

On-Line Data Security Implementation by Random Sequence

Xu Huang and Dharmendra Sharma

Abstract—Web services, with an emphasis on open standards and flexibility, may provide benefits over existing capital markets integration practices. However, web services must first meet certain technical requirements including performance, security and fault-tolerance. There has recently been a great focus from research projects towards providing unlimited, correct, numerical responses to ad-hoc queries to an on-line database, in particular multimedia data, while not compromising confidential numerical data. Unlike traditional random data perturbation (RDP), a new approach titled pseudo random binary sequences covered RDP is carefully first established in this paper. It is also demonstrated how numerical confidential data, such as multimedia data, can be protected against insider threat based on data. As described, it is easily extended this method to different applications, in particular multimedia online systems, such as telemedical systems, e-government, e-commerce, etc.

Index Terms—On-line multimedia data systems, online data protecting, random sequence, e-medical system, e-commerce.

I. INTRODUCTION

The Internet is critical to business. Companies have no choice but to connect their internal networks to the rest of the world to link with customers, suppliers, partners, and their own employees. But with that connection comes new threats: malicious hackers, criminals, industrial spies. These attackers regularly steal corporate assets and intellectual property, cause service breaks and system failures, mar corporate brands, and scare customers. Recently there has been great attention paid to approaches to disclosure limitation while answering queries to a database, protecting numerical confidential data against insider threat based on data or algorithms. Some papers such as [1-5, 15] have carefully investigated a few methods such as *confidentiality via camouflage* (CVC), where an economic model of the intermediation of an electronic market for private information and outlined the economic desiderate for a viable market.

In fact, markets transacting information, in particular via Internet, have grown rapidly and become a multi-billion dollar industry. In 2000 it was estimated that the size of the information markets was \$33.5 billion and a large slice of this market was serviced by database vendors [5]. As West said, the public good nature of information that is characterized by nonrivalrous consumption and nonexcludability, present

challenges in pricing of information products [5]. However, as [1] pointed out, this can cause significant unauthorized re-use of information and may result in market failure, where there is no economic incentive to provide such products despite significant consumer demand for them.

Government agencies such as the Census Bureau, which are responsible for gathering and disseminating information, adopt many techniques including the masking of microdata, to limit the disclosure of confidential information [7]. A snooper may be able to infer confidential information associated with a particular individual through a clever choice of queries, leading to disclosure [8]. It is well known that databases often use random data perturbation (RDP) methods to protect against disclosure of confidential numerical attributes. Muralidhar and Sarathy in their paper [6] showed how security for the three random data perturbation methods described by Traub et al. [10], Kim [12] and Tendick [11], and Tendick and Mastloff [9].

The conclusion are, under the attacks from “professional” snoopers, the later two methods are the same level of security in both univariate and multivariate case the first method can vary depending on the characteristic of the specific database being perturbed. If the condition is under the attacks from “casual” snoopers, the third of the method will offer lower level security [6]. This is because the simple linear relationships were employed.

However, Garfinkel et al. presented a more complex linear function in their paper [1], but there is vulnerable to insider data information. Since it would be not too hard to get the “real” exact information by solving multi-variables equations obtained from the appearing values that formed from a linear or semi-linear relations as shown by [1], such as \mathbf{P}^3 vector, where two less than unit factors, λ_1 and λ_2 , are helpful against vulnerable to insider data information. But they are limited except for protecting private information.

We shall illustrate “a case study” of an online medical surveys as an application of protecting confidential data, where the whole system consists of four different security levels, namely *top protection* that protected by PRBSCI, in the case study this is for only particular members can use it, *protection level* that presented by linear hidden function such as *confidentiality via camouflage* (CVC) [1], in the case study, this is for member’s uses, *classified level*, in our case study, this level is defined as for individual patent for his/her won data checking, and *staff level* that can be observed by multi-level information that controlled by the manager.

As an example, the information of the average life-time for a patent in a particular cancer can be varied: for a doctor (“staff level”) the data may show 10 to 15 months, denoted as [10, 15] months, (defined as 1st first range) but for patents

Xu Huang and Dharmendra Sharma are with Faculty of Information Sciences and Engineering, University of Canberra, ACT Canberra, Australia e-mail: {Xu.Huang, Dharmendra.Sharma}@canberra.edu.au.

may show [2, 30] months (called 2nd range), which indicates different accuracy and can be denoted as range 1 \subset range 2.

II. RDP AND LINEAR FUNCTIONAL OPERATIONS FOR DATA HIDDEN

In this section we briefly introduce the RDP and linear functional operations for data hidden. Following the setting used in Tendick and Matoff [9], let the variables $\mathbf{A} = \{A_1, \dots, A_p\}$ represent the set of p attributes of the database. Some of these attributes are confidential and will be perturbed using RDP. It is also assumed that \mathbf{A} is a realization from a multivariate normal distribution with mean vector μ and covariance matrix Σ . For a simplest form, a single confidential attribute A (with mean μ and variance σ^2) RDP involves the addition of random noise (ε) to result in the perturbed attribute $A^\#$ as follows:

$$A^\# = A + \varepsilon \quad (1)$$

where ε has a mean of zero, or called no base, and variance of $d\sigma^2$. Here, d is defined as perturbation level, representing the extent of protection against partial disclosure that the database administrator intends to provide. If both A and ε have a normal distribution, then the perturbed attribute $A^\#$ also has a normal distribution with mean μ and variance $(1+d)\sigma^2$. Muralidhar and Sarathy discussed three methods of random data perturbation based on equation (1) with different approaches, such as “independent noise”, “correlated noise” and “no based” RDP, “based” RDP [6].

Garfinkel and Rice raised a more complex algorithm but it is still linear functional operation [1]. Since they discussed a given query in functional form f and a subset $T \subset N$, the *confidentiality via camouflage-star* (CVC-STAR) query answer is given by an interval $[f^-, f^+]$ containing $f(a, T)$, which we are interested but with totally different algorithm. We need to have a look about their method. Following their notations, the answer has a range:

$$R(f) = f^+ - f^- \quad (2)$$

The interval is constructed by minimizing and maximizing the query function over a compact set S defined as

$$S := \bigcup_{i \in N} L_i \quad (3)$$

$$\text{where } L_i := \mathbf{a} + \{\alpha u_i + (1-\alpha)l_i - a_i\} \mathbf{e}_i : \alpha \in [0,1] \quad (4)$$

for all $i \in N$. The set S is a concentric union of n line segments in n -space, and can be thought of as resembling a star with center at \mathbf{a} . Hence, provided that f is defined and bounded on S for a fixed T , the answer interval is found by setting

$$f^- := \inf \{f(x, T) : x \in S\} \quad (5)$$

$$f^+ := \sup \{f(x, T) : x \in S\} \quad (6)$$

In particular, for a continuous or a function f such as MEAN (SUM) defined on S , the corresponding answer interval can be found by a one-dimensional minimization and maximization of f over each line segment L_i , $i \in T$, and then concatenating the resulting optima.

Therefore, we have

$$f^- := \min \{f(x, T) : x \in L_i\} \quad (7)$$

$$f^+ := \max \{f(x, T) : x \in L_i\} \quad (8)$$

for all $i \in T$, and then we have

$$f^- := \min \{f(x, T) : x \in T\} \quad (9)$$

$$f^+ := \max \{f(x, T) : x \in T\} \quad (10)$$

It is shown in [1] that, as an example, we have

$$\Delta_i^- := l_i - a_i$$

$$\Delta_i^+ := u_i - a_i$$

for all i . Then it has the following equations

$$f^- = \sum_{i \in T} a_i + \min \{\Delta_i^- : i \in T\}, \quad (11)$$

$$f^+ = \sum_{i \in T} a_i + \max \{\Delta_i^+ : i \in T\}, \quad (12)$$

Let us use the same example, refereeing Table 1, if we have the query $f(x, \{168\}) := \text{SUM}(1,6,8)$ the CVC-STAR would answer it with $[f^-, f^+] = [148, 202]$ with $R(f) = 54$.

In order to control the answer quality, [1] also introduces a factor, titled *over quality of the answer* to query f after the reduction of the subjects' protection intervals is given as

$$q(\Theta) := 1 - \frac{f^+(\Theta) - f^-(\Theta)}{R(f)} \quad (13)$$

where $\Theta := (\theta_1, \dots, \theta_n)$ and

$$f^-(\Theta) := \inf \{f(x, T) : x \in S(\Theta)\} \quad (14)$$

$$f^+(\Theta) := \sup \{f(x, T) : x \in S(\Theta)\} \quad (15)$$

and $S(\Theta)$ is the star protection set obtained using the subject intervals $[l_i(\theta_i), u_i(\theta_i)]$, $i \in N$.

Table 1: Example Database Table

Rec	Name	Age	ST	Job	Sal	l	U
1	Kelley	32	CT	Eng	65	50	90
2	Burrell	56	OH	Mgr	25	6	46
3	Allen	45	NY	Mgr	98	80	120
4	Marsh	27	NJ	Eng	87	71	111
5	George	45	CT	Mgr	54	34	74

6	Ollie	54	OH	Eng	27	5	45
7	Corley	37	NY	Jou	45	10	50
8	Dropo	34	CT	Eng	78	70	110
9	Biake	28	CT	Mgr	56	36	76
10	Yoka	47	OH	Mgr	30	11	51

As we can see that CVC is vulnerable to insider data information, even though [1] claimed that it is only “marginally vulnerable” to insider algorithm information due to the rage may not be found by those people who have insider data information. But if we are talking about “malicious attacker”, which is always the case for any web set protections, we do need to think about the “security” problem more seriously. For example, since the relationship between \mathbf{L} and \mathbf{a} and \mathbf{U} and \mathbf{a} are “linear”, even the whole “linear” relationship seems to be more complex than that discussed in Muralidhar and Sarathy [6], it would not be so hard to find out this relation by a programmed calculations. For example, attackers may insert a few variables for the list values obtained from a few queries and form a set of equations in various ways that approaches to the targeted parameters (such as “predictive deconvolution” method to obtain the so called “prediction distance” as small as possible) and ultimately approaches the true values they wanted. In fact for any database the larger numbers of data, which make malicious attackers easily to break the linear “cover” to “exact” data.

III. PSEUDO RANDOM BINARY SEQUENCES COVERED RDP

As we have seen that the major vulnerable for functional operations (or covers) is the nature of the linear functions we have used, such as equation (1) or (3).

It is obvious that if we use a non-linear functions to operate (or to cover) the true vales, it would be much harder for a attacker to break the protected data. It is obvious that the more complex of the function we used, the harder to break the protected data.

Therefore, we are going to establish a semi-random operation system, by which we can almost fully protect our stored database (true values) at the same time it is easy to obtain the true values for those people whom are servicing for, we called this system as *pseudo random binary sequences covered RDP*. Assume we have a vector as below

$\mathbf{a} := (a_1, a_2, \dots, a_n)$, the real (true) values
 $\mathbf{s}(t) := (1, 1, \dots, -1, 1, -1, \dots)$, pseudo random binary sequence consists of ± 1 randomly for example it can be produced by random electronic signal generator.

Here, $\mathbf{n}(t) :=$ sampled random noise, for example it may be a Poisson distribution that we are focusing on due to its discrete nature or Gaussian distribution, etc.

In our pseudo random binary sequences covered RDP system as shown in Figure 1, we can first let the “true value” \mathbf{a} (line 1) be modulated by the pseudo random binary sequence $\mathbf{s}(t)$ (line 2) by a multiplier. The symbols “ \times ” and “ Σ ” in Figure 1 are “multiplier” and “adder” respectively. Then the modulated values (line 6) will be added a vector $\mathbf{n}(t)$ (line 4) that is sampled from a real noise (line 5; either Gaussian or Poisson noise). The output of this adder (line 7), denoted as port 1 in Figure 1, is the output designed for the

non-member of the system (or it is a protected output information) as the information from port 1 is very noisy, or the obtained information is meaningless at all. In our case study this level is for those people who are login without membership. For those people who are allowed to obtain the true values, or accurate information, they will allow to take the output in port 2 (output 2) in Figure 1 (line 11), where there are the two lines added (line 9 and line 10) to the adder. The “-1” in Figure 1 is an inverter, which makes the output value of the inverter equal to the “negative” input.

In Figure 1, the line 1 is \mathbf{a} vector (true values), line 2 is the pseudo random binary sequence. Both tow vectors put into “multiplier” the output of the multiplier (line 6) is $\mathbf{a} \cdot \mathbf{s}(t)$ that is added to the sampled noise (line 4) by the adder as shown in Figure 1. The line 7 is the output 1, which is “fully protected data”, which can be expressed as follows:

$$\text{data of output 1} = \text{data (line 7)} = \mathbf{a} \cdot \mathbf{s}(t) + \mathbf{n}(t) \quad (16)$$

It is noted that the data in line 8 (equal to line 2) is multiplied to the data in line 7 by a multiplier and the data of this multiplier is in line 9, which can be expressed as

$$\text{data in line 9} = [\mathbf{a} \cdot \mathbf{s}(t) + \mathbf{n}(t)] \cdot \mathbf{s}(t) \quad (17)$$

We also note that the data in line 3, sampled noise, will become negative sampled values by the “inverter”, the output of the inverter is the data in line 10, which is

$$\text{data in line 10} = -\mathbf{n}(t) \Rightarrow -\mathbf{n}'(t) \quad (18)$$

where it is important that we defined the delayed data in line 10 as that different from original sampled values, denoted as $-\mathbf{n}'(t)$.

We can easily replace the pure “inverter” as “amplifier” or equal to $(-c)$, where $c =$ a designed constant, rather than negative unit, -1 . This “ c ” will be used for designing multi-level quantity of protections as we described in the first section. Hence, the output 2 in Figure 1, sitting on line 11, can be expressed as

$$\text{data in line 11} = \mathbf{a} + [\mathbf{n}(t) - \mathbf{n}'(t)]\mathbf{s}(t) = \mathbf{a} + \mathbf{k} \cdot \mathbf{s}(t) \quad (19)$$

It is worth noting that the relation $[\mathbf{s}(t)]^2 = 1$ is used and the \mathbf{k} is controllable semi-constant vector. Because the difference between $\mathbf{n}(t)$ and $\mathbf{n}'(t)$ is very small or we can use a so-called “threshold” amplifier to control, we can control the exact value vector as the true vector, in which we may let $\mathbf{k} = \mathbf{0}$.

It is important to note that the output 2 can be divided into several levels depending on the parameters in equation (17), which will be discussed later (section 4). In order to illustrate our pseudo random binary sequences covered RDP system, now we pick the noise from the Poisson and Gaussian described in [13-14]. To make the “covering” nicely, we pick the “reasonable” mean values responding to the Table 1 presented in previous section.

It is important to note that it is to compare the case described in table 1, which implicates we have to establish a

small sample space in responding to the size of table 1.

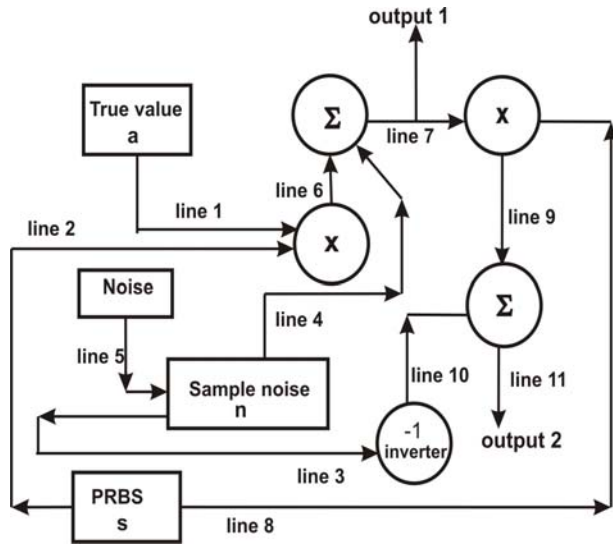


Figure 1: The block diagram for the pseudo random binary sequences covered RDP system

In table 2, an example for explain how our pseudo random binary sequences covered RDP system works responding to the previous table 1.

The mean of the true values is 56.5, for the pseudo random binary sequences covered RDP system, the Gaussian mean was used 56.1 and the Poisson mean was 55.5 and the output 1 is either “Poi” covers by Passion noise or “Gau” covers by Gaussian.

As we can see that we can easily to just the output by the constant vector **k**, if we would like to build them for different answer quality. Now if the names replace by the multimedia materials, for example, patient’s X-ray images that can be protected by different levels, for example, full information (accurate image e.g. for a doctor), rough information (some related information e.g. for reasonable working relationship), noisy information (the right information but noisy e.g. for some relatives), no information (cannot obtaining information due the required privacy).

The mean of the true values is 56.5, for the pseudo random binary sequences covered RDP system, the Gaussian mean was used 56.1 and the Poisson mean was 55.5 and the output 1 is either “Poi” covers by Passion noise or “Gau” covers by Gaussian.

Table 2: Data for the pseudo random binary sequences covered RDP system

Rec	Name	Age	ST	Job	Sal	Poi	Gau	O/P2
1	Kelley	32	CT	Eng	65	43	78	65.01
2	Burrell	56	OH	Mgr	25	78	-35	24.9
3	Allen	45	NY	Mgr	98	54	30	98
4	Marsh	27	NJ	Eng	87	63	200	87.03
5	George	45	CT	Mgr	54	45	57	54.01
6	Ollie	54	OH	Eng	27	81	-84	26.99

7	Corley	37	NY	Jou	45	35	149	45
8	Dropo	34	CT	Eng	78	66	-29	78.01
9	Biake	28	CT	Mgr	56	53	106	56
10	Yoka	47	OH	Mgr	30	37	89	30

IV. CONCLUSION

We have first designed a pseudo random binary sequences covered RDP system with multi-levels to disclosure limitation while answering queries to a database.

In this system, with the descriptions in section 3, we have demonstrated the fact that this system can be almost fully protected to the stored confidential information, even to those people who have insider data information that is very hard for the linear operation functions.

As the electronic multimedia online systems have sharply increasing, in particular for e-medical, e-governor, e-commerce, etc., the discussed system will play an important role in the future.

REFERENCES

- [1] R. Garfinkel, R. Gopal, and D. Rice, “New approaches to disclosure limitation while answering queries to a database: protecting numerical confidential data against insider threat based on data or algorithms,” Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.
- [2] R. Sarathy and K. Muralidhar, “The Security of Confidential Numerical Data in Databases,” Information Systems Research, **13** 389-403.
- [3] R. Gopal, P. Goes, and R. Garfinkel, “Interval protection of confidential information in a database.” Informs Journal on Computing, **10**, 309-322, 1998.
- [4] R. Gopal, P. Goe, and R. Garfinkel, “Confidentiality via camouflage: the CVC approach to databss security.” Operation Research, **50**, 3 (2002).
- [5] L. West, “Private markets for public goods: pricing strategies of online database vendors.” Journal of Management Information Systems, **17**, 1 59-86, 2000.
- [6] Krishnamurty Muralidhar and Rathindra Sarathy, “Security of Random Data Perturbation Methods”, ACM Transactions on Database Systems, Vol. 24, No.24, December 1999, pp487-493.
- [7] W. A. Fuller, “Masking procedures for microdata disclosure limitation” J. Official Stat. 1993. 9, 2, pp.383-406.
- [8] N. R. Adam and J. C. Wortmann, “Security-control methods for statistical databases: A comparative study” ACM Comput. Surv. **21**, 4 1989. pp 515-556.
- [9] P. Tendick and N. Matloff, “A modified random perturbation method for database security” ACM Trans. Database Syst. **19**, 1 1994, pp47-63.
- [10] J. Traub, Y. Yemini and H. Wozniakowski, “Statistical security of a staitical database” ACM Trans. Database Syst. **9**, 4, 1984, pp672-679.
- [11] P. Tendick, “Optimal noise addition for preserving confidentiality in multivariate data”, J. Stat. Plan. Inference **27**, 2, 1991 pp.341-353.
- [12] J. Kim, “A method for limiting disclosure in microdata based on random noise and transformation” In Proc. Of the American Statistical Association on Survey Research Methods, American Statistical Association, Washington, DC, 1986. pp370- 374.
- [13] Xu Huang, Dharmendra Sharma and Allan C. Madoc, “Numerical Information in Databases Online Protecting via Random Multi-Level Perturbations,” Journal of WSEAS Transactions on Information Science and Applications, ISSUE 5, Volume 4, May 2007. ISSN 1790-0832, pp1105-1110.

- [14] Xu Huang, Allan C. Madoc, Dharmendra Sharma, and Nazni Farooq, "Pseudo Random Binary Protecting On-line Data Communications against Insider Threat" IEEE, the 9th International Conference on Advanced Communication Technology (ICACT 2007), February 12-14, 2007, Phoenix Park, Korea. IEEE Catalog Number 07EX1671, Proceedings Volume II, pp.1347-1352.
- [15] Chales P. Pfleeger, and Shari L. Pfleeger., "Security in computing", Pearson Education, Inc. Pbulishing as Prentice Hall professional Technical Reference. 2003.