Information Theory and the Security of Binary Data Perturbation

Poorvi L. Vora

George Washington University, Washington DC 20052
poorvi@gwu.edu

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. It has recently gained popularity as a privacy technique in data mining. To our knowledge, attacks on binary RDP have not been completely characterized, its security has not been analyzed from a complexity-theoretic or information-theoretic perspective, and there is no privacy measure of binary RDP that is related to the complexity of an attack. We characterize all inference attacks on binary RDP, and show that if it is possible to reduce estimation error indefinitely, a finite number of queries per bit of entropy is enough to do so. We define this finite number as the privacy measure of the binary RDP.

1 Introduction

The general problem solved by random data perturbation (RDP) is that of providing statistics while protecting individual values. This is done by adding noise to the individual values. The larger the probabilistic perturbation of the data, the more privacy provided to the individual values, and the less accurate the statistics. RDP has been in use for about twenty years in statistical database security [1, 12], and has recently been proposed as a means of personal privacy protection in data mining applications [2, 3]. In this paper, we analyze the amount of privacy provided by binary RDP.

The purpose of a statistical database is to provide statistics to researchers while keeping individual values "private". For example, a health database would keep "private" whether individual X had Hepatitis A or not, but would reveal how many members in a community had Hepatitis A. The general technical problem is as follows:

- Database A contains two (possibly intersecting) sets of binary variables: $\mathcal{Q} = \{q_1, q_2, ...q_i, ...\}$ (queryable bits) and $\mathcal{S} = \{s_1, s_2, ...s_i, ...\}$ (sensitive bits).
- Data collector B queries the value of $f(a_1, a_2, ..., a_k)_{a_i \in \mathcal{Q}} = X \in \{0, 1\}$, for *any* f. In particular, B can query combinations of queryable values across records such as "the most common gender among records 1, 2 and 3".

A. Canteaut and K. Viswanathan (Eds.): INDOCRYPT 2004, LNCS 3348, pp. 134–144, 2004.

- $\mathcal{Q}$ and $\mathcal{S}$ are probabilistically-related, i.e. the mutual information between the two (the change in uncertainty in one on knowing values in the other) is non-zero: $\mathcal{I}(\mathcal{S};\mathcal{Q}) \neq 0$.
- The bits in $\mathcal{S}$ are to be "protected".

One approach is to compute the value of $f(a_1, a_2, ...a_k)$ so that no other information is revealed. This can be done using secure multiparty computation [9]. However, secure multiparty computation cannot prevent *inference attacks* [6, 12], which involve the determination of information on bits in $\mathcal{S}$ from several queried values $A_i = f_i(a_1, a_2, ...a_k)$, $i = 1, 2, ...n$. It is not straightforward to recognize such attacks.

Example 1. The general inference attack. Consider

s_1 "gender"
s_2 "Over 40"
q_1 "Losing Calcium"
q_2 "Balding"
q_3 "Greying"
q_4 "Gaining weight"

Suppose B wishes to determine bits s_1 and s_2. To do so, B may query functions of the bits q_1, q_2, q_3, q_4, which would reveal information about the sensitive bits, but would not determine them completely. For example, women over 40 are more likely to be losing Calcium than any of the three other categories. Similarly, men over 40 are almost the only category balding. However, it is possible for a man over 40 to have the same responses as a man under 40.

The RDP of $A_i = f_i(a_1, a_2, ...a_k)$ before revealing the values, whether computed using trusted multiparty computation or not, makes the task of inference more difficult. Binary RDP proceeds as follows:

> B requests bit X from A and receives the variable $\phi(X) = Y \in \{0, 1\}$ generated according to $P(Y|X)$,
>
> $$P(Y|X) = \begin{cases} \rho & Y = X \\ 1 - \rho & Y \neq X \end{cases}$$
>
> i.e. B receives the requested bit flipped with probability $1 - \rho$. There are no conditions on the amount of information A has about what B received, i.e. we do not consider private information retrieval.

The simplest attack on RDP is the repeated query attack, where B repeatedly asks for the same bit $x \in \mathcal{Q} \cap \mathcal{S}$ and guesses the correct value of x to be the one received most often (assuming $\rho > \frac{1}{2}$). Clearly, the estimation error can be decreased without bound by increasing queries without bound, i.e. if ω_m is the probability of error using m repeated queries of a bit, $\lim_{m\to\infty} \omega_m = 0$.

However, this is the best B can do with repeated queries, i.e. if η_m is the number of queries per bit determined, η_m increases indefinitely if ω_m is not bounded below.

$$\lim_{m\to\infty} \omega_m = 0 \Rightarrow \lim_{m\to\infty} \eta_m = \infty \tag{1}$$

There are a number of attacks other than a repeated query attack, which is very recognizable. However, because these are not well characterized, it is typically assumed that expression (1) represents a best-case scenario for B. In [16] it is stated that, if $\mathcal{C}$ is the channel capacity of the protocol viewed as a communication channel, attacks in which

$$\lim_{m\to\infty} \omega_m = 0 \; for \; \eta_m = \frac{1}{\mathcal{C}} \tag{2}$$

exist, i.e. that inference attacks can be more efficient than the repeated query attack - if the queries $\mathbf{x}$ are functions of the bits B wants to determine, and if B and A are willing to participate in a large enough number of queries.

In this paper we show that expression (2) *is* the best an attacker can do, i.e. that

$$\lim_{m\to\infty} \omega_m = 0 \Rightarrow \lim_{m\to\infty} \eta_m \geq \frac{1}{\mathcal{C}}$$

for all inference attacks if $\lim_{m\to\infty} \eta_m$ exists. Note that the lower bound obviously does not hold for attacks that do not seek to reduce error arbitrarily. We define the asymptotic lower bound on η_m as the privacy measure of the protocol; it is the inverse of the channel capacity of the protocol viewed as a communication channel. Note that, we use "asymptotic" as used by mathematicians, to mean: "in the limit".

When the protocol has a small bias β (i.e. each bit is flipped with probability $0.5 + \beta$, β small), Chernoff-type bounds [11, 10] provide estimates of the query complexity of a repeated query attack. For example, from the Chernoff bound:

$$m = \eta_m = \frac{[ln(\frac{2}{\delta})]}{0.38\beta^2} \Rightarrow \omega_m \;\leq\; \delta$$

We show that η_m for an inference attack can be *independent of* δ, i.e. while $m \to \infty$, η_m can be finite, though bounded below. In particular we show that η_m is $\Theta(\frac{1}{\beta^2})$.

Our main contributions are: (a) the framework we have used to study the security of binary RDP, and the corresponding definitions and associations with information theory and coding; (b) a general characterization of inference attacks, and (c) the use of our framework in deriving a very general efficiency result that changes some of the view of the efficiency of inference attacks.

The paper is organized as follows. In section 2 we present a short review of existing work, and in section 3, definitions motivated by the statistical database security problem. Section 4 presents our results with proofs. The conclusions are presented in Section 5.

2 Related Work

The database community has measures of the privacy of randomization [8, 3, 2]; these are, however, not motivated by a security analysis. The security analyses that do exist [12] focus on the variance of the estimation error. [2] proposes the use of the differential mutual information between the original and perturbed continuous-valued data points as a measure of "conditional privacy loss", which inspires our measure. The mutual information between two variables is the change in uncertainty of one on knowing the other. Thus the measure of conditional privacy loss has some useful properties: (a) it addresses the *change* due to a protocol instance, and (b) because it is based on entropy, it distinguishes among situations where the two possibilities are almost equally likely and situations where this is not so. The measure does, however, depend on the original pdf, and not only on protocol parameters. Our privacy measure, the inverse of the protocol channel capacity, is closely related to this measure, but improves on it by being independent of the input pdf (channel capacity is the maximum value of the mutual information, taken over all possible input pdfs). Unlike [2], our work also provides a connection between our privacy measure and the complexity of certain types of attacks.

3 Definitions

In this section we provide the definitions we shall need to prove our results. We provide a list of symbols in the appendix.

Consider a query sequence $\mathbf{x}$ of length m. The number of possible values of the true responses need not be 2^m, because certain bit 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. Clearly, a "most efficient" query sequence would use exactly $log_2 M$ bits to distinguish among the M values, but most effective query sequences would want to correct for the RDP and would hence consist of more that $log_2 M$ queries.

Definition 1. *The* query complexity per bit*, of query sequence* $\mathbf{x}$ *of length* m*, as a means of distinguishing among* M *possible values of* $\mathbf{x}$ *is* $\eta_m = \frac{m}{log_2 M}$.

We define the most general inference attack, such as the one of example 1, next.

Definition 2. *An* inference attack *is a set of queries* $\mathbf{x}$ *such that* $\mathbf{x}$ *and the set of sensitive bits* $\mathcal{S}$ *are not independent, i.e.* $I(\mathcal{S}; \mathbf{x}) \neq 0$.

The definition is intentionally broad, as we show a lower bound on the query complexity per bit for an inference attack for which $\lim_{m\to\infty} \omega_m = 0$. The definition also assumes nothing about the relationship between queried bit x_i and previously received responses: $\phi(x_1), \phi(x_2), ...\phi(x_{i-1})$, and, hence includes adaptive inference attacks.

B cannot do any better in reducing the uncertainty of $\mathcal{S}$ than is possible through accurate knowledge of all of $\mathcal{Q}$ and unlimited computing power. Assume, wlog, that B wishes to determine the k bits $\mathbf{p} = (p_1, p_2, ...p_k) = \{g_i(a_1, a_2, ..a_j...)_{a_j \in \mathcal{Q}}\}_{i=1}^{k}$ from each record in database A. The RDP limits B in determining $\mathbf{p}$ accurately, but does not affect the uncertainty reduction in $\mathcal{S}$ from complete knowledge of $\mathbf{p}$. In evaluating the RDP, hence, we focus on the accurate determination of $\mathbf{p}$. We denote the entropy of $\mathbf{p}$ as *queryable entropy*, (that which can be reduced to zero through queries if there is no RDP). Contrast this to the entropy of bits in $\mathcal{S}$, which, in general, cannot be reduced to zero through queries of functions of queryable bits even if there were no RDP.

The maximum probability of estimation error[1], denoted ω_m, is a measure of the success of query sequence $\mathbf{x}$, of length m, in estimating $\mathbf{p}$. As the value of m grows, it is reasonable to expect the error to reduce, or, at least, not increase. We define attacks in which asymptotic error is zero as *small error* attacks.

Definition 3. *A* small error inference attack *is one in which* $\lim_{m\to\infty}\omega_m = 0$.

Clearly, error and query complexity are related, and a lower error could require a higher query complexity per queryable bit. An RDP that forces a higher query complexity to reduce error is better from the privacy point of view. We propose that the measure of the privacy of binary RDP be the minimum value of the query complexity per bit of queryable entropy required for a small error attack.

Definition 4. *The* privacy *of binary RDP is the (tightest) asymptotic lower bound on the query complexity, on average, per bit of queryable entropy, for a small error attack.*

We now review some definitions from information theory necessary for our results.

Definition 5. *[5] A* communication channel *is a triplet of the following: a set of input variables, $\mathcal{X}$, a set of output variables, $\mathcal{Y}$, and a* a posteriori pdf, *$P(Y|X)$, and is denoted $(\mathcal{X}, P(Y|X), \mathcal{Y})$.*

We denote the channel corresponding to a protocol by Φ, and the channel corresponding to binary RDP with probability of lie $1-\rho$ by $\Phi_{\mathcal{B}}(1-\rho)$.

Definition 6. *The* channel capacity *of protocol Φ is the maximum decrease in entropy of variable X due to the protocol, and is denoted $\mathcal{C}(\Phi)$.*

The channel capacity of the binary symmetric protocol with probability of a lie $1-\rho$ is

$$\mathcal{C}(\Phi_{\mathcal{B}}(1-\rho)) = 1 - \mathcal{H}(\rho) = 1 + \rho log_2\rho + (1-\rho)log_2(1-\rho)$$

bits, where $\mathcal{H}(\rho)$ is the entropy of the binary variable with ρ being the probability of one of its values. When the protocol has a small bias, i.e. $\rho = 0.5+\beta$ for small β, its capacity is determined by the second order term of the Taylor expansion (zeroth and first order terms are zero):

$$\mathcal{C}(\Phi_{\mathcal{B}}(0.5 \pm \beta)) = \frac{2\beta^2}{ln2}, \beta\ small \tag{3}$$

[1] The estimation error calculation assumes a maximum likelihood estimation.

4 Our Results

We wish to determine the privacy of binary RDP. To do so, we demonstrate an asymptotic lower bound on the query complexity, per bit of queryable entropy, for a zero error inference attack. [16] implies that the bound is tight. However, attacks that achieve the bound might be recognizable.

We approach the problem by viewing binary RDP as a communication channel as in [16]. The analogy with communication over a channel is as follows: the protocol is a channel and **p** a message. The channel coding ("Shannon's second") theorem [14, 5] provides a tight upper bound of channel capacity on the inverse of η_m for a zero error attack - if each query is a function of **p**, and η_m is constant as m increases. Hence, when the query sequence **x** is a function of **p**, inference attacks are channel codes; η_m^{-1} are the rates of the codes; when such attacks are zero-error with constant $\eta_m = \eta$, the inverse of channel capacity is the minimum value of η, achieved by attacks that correspond to Shannon codes.

The most general inference attack (see example 1, section 1 and definition 2, section 3) is not one in which the query sequence is a function of the required values **p**. Nor does an inference attack require constant η_m as m increases. By modifying the proof of the converse of the channel coding theorem using Fano's inequality [5–pg. 205] - the main ingredient for demonstrating channel capacity as a bound on the rate of a code - we show that the tight asymptotic lower bound on the query complexity per bit for the (more general) small error inference attack is also the inverse of the channel capacity of the protocol. Fano's inequality provides the *asymptotic lower bound* on η_m, and the result in [16] and the channel coding theorem provide the *existence* of zero error inference attacks that achieve it.

Theorem 1. *Given a binary RDP Φ, an asymptotic lower bound on η_m, for a small error inference attack, is $\frac{1}{\mathcal{C}(\Phi)}$. More formally,*

$$\lim_{m\to\infty} \omega_m = 0 \Rightarrow \exists\ \{\Lambda_m\}_{m=1}^{\infty} \text{ such that } \eta_i \geq \Lambda_m \forall i \geq m \text{ and } \lim_{m\to\infty} \Lambda_m = \frac{1}{\mathcal{C}(\Phi)}$$

Proof. The proof is similar to the proof of the converse of the channel coding theorem [5], except for two differences: (a) in an inference attack, queries **x** are not necessarily a function of bits required **p**, 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. Consider the case when the values of $\mathbf{p}_m$ are equally likely. Then,

$$log_2 M = \mathcal{H}(\mathbf{p}_m) = \mathcal{H}(\mathbf{p}_m|\phi(x_1), \phi(x_2), ...\phi(x_m)) + I(\mathbf{p}_m; \phi(x_1), \phi(x_2), ...\phi(x_m))$$

From equation (8.95) (Fano's inequality), [5–pg. 205],

$$\mathcal{H}(\mathbf{p}_m|\phi(x_1), \phi(x_2), ...\phi(x_m)) \leq 1 + E_m log_2 M$$

and hence,

$$log_2 M \leq 1 + E_m log_2 M + I(\mathbf{p}_m; \phi(x_1), \phi(x_2), ...\phi(x_m)) \tag{4}$$

Further,

$$\begin{aligned}
&I(\mathbf{p}_m; \phi(x_1), \phi(x_2), ...\phi(x_m))\\
&= \mathcal{H}(\phi(x_1), \phi(x_2), ...\phi(x_m)) - \mathcal{H}(\phi(x_1), \phi(x_2), ...\phi(x_m)|\mathbf{p}_m)\\
&= \mathcal{H}(\phi(x_1), \phi(x_2), ...\phi(x_m)) - \textstyle\sum_i \mathcal{H}(\phi(x_i)|\phi(x_1), \phi(x_2), ...\phi(x_{i-1}), \mathbf{p}_m)\\
&\leq \mathcal{H}(\phi(x_1), \phi(x_2), ...\phi(x_n)) - \textstyle\sum_i \mathcal{H}(\phi(x_i)|\phi(x_1), \phi(x_2), ...\phi(x_{i-1}), \mathbf{p}_m, x_i)\\
&= \mathcal{H}(\phi(x_1), \phi(x_2), ...\phi(x_m)) - \textstyle\sum_i \mathcal{H}(\phi(x_i)|x_i)\\
&\leq \textstyle\sum_i \mathcal{H}(\phi(x_i)) - \sum_i \mathcal{H}(\phi(x_i)|x_i)\\
&= \textstyle\sum_i I(x_i; \phi(x_i))\\
&\leq m\mathcal{C}(\Phi)
\end{aligned}$$

From equation (4),

$$log_2 M \leq 1 + E_m log_2 M + m\mathcal{C}(\Phi)$$

Hence,

$$\eta_m = \frac{m}{log_2 M} \geq \frac{1 - E_m}{\frac{1}{m} + \mathcal{C}(\Phi)} = \Lambda_m$$

and

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

$$\eta_m = \Omega(1)$$

Theorem 2. *For a binary RDP Φ, $\forall \Lambda > \frac{1}{\mathcal{C}(\Phi)}$, there exists a small error inference attack on Φ with $\eta_m = \Lambda$, $\forall m$.*

Proof. Follows from the channel coding theorem [14].

Theorem 1 indicates that $\eta_m = \Omega(1)$. Theorem 2, that $\eta_m = \Theta(1)$.

Attacks that correspond to codes are those where the queries $\mathbf{x}$ are deterministic functions of the desired bits $\mathbf{p}$. These are rare but not impossible. We provide an example of such an attack.

Example 2. The deterministically-related query attack. Consider a database of records of all residents of a county. From each record, consider the set of the following bits:

x_1. “location = North”;
x_2. “virus X test = positive”;
x_3. “gender = male” AND “condition Y = present”.

Suppose it is also known that, for this county,

$$(location\ =\ North) \oplus (virus\ X\ test\ =\ positive) \Leftrightarrow (gender\ =\ male) AND \\ (condition\ Y\ =\ present) \tag{5}$$

i.e,

$$x_1 \oplus x_2\ =\ x_3 \tag{6}$$

for all records, where $\oplus$ represents the XOR operation. This could be determined, for example, from county health statistics.

Suppose B chooses as desired bits $\mathbf{p} = (x_1, x_2)$ for all records, and designs an over-determined query sequence by also requesting x_3. Without randomization, B would not need to do so; with randomization, x_3 serves as a parity 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 over-determined sequence of m queries whose values are completely determined by $\mathbf{p}$ - through a set of m equations known to be satisfied by $\mathbf{p}$ and $\mathbf{x}$. Equation (6) is one such equation.

If the attack is recognized, A could:
(a) refuse to respond
(b) respond with $\phi(x_1) \oplus \phi(x_2)$ instead of $\phi(x_1 \oplus x_2)$.

Recognizing the attack is not trivial. If, instead of "male with condition Y", x_3 were, "$(location = North) \oplus (virusXtest = positive)$", it may be recognized by A, through extensive record keeping, as a logical combination of previously provided bits. But in the form of a request for a bit about gender and condition Y, and in the absence of knowledge of the specific relationship of equation (5), or a causal relationship - as opposed to a statistical one in a limited population - gender and condition Y are not readily seen to be revealing information regarding infection with virus X. Such an attack is fairly difficult to recognize, and hence to counter.

An approach like that of the source-channel coding theorem shows that B cannot do better using another procedure. This gives our final result, that the tight asymptotic lower bound on query complexity for zero asymptotic error is the ratio of queryable entropy to protocol channel capacity. As a corollary, the privacy of binary RDP is the inverse of its channel capacity.

The values of $\mathbf{p}_m$ are not necessarily uniformly distributed, and hence the entropy of $\mathbf{p}$, the queryable entropy, is not necessarily $log_2 M$. From the source coding theorem, if the entropy of $\mathbf{p}$ is $\mathcal{H}(\mathbf{p})$, then $\mathbf{p}$ is represented by $\mathcal{H}(\mathbf{p})$ bits on average (over many records). This observation can be combined with a reasoning similar to that in Theorem 1 to obtain a result similar to that of the source-channel coding theorem, except, as with Theorem 1, inference attacks are not of constant η_m, and do not consist of queries $\mathbf{x}$ that are deterministic combinations of the required bits $\mathbf{p}$. Again, we derive the asymptotic lower bound, and Shannon's results show it is tight.

Theorem 3. *The tight asymptotic lower bound on the query complexity, on average, per record, for a small error inference attack, is $\frac{\mathcal{H}(\mathbf{p})}{\mathcal{C}(\Phi)}$ if the record sequence is stationary, i.e. if the number of records is N_r, and γ_m the number of queries per record of sequence $\mathbf{x}$,*

$$\lim_{m\to\infty} \omega_m \to 0 \Rightarrow \exists\ \Gamma_m \text{ such that } \gamma_m \geq \Gamma_m\ \forall\ i \geq m \text{ and } \lim_{N_r\to\infty} \Gamma_m = \frac{\mathcal{H}(\mathbf{p})}{\mathcal{C}(\Phi)}$$

Proof. $\frac{\mathcal{H}(\mathbf{p})}{\mathcal{C}(\Phi)}$ *is an asymptotic lower bound*: Assume the existence of a small error attack with asymptotic query sequence length $K = \frac{\mathcal{H}(\mathbf{p})}{\mathcal{C}(\Phi)} - \Delta$ 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 1, 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 $\mathbf{p}$, $\frac{m}{\eta_m}$, at most $\frac{N_r(K+\epsilon)}{(\frac{1}{\mathcal{C}(\Phi)}-\nu)} = \frac{N_r(\mathcal{H}(\mathbf{p})-\mathcal{C}(\Phi)(\Delta-\epsilon))}{1-\nu\mathcal{C}}$, i.e. each record is represented, on average, by a number of bits strictly smaller than the record entropy for small enough ϵ, δ, ν. This violates Shannon's source coding theorem [5–pg. 89, Thm. 5.4.2] and [14].

$\frac{\mathcal{H}(\mathbf{p})}{\mathcal{C}(\Phi)}$ *can be achieved from above (i.e. tightness)*: straightforward from Shannon's source-channel coding theorem [5].

Thus Theorem 3 says that the query complexity per record, on average, for a zero error attack, is independent of the error.

Theorem 2 says that small error attacks in which η_m remains the same (but decrease in error is paid for by increase in total query complexity) exist if $\eta_m \geq \frac{1}{\mathcal{C}(\Phi)}$. It does not say anything about how the attacks will be constructed, and while the query complexity is tightly bounded below, the information-theoretic result does not indicate whether the processes of determining the values of $\mathbf{x}$ and $\mathbf{p}$ are computationally feasible. Recall that the value of $\mathbf{x}$ is computed by the database, A, and its complexity is measured by the number of logical operations performed to produce a response to a query from points in the database.

Some results since Shannon's work help address the issue of feasibility and construction. Forney's work, originally published in [7], shows that Shannon codes that are encodable and decodable in polynomial time exist. This implies that polynomial-time small-error attacks of constant finite η_m exist. More recent work, that of Spielman, [15] shows how to construct linear time encodable and decodable codes that approach the channel coding theorem's limits. Thus, linear time attacks with η_m approaching η_{min}, and arbitrarily low error, can be constructed. It is likely that attacks modeled on good, computationally feasible, error-correcting codes would consist of queries $\mathbf{x}$ that are rather contrived combinations of queryable bits from $\mathcal{Q}$. It is not clear how easy it would be to recognize such attacks. Recognizability constraints, ignored by us, could affect the existence result.

Corollary 1. *The tight asymptotic lower bound on the query complexity, per bit of queryable entropy, for a small error inference attack on $\Phi_{\mathcal{B}}(0.5 \pm \beta)$ is $\frac{ln2}{2\beta^2}$. Hence η_m is $\Theta(\frac{1}{\beta^2})$.*

Proof. The result follows from Theorem 1-3 and equation (1).

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

Proof. Follows from Theorems 1-3 and Definition 4.

Corollary 3. *The privacy of $\Phi_{\mathcal{B}}(0.5 \pm \beta)$ is $\Theta(\frac{1}{\beta^2})$.*

Proof. Follows from Corollary 1 and Definition 4.

In statistical databases, it is typically assumed that a larger number of queries (per attribute desired) is required for a lower error. Our proof of the existence of small error attacks for all asymptotic rates below channel capacity implies that a finite, fixed number of queries, per attribute desired, can ensure asymptotic error is zero; i.e. while total cost needs to increase to reduce error, the cost per bit of entropy need not.

Further, our work demonstrates that some inference attacks, which may not be as recognizable as repetition queries, are less expensive per bit. Last, at first glance it might appear that combinations of a greater number of bits for a query provides greater protection of the bits. But we have shown that combinations of a greater number of bits may also reduce error considerably through B's use of efficient error correcting codes.

Though our results follow very easily from classical results in information theory and coding, our view of the protocol 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 protocol is to decrease the information transfer over the protocol given certain constraints (such as the error in statistics that use these perturbed data points). Because of this, A would be interested in channels with small capacity, i.e. "good" privacy protocols. On the other hand, B is interested in the efficient transfer of bits over a particular protocol, typically a channel with small capacity, and a number of the constructive results from the theory of coding are of interest to him.

5 Conclusions

Our result on the correspondence between channel codes and certain types of inference attacks is an example of the study of attacks on non-perfect protocols using results from coding theory. Interesting further results could follow from viewing non-perfect anonymous delivery protocols - such as Crowds [13] and non-perfect combinations of mixes - as channels. Ramp secret sharing schemes [4] might also be amenable to this approach. An even more interesting direction of further work is to determine if our approach provides ingredients for a theory of statistical attacks on block and stream ciphers known to leak information.

References

1. Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, Vol. 21, No. 4, pp. 515-556, December 1989.
2. D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. *Proceedings of the Twenteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Santa Barbara, California, USA, May 21-23 2001.

3. R. Agrawal and R. Srikant. Privacy-Preserving Data Mining. *Proc. of the ACM SIGMOD Conference on Management of Data*, Dallas, May 2000.
4. G. R. Blakley and C. Meadows. Security of ramp schemes. *Proc. of Crypto'84*, Lecture Notes on Comput. Sci., 196, pp. 242–268, Springer Verlag, 1984.
5. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.
6. Csilla Farkas and Sushil Jajodia. The inference problem: a survey. *ACM SIGKDD Explorations Newsletter* Volume 4, Issue 2, pp. 6-11, December 2003.
7. David G. Forney. *Concatenated Codes*, MIT Press, Cambridge, Mass., 1966.
8. Diane Lambert. Measures of Disclosure Risk and Harm. *Journal of Official Statistics*, 9, pp. 313-331, 1993.
9. Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. *Journal of Cryptology*, 15 (3), 177-206, 2002.
10. Michael Luby. *Pseudorandomness and cryptographic applications.* Princeton Computer Science Notes, 1996.
11. Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*, pp. 67-73, Cambridge University Press, New York, NY, 1995.
12. Krishnamurty Muralidhar and Rathindra Sarathy. Security of random data perturbation methods. *ACM Transactions on Database Systems (TODS)* vol. 24, no. 4, Dec. 1999, pp. 487 - 493
13. Michael K. Reiter and Aviel Rubin. Crowds: Anonymity for Web Transactions. *ACM Transactions on Information and System Security*, Vol. 1, No. 1, pp. 66-92, November 1998.
14. Claude Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, vol. 27, pp. 379-423, July 1948.
15. Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, Vol 42, No 6, pp. 1723-1732, 1996.
16. Poorvi Vora. The channel coding theorem and the security of binary randomization. Proc., 2003 IEEE International Symposium of Information Theory, Yokohama, Japan, June 30 - July 4, pp. 306, 2003.

A Appendix: List of Symbols

A	Database
B	Data collector
$\mathcal{Q}$	set of queryable bits
q_i	a queryable bit
$\mathcal{S}$	set of sensitive bits
s_i	a sensitive bit
X, x_i, A_i	a single queried bit
$\mathcal{I}(\alpha;\beta)$	the mutual information between α and β
$Y = \phi(X)$	a single response to a query X
ρ	probability of truth
Σ	$\{0,1\}$
$P(Y\|X)$	posterior pdf, (or *a posteriori* pdf) of protocol/channel
m	number of queries or length of query sequence $\mathbf{x}$
ω_m	maximum probability of error for $\mathbf{x}$
η_m	ratio of queries to bits determined for $\mathbf{x}$
$\mathbf{x}$	a query sequence
M	number of values of $\mathbf{x}$
$\mathbf{p}$	sequence of queryable bits B wishes to determine
k	number of required bits or length of $\mathbf{p}$
Φ	protocol/channel
$\mathcal{C}(\Phi)$	capacity of Φ
$\Phi_{\mathcal{B}}$	binary protocol
β	small bias of a binary protocol
E_m	average probability of error of protocol using $\mathbf{x}$
$\mathcal{H}$	entropy of $\mathbf{q}$
N_r	number of records