The channel coding theorem and the security of binary randomization

Poorvi L. Vora
poorvi@ieee.org
Hewlett-Packard Co., USA.

Abstract

We propose that the randomization protocol for privacy protection be viewed as a communication channel, with lower channel capacity implying greater privacy. Focusing on binary symmetric randomization, and attacks consisting of the querying of deterministically related attributes, we show that a one-to-one correspondence exists between (i) the patterns of all attacks of "rate" r nd (ii) invertible channel codes of the same rate. The "rate" of an attack is defined as the number of independent target attributes determined per query, and is typically less than unity. We define reliable attacks as those whose probability of error can be made arbitrarily small by increasing the number of queries while maintaining rate. From the channel coding theorem, its converse and other more recent results on error-correcting codes, we conclude that (a) reliable attacks exist for all rates below protocol capacity (b) reliable polynomial-time attacks exist for all rates below protocol capacity (c) linear time attacks with arbitrarily small error and rates approaching protocol capacity can be constructed and (d) the rate of a reliable attack consisting of querying deterministically-related attributes to determine uniformly distributed target attributes cannot be higher than the protocol capacity.

1. Introduction

Randomization for the purposes of privacy protection refers to the probabilistic perturbation of data points to introduce protection through information-theoretic uncertainty, and to allow the possibility of the use of the perturbed data points for the estimation of the statistics of a population. It has been in use for about twenty years in public surveys and in statistical database security [1], and has recently been proposed as a means of personal privacy protection [2]. A randomization protocol consists of (i) a query of an attribute (a data point from an individual's personal profile), followed by (ii) a response, usually revealing a probabilistically perturbed value of the attribute. The protocol is defined by the parameters of the perturbation. It provides privacy protection of individual data points to individual users, and aggregate statistical information to data collectors. The amount of privacy and the accuracy of the aggregated information both depend on the extent of the perturbation. When used by individuals, randomization has the potential of enabling user-controlled, variable privacy. It also provides the opportunity of trading privacy for benefits, such as recommendations, that depend on the statistical or structural properties of the user's profile.

The secrecy of randomization is neither information-theoretically [3, pg. 44-48] nor computationally [4] perfect. This is intentional and often provides benefit to the user, e.g. when randomization leaks information for use with recommendation algorithms. A measure of the secrecy of randomization is required to calibrate it with respect to other privacy protocols, to compare various randomization protocols among themselves, and to measure the cost of a randomization benefit such as recommendations. In this paper, we mathematically define and study a class of common attacks on binary randomization, and derive results with respect to this class of attacks.

2. Literature Survey

Most existing measures of randomization are based on the error measure of the protocol. There are some recent measures of interest that are slightly different. In [5], the measure suggested is the maximum *a posteriori* probability of a particular record belonging to a particular individual. This measure depends on *a priori* probabilities and not only on the protocol's explicit effect on privacy/anonymity. In [6], the suggested measure is the size of the *a posteriori* range of possible values of a continuous-valued data point given a confidence level. It does not address the fact that some values may be more likely than others. In a different approach, [7] suggest the use of the differential mutual information between the original and perturbed continuous-valued data points. Larger values of mutual information imply larger privacy invasion. Unlike all other work we are aware of, this measure addresses the *change* in uncertainty due to a protocol instance. Mutual information does, however, depend on the original pdf, and not only on protocol parameters. None of the existing work on the security/privacy of randomization addresses specific attacks, such as those consisting of making related queries, in a satisfactory manner. We propose that one think of the protocol as a channel and show that this gives access to a number of major theorems in communication and coding theory that have implications for attacks and the security of randomization.

3. The protocol as a channel

Consider a simple classical example of binary randomization. In a telephonic public health survey, a data collector asks the respondent to roll a fair die, and, based on whether the rolled die shows a number divisible by 3 or not, to provide a false or true answer to the question, "Do you have HIV?". The data collector obtains individual answers which are true with a probability of truth $= \frac{2}{3}$. If the protocol were replaced by a memoryless binary symmetric communication channel with probability of error $= \frac{1}{3}$, the outputs of the channel and the protocol would be (information-theoretically, and hence, also computationally) indistinguishable.

A symmetric binary protocol consists of random variables $X \in \mathcal{X}$ representing personal data points; and $Y \in \mathcal{Y}$ representing the randomized responses to the query. It is characterized by the probabilistic relationship between X and Y, $P(Y|X)$. We denote the protocol by $\phi = (\mathcal{X}, \mathcal{P}(\mathcal{Y}|\mathcal{X}), \mathcal{Y})$. This is identical to the definition of a channel [8]. In fact, $P(Y|X) = 1-q$, $Y = X$; and $P(Y|X) = q$, $Y = \overline{X}$; where q is the probability of a lie. It is easy to see that the probability of a lie is the probability of error in the corresponding channel. We use $\phi(X)$ to denote the randomized value of X according to $P(Y|X)$.

A natural measure of a channel is its capacity, which measures the worst-case privacy *invasion* of the protocol. While the goal of communication theory is to increase information transfer over a channel given certain constraints, in this framework the goal of the theory of privacy is to decrease the information transfer over the protocol given certain constraints (such as the error in recommendations that use these perturbed data points). Because of this, privacy theory would be interested in channels with small capacity (i.e. "good" privacy protocols). As an example, the capacity of a binary symmetric protocol with small bias, i.e. one in which the probability of truth is $0.5 \pm \beta$ for small β, is determined by the second order term of the Taylor expansion (zeroth and first order terms are 0), which is easily seen to be $\frac{2\beta^2}{ln2}$. In the next section we demonstrate the relationship between channel capacity and protocol security with respect to a class of common attacks.

4. A class of attacks

Most attacks on randomization are in the form of related queries. Before we define attacks constructed from related queries, let us establish some notation. A binary attribute q is a binary function of parameter θ, and a member of $\mathcal{Q}$, the set of binary attributes under consideration. The value $q(\theta)$ is the binary value of the attribute for parameter value θ, $q : \Theta \rightarrow \Sigma = \{0, 1\}$. For an example, consider the binary attribute q, "capitalist", and parameter value θ, "Karl Marx". $q(\theta)$ is then 0, or false. As another example, consider the binary attribute "visited a news site", with parameter $\theta = \{JohnSmith, 11Nov.2001, 3pm\}$. The value of $q(\theta)$ is then the truth value of "John Smith visited a news site on 11 Nov. 2001 at 3 pm". We now define the class of attacks we address.

Definition 1: A $(k, n), k \leq n$ deterministically-related attribute sequence (DRAS), is a set of n "queried" binary attributes whose values are completely determined by corresponding values of a set of k binary "target" attributes, such that all k-tuples of target attribute values exist. Its rate is $\frac{k}{n}$.

More formally, a set of n binary attributes, $\mathbf{Q} = (Q_1, Q_2, ...Q_n) \in \mathcal{Q}^n$, is a (k, n) DRAS if there exists (a) a set of k binary attributes, $\mathbf{q} = (q_1, q_2, ...q_k) \in \mathcal{Q}^k$, $\mathbf{q}(\Theta) \;=\; \Sigma^k$, and (b) a function $\Lambda : \Sigma^k \rightarrow \Sigma^n$, mapping each k-tuple of bits to an n-tuple of bits, such that $\mathbf{Q}(\theta) = \Lambda(\mathbf{q}(\theta))\ \forall(\theta)$. Λ is the DRAS map.

Thus, Λ is the function taking the set of "target attributes", the one the data collector is really interested in, $\mathbf{q}$, to the set of queried attributes, $\mathbf{Q}$. Note that the image of $\mathbf{q}$ should be all of Σ^k. We require this so that none of the target attributes is redundant, and so that the rate of a DRAS is well-defined. If we did not have this requirement, any (k, n) sequence with a constant attribute added to the sequence would be a $(k + 1, n + 1)$ sequence, or a $(k, n + 1)$ sequence. With this requirement, such a sequence would be only a $(k, n+1)$ sequence. Whether the requirement is met or not can be checked by testing the range of $\mathbf{q}(\Theta)$. Changing Θ (the population of interest) may change whether the requirement is satisfied.

Example 1 Consider a database of attributes of all residents of a county. Consider the set of attributes: Q_1. "location = North"; Q_2. "virus X test = positive"; Q_3. "gender = male". Suppose it is also known that, for this county, all men with virus X are in the south, and all females with the virus are in the north. This could be determined, for example, from the number of each gender that tested positive for the virus at the county's medical centers, and implies that (location = North) XOR (virus X test = positive) = (gender = male). The attributes Q_1, Q_2, Q_3 form a DRAS of rate $\frac{2}{3}$, as the value of Q_3 is totally determined by the values of Q_1 and Q_2. The target attributes are "location = North" and "virus X test = positive". The queried attributes are "location = North", "virus X test = positive", and "gender = male". $\mathbf{Q} = \Lambda(q_1, q_2) = (q_1, q_2, q_1 XOR q_2)$. We have chosen a simple example where the target attributes are a subset of the queried attributes, but this need not always be the case.

$Q_3(\theta)$ serves as a parity check for the values of $Q_1(\theta)$ and $Q_2(\theta)$, or, in the communication channel framework, as an error-correcting symbol. The target attributes may be thought of as the "message" bits - the variables required to be transmitted by the protocol. The queried attribute values may be thought of as the code bits - the actual variables transmitted by the protocol. The rules of the protocol allow the data collector to design the queries, and hence the code and the function Λ, to best determine the target attribute values (the message bits) from randomized responses (the code bits after transmission). He does this without knowing the values of the code bits (i.e whether a particular response is 0 or 1). We wish to determine how well he can really do.

In this paper, we do not address the element of the attacker's choice of specific questions, simply the pattern among them, the DRAS map Λ. It is a key design element of an attack. All DRAS do not correspond to attacks. An attack requires: (a) that the map Λ is *one-to-one* and (b) a way of *estimating* the value of the target attributes from the randomized responses to queries. This leads us to the following definition of an attack:

Definition 2: A (k, n) DRAS attack for binary protocol ϕ is a one-to-one DRAS map Λ and an estimation map $\Psi : \Sigma^n \rightarrow \Sigma^k$ for estimating $\mathbf{q}(\theta)$ from $\phi(\mathbf{Q}(\theta))$. Its rate is $\frac{k}{n}$.

The rate of an attack is its efficiency, as it is the number of target data values obtained per query (this is typically less than one). It is clear that reducing the efficiency of an attack may help increase the estimation accuracy.

Recognizing a DRAS, attack or not, is not trivial. If, instead of "gender = male", Q_3 were, "(location = North) XOR (virus X test = positive)", it may be recognized, through extensive record keeping, as a logical combination of previously answered queries. But in the form of a query about gender, and in the absence of knowledge that (location = North) XOR (virus X test = positive) = (gender = male), gender is not readily seen to be revealing information regarding infection with virus X. Such a DRAS is fairly difficult to recognize, and hence to counter. If the attack is recognized, the user would provide not $\phi(Q_1(\theta) XOR Q_2(\theta))$, but $\phi(Q_1(\theta)) XOR \phi(Q_2(\theta))$.

We do not explicitly address recognizability in this paper. We are not aware of models of recognizable attacks, and it appears that if such models existed, they could not be static, and would not provide realistic limits to the ability of an attacker to disguise an attack. Not enforcing non-recognizability as a requirement on the definition of an attack ensures we obtain worst-case bounds (though perhaps not tight ones) - i.e. we assume an attacker powerful enough to always ensure that a DRAS attack is not detectable as an attack.

5. One-to-one channel codes correspond to DRAS attacks

We use the communication channel model of the protocol to study the rates of DRAS attacks. In this model, one-to-one channel codes, because they increase the efficiency of data transmission over a channel, are attacks. We draw a correspondence between DRAS attacks and channel codes, and then use the channel coding theorem to provide a bound on the rate of a DRAS attack.

Definition 3: A (k, n) binary channel code for a binary channel $(\Sigma, P(y|x), \Sigma)$ consists of (a) a domain of messages, Σ^k (b) an encoding function taking messages to codewords, $f : \Sigma^k \rightarrow \Sigma^n$ and (c) a decoding function acting on all possible channel output, $g : \Sigma^n \rightarrow \Sigma^k$. The rate of the code is $\frac{k}{n}$, the number of message bits per codeword bit.

Theorem 1: A one-to-one correspondence exists between the set of all one-to-one (k, n) binary channel codes and the set of all (k, n) DRAS attacks.

Sketch of proof: The hard work was in the definitions, the proof is almost trivial.

Binary DRAS attack ⇒ one-to-one binary channel code: Given the (k, n) DRAS attack of Definition 2. While the DRAS map Λ takes the target attributes to the queried ones, it induces a corresponding map on the attribute values for a parameter θ. The code map f is the map induced on the attribute values for parameter θ, $f(\mathbf{q}(\theta)) = [\Lambda(\mathbf{q})](\theta) = \Lambda(\mathbf{q}(\theta))$. The domain of f is Σ^k, as $\mathbf{q}$ is such that $q(\Theta)$ covers all of Σ^k. The estimation map g for channel ϕ is the estimation map ψ for the DRAS on protocol ϕ. The code is one-to-one because Λ is one-to-one from Definition 2.

One-to-one binary channel code ⇒ binary DRAS attack: Given the (k, n) binary code of Definition 3, and any set of target queries $\mathbf{q} \in \mathcal{Q}^k$, such that $\mathbf{q}(\Theta) = \Sigma^k$, let $\mathbf{Q} = f(\mathbf{q})$. $\mathbf{Q}$ is a (k, n) DRAS, and f is one-to-one, hence f, with estimating function g, is a (k, n) DRAS attack for protocol ϕ corresponding to the binary channel defined by $P(y|x)$. $\square$

There are as many DRAS attacks as there are error-correcting codes. Among these, is it possible for the attacker to obtain attacks that do not sacrifice rate for estimation accuracy?

6. Existence and bounds on attacks We *define a reliable attack* as one which can be made to decrease estimation error by increasing the total number of queries while maintaining rate - i.e one in which the price of increased accuracy is not a decrease in rate, but an increase in number of attributes targeted and number of total queries.

Definition 4: A reliable DRAS attack of rate r is said to exist on a binary protocol ϕ if $\exists$ a sequence of (rn, n) DRAS attacks in which the maximum probability of estimation error $\to 0$ as $n \to \infty$.

A reliable DRAS attack has quite a strong property: the probability of estimation error should decrease to 0 as more queries are performed, but the number of target attributes per query should remain the same and not decrease. Such an attack provides a significant benefit to the attacker: the probability of error can be decreased at will while keeping the cost per target attribute constant, and paying for error decrease through an increase in total number of queries. Statistical disclosure control, defined in [1], is the ability to require a large number of independent data points for small estimation error of a single attribute. It is considered a desired quality of statistical database security techniques. The existence of reliable attacks of rate r would imply that a fixed number of independent data points per independent target attribute $(\frac{1}{r})$ is sufficient to decrease error as much as desired. Statistical disclosure control would be possible only if r were small enough. Of course attacks that are not reliable would be of interest to the attacker, but they would have a limitation: the probability of error would have a non-zero lower bound.

Theorem 2: Given a protocol ϕ, reliable DRAS attacks of rate r exist $\forall r < \mathcal{C}(\phi)$.

Proof: Follows immediately from Theorem 1, the channel coding theorem [9] (a more modern version in [8, pg. 198]), and the observation that a good code is one-to-one.

Theorem 2 above says that DRAS attacks in which the rate remains the same (but decrease in error is paid for by increase in number of queries) exist if the rate is below the protocol capacity. It does not say anything about how the attacks will be constructed, and whether the encoding and decoding functions, Λ and ψ, are computationally feasible. Some results since Shannon's work help address the issue of feasibility and construction. Forney's work, originally published in [10] - a short summary of which is accessible in [11, pg. 129] - says that Shannon codes that are encodable and decodable in polynomial time ($O(n^4)$) exist. More recent work, that of Spielman, [12] shows how to construct linear time encodable and decodable codes that approach the channel coding theorem's limits. Thus, linear time constructible and decodable DRAS attacks with rates approaching protocol capacity can be constructed.

Theorem 3: Given a protocol ϕ, reliable DRAS attacks of rate r do not exist for uniformly distributed inputs and $\forall r \geq \mathcal{C}(\phi)$.

Proof: Follows immediately from Theorem 1 and the converse of the channel coding theorem [9] and [8, pg. 198].

Applying Theorem 3 to the binary symmetric protocol of probability of truth $0.5 \pm \beta$, with channel capacity $\frac{\beta^2}{ln2}$, (see section 3) the upper bound on the rate of a reliable DRAS attack on a symmetric binary protocol with uniformly distributed input and small bias β is $O(\beta^2)$.

Our results imply that channel capacity provides a measure of privacy invasion in this way: it is the supremum of all possible rates of a reliable DRAS attack with the uniformly distributed target attributes. Decreasing channel capacity decreases an attacker's efficiency in a specific manner - he needs to perform more queries per independent target attribute to decrease probability of error to zero with an increase in queries. Further, the upper bound on his rate goes as the square of the small perturbation of the probability of truth from 0.5. We do not say anything about other types of attacks.

7. Conclusions

We have proposed approaching randomization protocols as communication channels, and have shown a one-to-one correspondence between deterministically-related attribute sequence (DRAS) attacks and channel codes. We show, using the channel coding theorem, that this approach leads to the derivation of an upper bound on the the rate of all reliable DRAS attacks. We hope that this work will provide a framework for the study of attacks (through the correspondence with channel codes) and counter attacks.

Acknowledgements

We would like to thank Umesh Vazirani for discussions, constant encouragement, and an observation enabling a step in the proof of the correspondence between channel codes and DRAS attacks. We would also like to thank Gadiel Seroussi for encouraging the use of an information-theoretic approach, and for pointing the necessity of independence among target attributes.

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] Anne Eisenberg, "With False Numbers, Data Crunchers Try to Mine the Truth", *New York Times*, July 18, 2002.

[3] Douglas R. Stinson, *Cryptography Theory and Practice*, CRC Press, 1995.

[4] A. C. Yao, “Theory and Application of Trapdoor Functions”, 23^rd *IEEE Symposium on Foundations of Computer Science*, pp. 80-91, , Chicago, Illinois, 3-5 November 1982.

[5] Diane Lambert, “Measures of Disclosure Risk and Harm”, *Journal of Official Statistics*, 9, pp. 313-331, 1993.

[6] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining”, *Proc. of the ACM SIGMOD Conference on Management of Data*, Dallas, May 2000.

[7] 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.

[8] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[9] Claude Shannon, “A mathematical theory of communication”, *Bell Systems Technical Journal*, vol. 27, pp. 379-423, July 1948.

[10] David G. Forney, *Concatenated Codes*, MIT Press, Cambridge, Mass., 1966.

[11] Robert J. McEliece, *The Theory of Information and Coding*, Cambridge University Press, 2002.

[12] Daniel A. Spielman, “Linear-time encodable and decodable error-correcting codes”, *IEEE Transactions on Information Theory*, Vol 42, No 6, pp. 1723-1732, 1996.