On-line Data Protecting via Pseudo Random Binary Sequences

Xu Huang, Allan C. Madoc, and Dharmendra Sharma
School of Information Sciences and Engineering, University of Canberra, ACT 2617
Canberra, Australia

Abstract—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, while not compromising confidential numerical data. Unlike traditional random data perturbation (RDP), a new approach titled *pseudo random binary sequences* covered RDP is carefully established. It will be also demonstrated how numerical confidential data can be protected against insider threat based on data in this paper. As an application of our method, an online medical survey database is, as a case study, described, which can be easily extended to other application fields, such as e-government, e-commerce, etc.

Keywords- database management, against insider threat, data protecting, data security, online survey.

1 Introduction

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] 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. The main determinants of information pricing include the utilization of resources required to provide information products, value obtained by customers from the provided information, and competition between information vendors.

Transacting on private information requires careful consideration of privacy which is the right of an individual group or institution to determine for what purpose information is to be collected, when the collected information should be used, etc. A significant demand for private information exists since it is a valuable asset to private and governmental institutions and departments.

A market for private information will require the involvement of a trusted third party information intermediary. If the intermediary fails to create an environment of trust, the market will become unviable due to deficient volumes of supply and demand. Therefore, it is up to the trusted third party information intermediary to provide security, quality, and to add value sufficient to create a sustainable trustworthy marketplace.

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 know that databases often use random data perturbation (RDP) methods to protect against disclosure of confidential numerical attributes. Muralidhar and Sarathy in their paper [6]

shoed 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 very 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 sown 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.

In our current paper we are focusing on the case that the information is compiled from private (or confidential) resources and needs to be protected in multilevel depends on applications, such as high secure, secure, classified, etc.. In order to almost fully protect the stored information, we, in contrast normal linear RDP, first present pseudo random binary sequences covering information (PRBSCI), by which the protected information almost cannot be vulnerable to a attacker even to those people who have insider data information. Here we used a word "almost" is because (1) there is no perfect security system that never be broken (2) in terns of confidentiality our method is very hard to be broken due to a set of "random" numbers will be used to cover the "real" data, but this "random" is "pseudo random". We shall illustrate "a case study" of a 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 define 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.

In fact those multi-levels, which will be discussed in the following sections, at some stage we can take them as "multi-accuracy" information. 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 may show [2, 30] months (called 2nd range), which indicates different accuracy and can be denoted as range 1 $\subset$ range 2.

In the next section, it will briefly introduce random perturbation (RDP) and the linear functional operations such as CVC described in [1], and then the vulnerabilities of the linear functional operations will be discussed. After that the pseudo random binary sequences covering information (PRBSCI) is carefully designed. It can be shown that the multilevel accuracy (or security) can be controlled by some parameters for an online database. In the section four a case study of an online medical survey database with protecting numerical confidential data against insider threat is presented.

2 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 $\mathbf{A} = \{A_1, \ldots A_p\}$ 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 **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].

Since the equation is simplest form of RDP, and the "simplest" linear relation indeed kills the security level of all those three methods. 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

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

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 **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}):= SUM(1,6,8) the CVC-STAR would answer it with $[f^-, f^+]$ = [148, 202] with $R(f)$ = 54.

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

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 $\Theta := (\theta_1, \ldots, \theta_n)$ 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 $[l_i(\theta_i), u_i(\theta_i)]$, $i \in N$.

As we can see that CVC is vulnerable to insider data information, even though [1] clamed 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 **L** and **a** and **U** and **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.

3 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*.

For our pseudo random binary sequences covered RDP system we define the following notations:

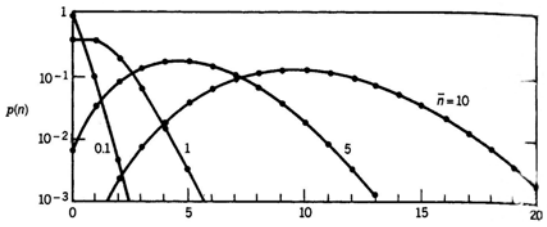

Figure 1: Various probability density functions (PDF)s of Poisson distribution with different means.

a := (a_1, a_2, … a_n), the real (true) values
s(t) := (1,1,…-1, 1, -1,…), pseudo random binary sequence consists of ±1 randomly for example it can be produced by random electronic signal generator.
n(t) := sampled random noise, for example it may be a Poisson distribution (Figure 1) 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 2, we can first let the "true value" **a** (line 1) be modulated by the pseudo random binary sequence **s**(t) (line 2) by a multiplier. The symbols "×" and "Σ" in Figure 2 are "multiplier" and "adder" respectively. Then the modulated values (line 6) will be added a vector **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 2, 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 2 (line 11), where there are the two lines added (line 9 and line10) to the adder. The "-1" in Figure 2 is an inverter, which makes the output value of the inverter equal to the "negative" input.

In Figure 2, the line 1 is **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 **a**· **s**(t) that is added to the sampled noise (line 4) by the adder as shown in Figure 2. The line 7 is the output 1, which is "fully protected data", which can be expressed as follows:

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

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

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

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

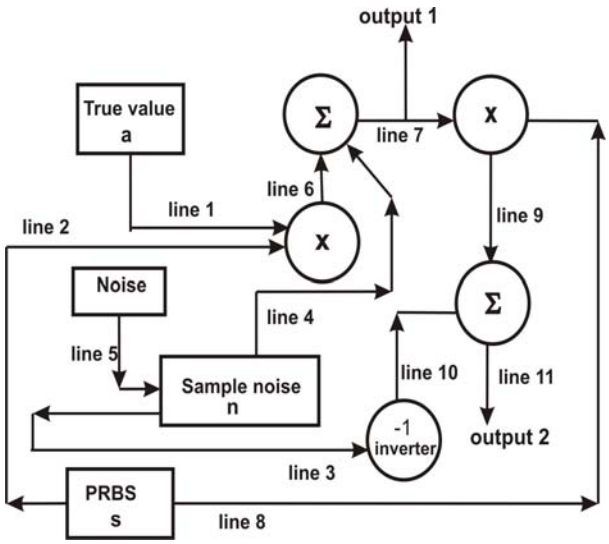

Figure 2: The block diagram for the pseudo random binary sequences covered RDP system

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 2, 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 $[s(t)]^2 = 1$ is used and the **k** is controllable semi-constant vector. Because the difference between **n**(t) and **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 **k** = **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 [13-17]. In order 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. The table is shown below and the sampled frequency is very low (because the values are very small and the statistical nature was far from the real cases due the sample space are too small). 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. Both are meaningless as we expected. The output 2 is very close to the true values (with its mean 56.495).

As we cane 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.

It is important to realize that this example is only to show how our pseudo random binary sequences covered RDP system works in terms of the processing. In real databases that have very large number data and the large sample space will show strongly the advantages of our pseudo random binary sequences covered RDP system in comparison with any linear functional presented in previous section.

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

In following section we are going to briefly introduce our pseudo random binary sequences covered RDP system in a case study of "online medical surveys".

4 Case study of a Pseudo random binary sequences covered RDP system: Online medical surveys

As a case study of the pseudo random binary sequences covered RDP system we shall briefly introduce online medical surveys, the focus will be the parts of private information protection by an established multi-level answer system for clients.

Medical research depends largely on surveys and questionnaires to trace patient progress and response to medication. The information so obtained is stored and analyzed for patterns, trends, possible cures and preventative strategies.

There are strict guidelines and legal requirements as to the ethical principles, confidentiality and anonymity of patient records. However, patient records need to be linked over time, to trace their progress in terms of medication, environmental changes, etc., so that patterns may be established, decisions taken on future treatment and even for life-saving measures in an emergency. It is therefore necessary to have some method of medically identifying and tracing patients over repeated surveys, while at the same time ensuring that their personal details are ever compromised.

Another aspect for surveys is the need for efficient presentation of Information to the various categories of users of the surveys. Information retrieval needs to be user-friendly and meaningful, so as to quickly and easily help patients, medical practitioners and researchers. In medical surveys, the information assists can be listed as

- Patients, for self-assessment, and for self-help
- Doctors, to assess patient progress and response to drugs, treatment
- Researchers, which also includes doctors, to find patterns, trends, leading to possible cures, prevention or arrest of disease
- Statisticians interested in information gathering and statistical analyses
- Potential patients, i.e., people with some susceptibility to the condition being surveyed
- Non-medical audience with an interest and philanthropists willing to help
- Government or public organisations with a view to regulating diseases, legislation, security issues

This paper is focusing on the need for private information, such as patient's lifetime for individual disease, medical drug's usages, job positions, salary, etc., which need to be protected or provided by different security levels for those people who have different login status. The hierarchy in our system followed [18] is shown in Figure 3.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

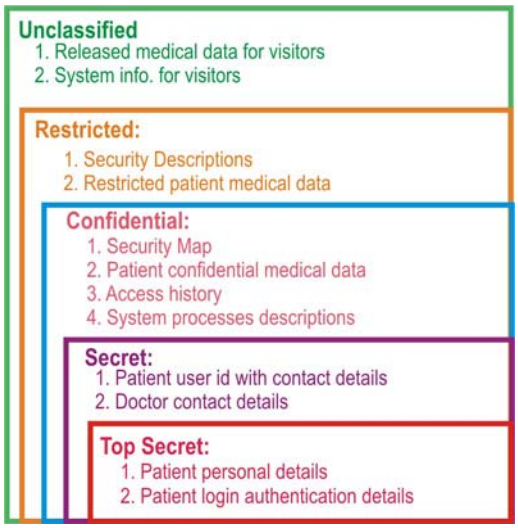

Figure 3: Security hierarchy in online medical surveys protected by pseudo random binary sequences covered RDP system.

Here we used the plan as the following

- The "top Secret" will be corresponding to the "output 1" in Figure 2 to all users but the people listed in Figure 3.
- The "output 2" we have marginal control to the different levels shown in Figure 3 as below:
  - "secret" $\in$ {-10, 10} $\cup$ {medical information}
  - "Confidential" $\in$ {-25, 25} $\cup$ {medical information}
  - "Restricted" $\in$ {output 1 in Figure 2} $\cup$ "Unclassified"
  - "Unclassified" $\in$ {general health information, similar to the general search engine}

5 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 case study of

"online medical surveys", 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 media online systems have sharply increasing, in particular for e-governor, e-commerce, e-hospital, 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 Pertur bation Methods", ACM Transactions on Databse 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 statiical 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] X. Huang and A. C. Madoc, "Image and its noise removal in Nakagami fading channels," IEEE 8th International Conference on Advanced Communication Technology, Phoenix Park, Republic of Korea, Feb, 20-22, 2006, Proceeding, Part I, pp.570.

[14] X. Huang and Dharmendra Sharma, "Analysis of the error probability of a binary receiver in wireless fading channels," IEEE 8th International Conference on Advanced Communication Technology, Phoenix Park, Republic of Korea, Feb, 20-22, 2006, Proceeding, Part II, pp.1087.

[15] X. Huang, "Multi-noise removal for images in wireless networks," International Journal of Computer Science and Network Security, Vol.6 No.1A pp.181-189. January 2006.

[16] X. Huang, A.C. Madoc, and Dharmendra Sharma, "Image Noise Removal in Nakagami Fading Channels via Bayesian Estimator," The Third International Workshop on Electronic Design, Test & Applications (DELTA 2006), Kuala Lumpur, Malaysia, Proceedings pp.31-34, 2006.

[17] X. Huang, "Investigations of The Error Probability of Digital in Nakagami Fading Channels," Computer and Its Application, GESTS International Transactions on Computer Science and Engineering, Vol.21 and No.1, pp677 -683. Nov. 30, 2005.

[18] Chales P. Pfleeger, and Shari L. Pfleeger., "Security in computing", Pearson Education, Inc. Pbulishing as Prentice Hall professional Technical Reference. 2003.