity is also an *asymptotic* lower bound on η_m of the small error inference attack (which is more general than a reliable binary code) for binary RDP. It does so by modifying the proof of the converse of the channel coding theorem using Fano's inequality ([5, p. 205])—the main ingredient for demonstrating channel capacity as a bound on the rate of a code. Thus, Fano's inequality and [20] provide the *asymptotic lower bound* on η_m, and the result in [19] and the channel coding theorem provide the *existence* of small error inference attacks that achieve it. Theorem 2 presents the proof from [20] for binary RDP, which also holds for discrete-valued RDP in our general framework. The proof for continuous-valued RDP simply involves replacing the entropy by the differential entropy.

Theorem 2: Given an RDP Φ, $\frac{1}{\mathcal{C}(\Phi)}$ is an asymptotic lower bound on η_m for a small error inference attack, i.e.

$$\lim_{m\to\infty} \omega_m = 0$$

$\Rightarrow$ $\exists$ strictly increasing sequence $\{\Lambda_m\}_{m=1}^{\infty}$ such that

$$\eta_i \ge \Lambda_m \quad \forall i \ge m$$

and

$$\lim_{m\to\infty} \Lambda_m = \frac{1}{\mathcal{C}(\Phi)}.$$

Proof: The proof is similar to the proof of the converse of the channel coding theorem ([5, p. 199 and p. 244]), except for two differences: a) in an inference attack, queries $\mathbf{X}$ are not necessarily a function of $\mathbf{S}$, and b) inference attacks do not have constant η_m as m increases.

Assume $\lim_{m\to\infty} \omega_m = 0$, i.e., the attack is small error. Then $\lim_{m\to\infty} E_m = 0$ where E_m is the average probability of error over all values $\mathbf{x}$ of r.v. $\mathbf{X}$. Then

$$\mathcal{H}(S_i) = \mathcal{H}(S_i|\phi(X_1),\phi(X_2),\dots\phi(X_m)) + \mathcal{I} \tag{6}$$

where

$$\mathcal{I} = I(S_i;\phi(X_1),\phi(X_2),\dots\phi(X_m)). \tag{7}$$

From Fano's inequality: ([5, p. 205, eq. (8.95)])

$$\mathcal{H}(S_i|\phi(X_1),\phi(X_2),\dots\phi(x_m)) \le 1 + E_m \log_2 M. \tag{8}$$

From (8), and the definition of $\eta_m (\eta_m = \frac{m}{\log_2 M})$

$$\mathcal{H}(S_i|\phi(X_1),\phi(X_2),\dots\phi(x_m)) \le 1 + \frac{E_m m}{\eta_m}. \tag{9}$$

Further, when the noise is discrete-valued

$$\begin{aligned}\mathcal{I} = {} & \mathcal{H}(\phi(X_1),\phi(X_2),\dots\phi(X_m)) \\ & - \mathcal{H}(\phi(X_1),\phi(X_2),\dots\phi(X_m)\,|\,S_i)\end{aligned} \tag{10}$$

and

$$\begin{aligned}&\mathcal{H}(\phi(X_1),\phi(X_2),\dots\phi(X_m)\,|\,S_i)\\ &\quad = \sum_i \mathcal{H}(\phi(X_i)\,|\,\phi(X_1),\phi(X_2),\dots\phi(x_{i-1}),S_i)\\ &\quad \ge \sum_i \mathcal{H}(\phi(X_i)\,|\,\phi(X_1),\phi(X_2),\dots\phi(x_{i-1}),S_i,X_i)\\ &\quad = \sum_i \mathcal{H}(\phi(X_i)\,|\,X_i)\end{aligned} \tag{11}$$

and

$$\mathcal{H}(\phi(X_1),\phi(X_2),\dots\phi(X_m)) \le \sum_i \mathcal{H}(\phi(X_i)). \tag{12}$$

Equations (10)–(12) give

$$\begin{aligned}\mathcal{I} &\le \sum_i \mathcal{H}(\phi(X_i)) - \sum_i \mathcal{H}(\phi(X_i)\,|\,X_i)\\ &= \sum_i I(X_i;\phi(X_i))\\ &\le m\mathcal{C}(\Phi).\end{aligned} \tag{13}$$

From (9) and (13), and the fact that the bound of (6) holds for the maximum value of its right-hand side (RHS) $\mathcal{H}(S_i) \le \log_2 M$

$$\frac{m}{\eta_m} \le 1 + \frac{E_m m}{\eta_m} + m\mathcal{C}(\Phi).$$

Hence

$$\eta_m \ge \frac{1 - E_m}{\frac{1}{m} + \mathcal{C}(\Phi)} = \Lambda_m$$

where $\{\Lambda_m\}_{m=1}^{\infty}$ is strictly increasing if $\{E_m\}_{m=1}^{\infty}$ is nonincreasing, and

$$\lim_{m\to\infty} \Lambda_m = \frac{1}{\mathcal{C}(\Phi)}.$$

The proof for continuous-valued noise follows easily by replacing entropy with differential entropy.

D. Asymptotic Bounds and the Privacy of RDP

The values of S_i are not necessarily uniformly distributed, and, hence, the entropy of $\mathbf{S}$ is not the maximum possible. From the source coding theorem, if the entropy of $\mathbf{S}$ is $\mathcal{H}(\mathbf{S})$, then $\mathbf{S}$ is represented by $\mathcal{H}(\mathbf{S})$ bits on average (over many records). This observation can be combined with a reasoning similar to that in Theorem 2 to obtain a result similar to that of the source-channel coding theorem, except, as with Theorem 2, inference attacks are not of constant η_m, and do not consist of queries $\mathbf{X}$ that are deterministic combinations of $\mathbf{S}$. Again, we derive the asymptotic lower bound, and Shannon's results show it is tight for query sequences that are deterministic functions of $\mathbf{S}$. As with Theorem 2, the proof for discrete-valued RDP is very similar to that for binary RDP in [20]. The proof for continuous RDP is the same as that for discrete-valued RDP.

Theorem 3: The tight asymptotic lower bound on the query complexity, on average, per record, for a small error inference attack, is $\frac{\mathcal{H}(\mathbf{S})}{\mathcal{C}(\Phi)}$ if $\mathcal{H}(\mathbf{S})$ and $\mathcal{C}(\Phi)$ are measured in the same units, and the record sequence is stationary. That is, if the number of records is N_r, and γ_m the number of queries per record

$$\lim_{m\to\infty} \omega_m \to 0$$

$\Rightarrow$ $\exists$ nondecreasing sequence $\{\Gamma_m\}$ such that

$$\gamma_m \ge \Gamma_m \forall i \ge m$$

and

$$\lim_{N_r\to\infty} \Gamma_m = \frac{\mathcal{H}(\mathbf{S})}{\mathcal{C}(\Phi)}.$$

Proof: $\frac{\mathcal{H}(\mathbf{S})}{\mathcal{C}(\Phi)}$ is an asymptotic lower bound. Assume the existence of a small error attack with asymptotic query sequence length $K = \frac{\mathcal{H}(\mathbf{S})}{\mathcal{C}(\Phi)} - \Delta$ per record on average, $\Delta > 0$. This means that, given $\epsilon, \delta > 0$, a query sequence of length at most $m = N_r(K + \epsilon)$ for N_r records, N_r large enough, can result in a probability of error at most δ. By

Theorem 2, for any given ν, η_m for the attack must be at least $\frac{1}{\mathcal{C}(\Phi)} - \nu$, for large enough m, and hence the length of $\mathbf{S}$, $\frac{m}{\eta_m}$, at most

$$\frac{N_r(K+\epsilon)}{(\frac{1}{\mathcal{C}(\Phi)} - \nu)} = \frac{N_r(\mathcal{H}(\mathbf{S}) - \mathcal{C}(\Phi)(\Delta - \epsilon))}{1 - \nu\mathcal{C}}$$

i.e., each record is represented, on average, by a number of symbols strictly smaller than the record entropy for small enough ϵ, δ, ν. This violates Shannon's source coding theorem ([5, p. 89, Theorem 5.4.2]) and [16]. $\frac{\mathcal{H}(\mathbf{S})}{\mathcal{C}(\Phi)}$ can be achieved from the above (i.e., tightness): straightforward from Shannon's source–channel coding theorem [5].

Corollary 1: The tight asymptotic lower bounds on the value of η_m for a small error inference attack on $\Phi_n(\frac{1}{n} \pm \beta)$, and $\Phi_{p_N}(\beta)$ are $\frac{2(n-1)ln2}{n^2\beta^2}$ and $\frac{2ln2}{\beta^2}$, respectively.

Proof: The result follows from Theorems 1–3 and (3) and (4).

Corollary 2: The privacy of Φ is $\frac{1}{\mathcal{C}(\Phi)}$.

Proof: Follows from Theorems 1–3 and Definition 5.

Corollary 3: The privacies of $\Phi_{\mathcal{B}}(0.5\pm\beta)$, $\Phi_n(\frac{1}{n}\pm\beta)$, and $\Phi_{p_N}(\beta)$ are all $\Theta(\frac{1}{\beta^2})$.

Proof: Follows from Corollary 1 and Definition 5.

E. Underestimating η_m

The value of η_m is not necessarily known to $\mathcal{A}$, because $\mathcal{A}$ is not even aware of $\mathbf{S}$, the attributes of interest to $\mathcal{B}$. $\mathcal{A}$ only knows the queries $\mathbf{X}$, and the RDP parameters: $P_{Y\,|\,X}$ or $p_{Y\,|\,X}$. Given Theorem 3, the parameters can be chosen so as to limit the asymptotic value of η_m of a small error attack. $\mathcal{A}$ can think of a single query response as providing at most $\mathcal{C}(\Phi)$ bits of information. We now provide two examples to illustrate what is *not* implied by our result.

Example 7: $\mathcal{B}$ uses a query sequence of length m per record, concentrating on k-tuples of desired bits (attributes) for an entire population. Every k-tuple is possible. $\mathcal{B}$ is interested in a small error attack with constant η_m (the equivalent of a reliable code). This is only possible if $\eta_m = \frac{m}{k} \geq \frac{1}{\mathcal{C}}$, or $\mathcal{C} \geq \frac{k}{m}$.

We now describe cases where our result does not imply that $\mathcal{C}$ needs to be as large as $\frac{k}{m}$, even though the number of queries is m, and the number of variables of interest k. Hence, in the following cases, small error attacks are possible for RDP with capacity smaller than $\frac{k}{m}$.

Case 1) $\mathcal{B}$ targets a few types of individuals, those who would respond in a certain way to the queries. This limits the number of true responses possible, i.e., $M < 2^k$, or $\log_2 M < k$. Then, $\eta_m = \frac{m}{\log_2 M} \geq \frac{1}{\mathcal{C}}$ implies that $\mathcal{C} \geq \frac{\log_2 M}{m} < \frac{k}{m}$.

Case 2) Suppose $\mathcal{B}$ is interested only in a subset of the target attributes, in, say, $k_1 < k$ bits. Again, the attack requires $\mathcal{C} \geq \frac{k_1}{m} < \frac{k}{m}$.

Case 3) Suppose some bits of $\mathbf{S}$ are completely determined from some others: say $k_1 < k$ bits are independent. Again, the attack requires $\mathcal{C} \geq \frac{k_1}{m} < \frac{k}{m}$.

Our final example is one in which there is a probabilistic relationship among the queries but not a deterministic one.

Example 8: Consider a last example [15] of a set of k desired bits, each highly correlated with the first. Assume that none of the desired bits is completely determined by any of the others, i.e., $\mathcal{H}(S_i\,|\,S_1, S_2, \ldots S_{i-1}, S_{i+1}, \ldots S_k) \neq 0 \forall i$. Assume that each bit is queried, i.e., that $m = k$ and $\mathbf{S} = \mathbf{X}$. Assume also that, for a given value of k, all k-tuples are possible. However, as k increases, a sequence of the desired bits $\{X_i\}_{i=1}^{k}$, is "typical", i.e., most bits are equal to the first. The value of η_m is unity, and the attack cannot be small error for $\rho \notin \{0, 1\}$ (i.e., for channel capacity lower than unity). However, $\mathcal{B}$ can learn a lot from the responses. What $\mathcal{B}$ cannot do is drive error to zero without increasing η_m.

F. Discussion

In statistical databases, it is typically assumed that a larger number of queries (per attribute desired) is required for a lower error. Our work shows that, while the total number of queries needs to increase to reduce error, the number of queries per bit of entropy need not.

As our measure is closely related to the conditional privacy loss measure of [2, eq. (1)] ($\mathcal{C} = \max_{P_X} I(\mathbf{X}; \mathbf{Y})$), it will also fail to detect privacy breaches that occur with low probability, which would be detected by the worst-case information measure of [6, eq. (2)]. This is because such breaches would occur too rarely to influence the asymptotic behavior of inference attacks. Each individual will see the complete distribution of $Z(y)$ and not just a single value, and the effect of privacy breaches will be more distributed throughout the entire population and not restricted to a few individuals, and the average will indeed represent the effect on a single individual's record.

Though our results follow very easily from classical results in information theory and coding, our view of RDP as a channel has one important point of difference from the view of a channel in communication theory. The goal of communication theory is to increase information transfer over a channel given certain constraints. The goal of a privacy protection technique is to decrease the information transfer given certain constraints (such as the desired accuracy of the statistics obtained using perturbed data points). Because of this, $\mathcal{A}$ would be interested in channels with small capacity.[3] On the other hand, $\mathcal{B}$ is interested in the efficient transfer of bits, typically over a channel with small capacity, and a number of the constructive results from the theory of coding are of interest to $\mathcal{B}$.

V. Conclusion

We have treated RDP as a channel, and channel codes as efficient attacks, to develop a framework for the study of the most general inference attack on RDP. We have demonstrated the existence of inference attacks with zero asymptotic error for all costs greater than the inverse of the RDP channel capacity. We are not aware of any other work that a) uses the channel coding theorem to study the privacy properties of data perturbation, or b) connects attacks on RDP to channel codes.

Appendix

$\mathcal{A}$	Database.
$\mathcal{B}$	Data collector.
Q_i, C, C_i	Database variable.
$\mathcal{Q}$	set of database variables
X, X_i	a single queried bit
$\mathbf{X}$	Query sequence.
m	Number of queries or length of $\mathbf{X}$.
S_i	Bit desired by $\mathcal{B}$.
$\mathbf{S}$	Bits desired by $\mathcal{B}$.
k	Number of desired bits or length of $\mathbf{S}$
Y	$= \phi(X)$, single perturbed response to query X.
ρ	Probability of correct response, discrete-valued RDP.
$P_{Y\,|\,X}, p_{Y\,|\,X}$	An *a posteriori* pdf of protocol/channel.

[3] Our work shows that this is a necessary condition for privacy, though, depending on how one defines privacy, it may not be a sufficient condition.

p_N	pdf of noise, continuous-valued RDP.
σ^2	Maximum average power in query sequence.
σ_n^2	Maximum average power in noise.
$\mathcal{P}$	Conditional privacy loss.
I_w	Worst-case information.
$g(.)$	ML estimate function.
$\hat{S}_i$	ML estimate of S_i.
$\hat{\mathbf{S}}$	ML estimate of $\mathbf{S}$.
ω_m	Maximum probability of error.
$\mathcal{M}$	Alphabet for X_i.
M	Number of possible values of $\mathbf{X}$.
η_m	Number of queries per bit determined.
Λ_m	Lower bound on η_i $\forall i \geq m$.
$\eta_{\min}$	$\lim_{m\to\infty} \Lambda_m$.
Φ	RDP channel.
$\mathcal{C}(\Phi)$	Capacity of Φ.
β	Small bias of discrete RDP.
E_m	Average probability of error.
$\mathcal{H}(.)$	Entropy.
γ_m	Query complexity per record.
Γ_m	Lower bound on γ_m $\forall i \geq m$.
N_r	Number of records.

ACKNOWLEDGMENT

The author would like to acknowledge the substantial contributions of the anonymous reviewers, including a pointer to [6].

REFERENCES

[1] N. R. Adam and J. C. Wortman, "Security-control methods for statistical databases: A comparative study," *ACM Comput. Surv.*, vol. 21, no. 4, pp. 515–556, Dec. 1989.

[2] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Syst.*, 2001.

[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Conf. Manage. Data*, 2000.

[4] M. K. C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1026–1037, Sep. 2004.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[6] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Syst.*, 2003.

[7] C. Farkas and S. Jajodia, "The inference problem: A survey," *ACM SIGKDD Explorat. Newsl.*, vol. 4, no. 2, pp. 6–11, Dec. 2003.

[8] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. 1st Int. Conf. Mobile Syst., Applicat. Services (MobiSys'03)*, 2003, pp. 31–42.

[9] M. Kantarcioglu, J. Jin, and C. Clifton, "When do data mining results violate privacy?," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2004, pp. 599–604.

[10] H. Kargupta, H. Dutta, S. Datta, and K. Sivakumar, "Privacy preserving data mining and random perturbation," in *Proc. Workshop on Privacy in the Electron. Soc. (WPES'03)*, 2003.

[11] D. Lambert, "Measures of disclosure risk and harm," *J. Official Stat.*, vol. 9, pp. 313–331, 1993.

[12] S. Merugu and J. Ghosh, "Privacy-preserving distributed clustering using generative models," in *Proc. 3rd IEEE Int. Conf. Data Mining (ICDM)*, 2003.

[13] K. Muralidhar and R. Sarathy, "Security of random data perturbation methods," *ACM Trans. Database Syst. (TODS)*, vol. 24, no. 4, pp. 487–493, Dec. 1999.

[14] S. Rizvi and J. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. 28th Conf. Very Large Data Base (VLDB'02)*, 2002.

[15] G. Seroussi, Personal Communication 2002.

[16] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul. 1948.

[17] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 434–447, 2004.

[18] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Rec.*, vol. 33, no. 1, Mar. 2004.

[19] P. L. Vora, "The channel coding theorem and the security of binary randomization," in *Proc. 2003 IEEE Int. Symp. Inf. Theory*, Yokohama, Japan, Jun. 30–Jul. 4 2003, p. 306.

[20] P. L. Vora, A. Canteaut and K. Viswanathan, Eds., "Information theory and the security of binary data perturbation," in *Progress in Cryptol. —INDOCRYPT 2004: 5th Int. Conf. Cryptol. India*, Chennai, India, Dec. 20–22, 2004, pp. 136–147.

Single-Symbol ML Decodable Distributed STBCs for Cooperative Networks

Zhihang Yi and Il-Min Kim, *Senior Member, IEEE*

***Abstract*—In this correspondence, the distributed orthogonal space–time block codes (DOSTBCs), which achieve the single-symbol maximum likelihood (ML) decodability and full diversity order, are first considered. However, systematic construction of the DOSTBCs is very hard, since the noise covariance matrix is not diagonal in general. Thus, some special DOSTBCs, which have diagonal noise covariance matrices at the destination terminal, are investigated. These codes are referred to as the row-monomial DOSTBCs. An upper bound of the data-rate of the row-monomial DOSTBC is derived and it is approximately twice higher than that of the repetition-based cooperative strategy. Furthermore, systematic construction methods of the row-monomial DOSTBCs achieving the upper bound of the data-rate are developed when the number of relays and/or the number of information-bearing symbols are even.**

***Index Terms*—Cooperative networks, distributed space–time block codes, diversity, single-symbol maximum likelihood (ML) decoding.**

I. INTRODUCTION

It is well known that relay terminal cooperation can improve the performance of a wireless network considerably [1]–[4]. A source terminal, several relay terminals, and a destination terminal constitute a cooperative network, where the relay terminals relay the signals from

Manuscript received September 8, 2006; revised March 21, 2007. This work was supported in part by the Ubiquitous Computing and Network (UCN) Project and by the Ministry of Information and Communication (MIC) 21st Century Frontier R&D Program in Korea.

The authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: 3zy1@qlink.queensu.ca; ilmin.kim@queensu.ca).

Communicated by R. R. Müller, Associate Editor for Communications.

Color version of Figure 1 is available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2007.901177