IJCCN International Journal of Computer Communications and Networks,
Volume 4, Issue 1, February 2014, ISSN 2289-3369
WWW.IARTC.NET

Privacy-preserving Data Mining for Personalized Marketing

Yong Jick Lee

Abstract—In electronic commerce markets, firms can easily achieve customers' personal information such as identity, demographic information, and shopping behavior. Researches in securing statistical database had introduced several tools and methods to secure such statistical database with sensitive personal information. Although the data perturbation methods secure the database very effectively, it is not applicable to the application beyond the simple statistical analysis on means, variances and covariance. In this paper, I suggest Multiple Staged Slice Perturbation Methods in order to apply them to RFM analysis. My study shows the possibility of applying a simple modification to perturbation methods in order to be able to perform the RFM analysis. My method of slicing the database into decile and perturbing each decile separately would maintain the mean and the standard deviation of each decile. I showed that current data security methods may not be applicable to some business analysis that deals with more than the mean, standard deviation and covariance between variables. Since perturbation method guarantees protection against exact disclosure, there is no threat of exact disclosure even if data is partitioned into small pieces and perturbed individually. However, because partitioning limits range of shuffling effect, partial disclosure is possible. Therefore, for achieving the maximum utility while preserving maximum security level, the number of partition should be minimized.

Index Terms— Data Mining, Data Perturbation, Privacy-preserving database, Statistical Database, Database Security.

1 Introduction

In electronic commerce markets, firms can easily achieve customers' personal information such as identity, demographic information, and shopping behavior. Such customer's private information is critical for direct marketing purposes [3, 12]. Firms use such information to profile the customer to infer the most profitable groups of customers. Personal information of customers is stored in a statistical database that has an ability of providing statistical information to users [4].

Although such profiling helps increasing the precision of marketing ability of a firm, a firm handling such statistical database also faces the risk of misuse of such information by unauthorized personnel. Unauthorized personnel not only include hackers or intruders who try to gain access to the data, but also include insiders who do not have enough privilege to access the data. Security issues arise when an unauthorized user obtains at least partial information that is sufficient to infer confidential information about any individual in the database with significant precision. Lately, securing such statistical database with customer information has become a critical issue for electronic commerce as privacy issues gain more attention [5].

Researches in securing statistical database had introduced several tools and methods to secure such statistical database with sensitive personal information. Such techniques suggested for securing statistical databases are implemented by either limiting the use of database or altering precision of database [1, 4, 9, 10]. For application purposes, trade-offs between accuracy and rate of disclosure of these solutions have to be considered. One of widely studied approach is data perturbation methods [8, 13]. It gains more attention as it is more effective in both ease of implementation and level of security that could be provided, alters data values to make the database less precise. This method adds random noise to the database entries so that exact confidential information can be protected while the original statistical information such as means, standard deviation, and correlation with other variables can be preserved [8].

Although the data perturbation methods secure the database very effectively, it is not applicable to the application beyond the simple statistical analysis on means, variances and covariance. In reality, most firms demand more than such simple analyses. Among the many techniques that have been used in the marketing practices, Recency, Frequency and Monetary (RFM) analysis, a well-known direct marketing analysis application, has been a popular tool for direct marketing campaign [1, 6, 7, 11].

Though it is not as sophisticated as many recent data mining techniques but only requires basic statistical tools for analysis, current data perturbation methods do not support RFM analysis as the analysis requires more than means, variances and covariance. It uses historic purchase data to study customer behaviors and segment customers in order to find the most re-

• *Y. J. Lee is with Department of Business Administration, Jungwon University, Goesan, Rep. of Korea. E-mail: yonglee@jwu.ac.kr.*

sponsive customers through analysis of three main variables of recency, frequency, and monetary value of each customer. Recency is defined as the time when the most recent purchase was made, frequency is defined as the number of purchases made by each customer, and monetary value is defined as the total dollar amount a customer has spent life-to-date. Using the RFM analysis, customers are ranked from highest to lowest profitable based on their RFM scores. For each variable, customers are classified into segments [7]. RFM analysis requires quintile or decile clustering information.

In this paper, I suggest *Multiple Staged Slice Perturbation Methods* in order to apply them to RFM analysis. Since most techniques available to secure statistical database limit the use of database at its full potential, existing methods restrict performing such RFM analysis. By slicing the database into pieces and applying the existing method multiple times, I try to preserve clustering information within the database while exploiting full aspects of the existing data perturbation method so that the method can be applied to secure the database while performing the RFM analysis.

The rest of this paper is organized as follows. Section 2 summarizes existing methods of securing such statistical database. Section 3 describes aspects of RFM analysis and privacy issues associated with the analysis. In section 4, I introduce the modified model of existing perturbation method. Analysis and the main results of my experiments are presented in Section 5, and the last section offers some concluding remarks.

2 THEORY

2.1 Data Perturbation Methods

Data perturbation methods preserve security while maintaining some statistical aspect of the original data set. The fundamental idea is to add a random "noise" to the confidential data to alter the database so that it can protect the original values from unnecessary disclosure [13]. For example, *General Additive Data Perturbation* (GADP) method by Muralidhar et al. [8] creates perturbed data using covariance between original confidential dataset and perturbed dataset. Therefore, GADP method could maintain security level while preserving mean, standard deviation of original dataset as well as covariance with other datasets.

Perturbation methods allow the analysis of the database at full potential without any limitation on access to the perturbed confidential or original non-confidential attributes. Moreover, perturbation method makes the statistical database portable. That is, after perturbation, statistical database can be exported for use outside of the firm, so that marketing analysis can be outsourced without risking the disclosure of confidential data. However, data perturbation sacrifices precision of query results. Although it tries to maintain some statistical information such as mean, variance and covariance, not all statistical information could be obtained due to the limitation of perturbation methods. For example, due to the "shuffling" effect of perturbation method, it does not maintain the rank of records and therefore a regression analysis also would be different after the perturbation. Table 1 summarizes various privacy-preserving methods.

2.2 The RFM Analysis

TABLE 1
SUMMARY OF PRIVACY PRESERVING METHODS

Method		Security		Accuracy			
		Exact Disclosure	Partial Disclosure	Use of Non-conf. Information	Statistical Information Available	Bias	Maintain "order"?
Query Restriction	Query Set Size Control	Yes	Yes	Partially	Yes	No	Yes
	Query Set Overlap Control	Yes	Yes	Partially	Yes	No	Yes
	Auditing	No	Yes	Partially	Yes	No	Yes
	Partitioning	No	Yes	Partially	Yes	Yes	Yes
	Cell Suppression	No	Yes	Partially	Yes	No	Yes
Data Perturbation	GADP	No	No	Fully	Yes	Yes	No
	Wavelet Transform	No	No	Fully	Yes	Yes	No

The RFM analysis is a widely used tool in marketing since the data on RFM variables are easy to collect and the analysis does not require complex statistical background or modeling techniques [6]. Several different segmentation methods for the RFM analysis have been introduced [9]. The most widely used segmentation method is to divide customers equally into five bins or cells for each variable. The RFM analysis maintains the information about the most recent time of purchase (recency), the number of times the customers made purchases (frequency), and the average money she or he spent (monetary). It does not have longitudinal information to track patterns or stream of customer behaviors. Rather, the RFM analysis only captures the "snapshot" of the customer behaviors. However, it can be used along with various other analyses to generate more insightful information.

In catalog sales industry, for example, the RFM analysis can be used to sort out existing customers for future marketing campaign [2]. The historical sales data can be sorted out based on the RFM ranks of each customer, and each segment is compared with the average response rate. For a certain segment, if the average response rate is greater than the threshold, the segment is retained for the next campaign. By eliminating less responsive customers and focusing on the most responsive customers, it enables marketers to reduce unnecessary costs and to increase the profitability.

Hughes [7] suggested using quintile for each variable. Customers are segmented into five groups based

on recency value, and then each segment is further divided into another five groups based on frequency value, and then divided even further into another five groups based on monetary value. Using this approach, customers are segmented into 125 (= 5 x 5 x 5) segments. If the dataset is large, decile could be used instead of quintile. In this case, the total number of cells would be 1,000 segments (= 10 × 10 × 10). Each segment can be evaluated using response rate, profitability, or other measures.

For illustration (and subsequent analysis) purposes, let us consider the example of an electronic commerce company. Assume that the company keeps sales history of customers. Customer identification number (custid), days since last purchase (recency), total number of purchase (frequency) and total dollars spent (monetary) values are collected for the RFM analysis. In addition, binary variable for whether the customer responded to the last email advertisement by making purchase or not (buyer) is included to predict response rate of each cell. The company maintains such data for 10,000 customers. For experimental analysis, database can be generated using random number generator reflecting typical aspects of RFM database as described in [7].

Summary of database is described in Table 2. Each percentile group of recency, frequency and monetary are computed from "recency", "frequency" and "monetary" variables and stored into "R", "F" and "M" variables, respectively. Each percentile group is divided into five subgroups (quintiles) and coded from "1" to "5." After classifying each RFM variable into quintile, three codes are aggregated into three digit codes. For example, if a record scored "1" in recency, "4" in frequency and "5" in monetary, three digit RFM code would be "145." Similarly, all records are classified into 125 distinct segments based on recency, frequency and monetary activity. Mean of the "buyer" variable of each RFM segment is the response rate of each segment.

Assume that the company is intended to send email only to the customers who are most likely to respond. If the company sets a goal of 2.5% response rate for upcoming email advertisement, the customers on the

TABLE 2
DATABASE FOR RFM ANALYSIS

CustID	Recency	Frequency	Monetary	Buyer	R	F	M	RFM
1	627	3	$81.54	0	2	4	5	245
2	364	2	$63.30	0	3	3	1	331
3	98	7	$75.97	0	5	5	3	553
...	...	...	...	...	...	...	...	...
9998	130	1	$61.03	0	5	1	1	511
9999	53	56	$71.64	0	5	5	2	551
10000	781	1	$69.01	0	1	1	2	112

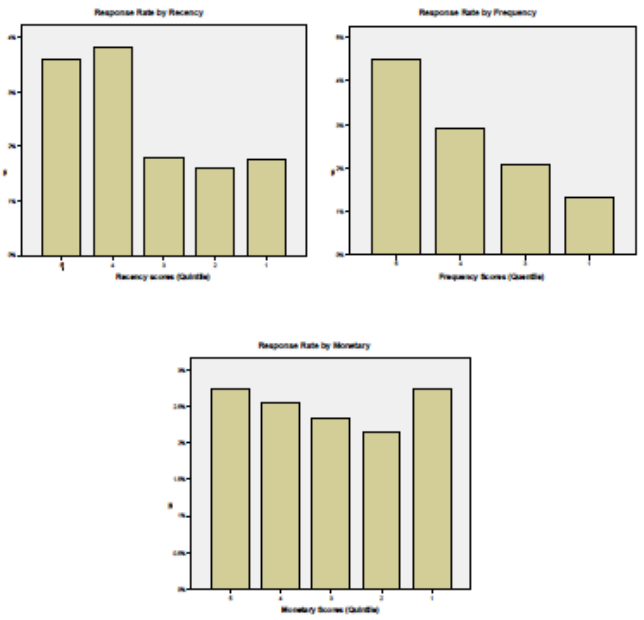

Fig. 1. Response Rate by each RFM Variables. (Set 1)

segments with a response rate higher than 2.5% would be selected for campaign and others would be excluded. In this example, 35 segments are selected and 90 segments are excluded. Top ten cells are listed on Table 3. Figure 1 depicts the relationship between RFM scores and response rate.

Clearly, some segments outperform others in terms of response rate. Hughes [6] argued that recency is the most influential, frequency is marginal, and monetary is almost flat (see Figure 1). However, by incorporating all three variables, the results accurately segment customers (see Figure 2).

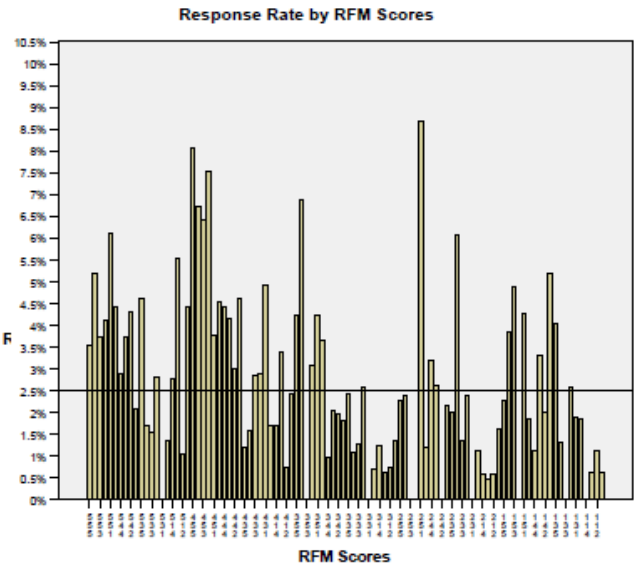

Fig. 2. Bar-chart showing response rate for each RFM segment (set 1)

3 Background and Assumptions

3.1 Privacy Issues in RFM Analysis

Is RFM data confidential? The answer is yes in many cases. Basic required variables are recency, frequency and monetary values that constitute a customer's shopping behaviors. Since the aim of RFM analysis is to develop a new direct marketing campaign, personal identifiable information such as name and address is included or at least linked to the RFM data. Even if some of RFM variables (such as recency and frequency) could be coded into their decile or quintile rank and may not require exact value, exact monetary value is necessary for the analysis since profitability is computed based on the monetary information. Moreover, it is common practice to outsource the marketing campaign. When outsourced, data that is considered non-confidential when it is used in-house might become confidential. For example, historical sales data of customers may be non-confidential for an internal analyst, but can become confidential for an outside analyst.

TABLE 3
TOP 10 RFM CELLS BY HIGHEST RESPONSE RATE

RFM	Response Rate	# of Customers
251	0.09	46
455	0.08	99
452	0.08	93
354	0.07	58
454	0.07	104
453	0.06	109
551	0.06	131
234	0.06	82
513	0.06	90
141	0.05	96

Consider a customer who recently made purchase from the electronic commerce company. If the customer had no purchase history with this merchant before, RFM data would show the exact time of the purchase on recency variable and the exact amount she spent on monetary variable. Frequency variable would confirm that she made the purchase with this store only once. All these values are considered to be private information of the customer since the data represent at least some part, if not all, of her shopping behaviors.

Due to the privacy issues described above, it is crucial to have proper protection on such confidential information. Strictly speaking, if the collection, analysis and storage of such customer's information are all carried out legitimately, there should be no legal issues. However, maintaining strict legitimacy is not always possible. Storing such sensitive information in non-secure storage space (such as personal computer) may also result in improper transfer – lost or stolen by snoopers. Transferring such information to a third party analyst may also cause concerns for information leaking. In many cases, mainly due to the need for accuracy or the inability to understand the importance of protecting such information, the database that contains private information has been used with no or little protection. As a result, preserving the privacy of customers while still being able to perform customer RFM analysis is important.

3.2 Applying Data Perturbation method in RFM analysis

In section 3.1, I argue that RFM data is confidential and therefore should be protected from unnecessary disclosure. Easiest way to secure the RFM data is to split the database into two segments, confidential and non-confidential attributes. Unauthorized personnel should not access to the confidential information, but is allowed to use non-confidential segments for her analysis.

However, splitting the database into two segments is not guaranteed to prevent security breaches. There is always a possibility of disclosure either by an accident or by an intended intrusion. Therefore, to be confident, some methods to protect RFM database should be implemented. Query restriction, however, is not an option as it not only limits comprehensive analysis of the data, but also does not provide full security as mentioned in Section 2.

I consider a data perturbation method as an alternative to secure RFM database. Since the RFM analysis is to study the relationship between RFM variables and target outcome, perturbation methods seem applicable to RFM analysis since it maintains covariance between recency, frequency and monetary values as well as covariance between these RFM variables and target outcome variable such as profit, response rate or etc. Moreover, portability of data perturbation methods facilitates the RFM analysis, even though they still cannot be used directly due to the possible misclassification errors described earlier.

However, since data perturbation methods only provide mean, standard deviation and covariance information, it m`ay also not be *directly* applicable for RFM analysis. For the RFM analysis, database should be classified based on the score of each attribute on the date of the most recent purchase, the number of purchases, and the monetary value that the customer spent on average. In all of the three cases of recency, frequency and monetary variables, data is analyzed based on the ranking of each category and classified into quintile or decile. Since the clustering information of data is destroyed after data perturbation, perturbed database would not be directly applicable in the RFM analysis.

Although perturbation method preserves the mean and standard deviation as well as the distribution of

the database, it destroys the rank order of each variable in the original database that is required to classify customers into decile cell membership for RFM analysis. However, data perturbation methods provide accurate mean, standard deviation and correlation information that is useful for the RFM analysis. In the RFM analysis, accuracy need not be precise to single records; only the proper placement of records into each "segment" plus the preservation of mean and standard deviation of each "cell" would be required for the RFM analysis.

Ideally, it would be best to maintain the mean, standard deviation and covariance as well as the order of the data while also achieving the maximum security level. However, security only can be achieved by sacrificing the accuracy of data, thus imposing the security-accuracy tradeoff. Therefore, it is important to find a way to modify data perturbation methods so that it can still perform the RFM analysis. Since the RFM analysis focuses on clustering, the key is to find ways to perturb data that preserve clustering.

To protect confidentiality while utilizing full aspects of RFM analysis, a technique that could (1) protect confidentiality and (2) preserve rank orderings is thus required. Issues with partial disclosure and unnecessary limitation on non-confidential information access prevent the use of the database at full potential. Although RFM analysis does not require disclosing individual entries in the customer database for analysis, full access to the database for segmenting customers is required for such analysis.

Data perturbation methods prevent the direct use of data for RFM since they "shuffle" data and can destroy the clustering of the original dataset. The order of the data is not preserved due to the random noise effect of perturbation methods. In this paper, I "slice" the whole dataset into decile and apply perturbation technique to each decile. It would shuffle data only within the decile so that it would minimize distortion of clustering. To verify feasibility of the proposed solution, I tested the methodology against the original GADP method proposed by Muralidhar et al. [8] using the above example of randomly generated 10,000 records.

4 MODEL

4.1. Multiple Staged Slice Perturbation Model

I explore the possibility of preserving security and characteristics of data perturbation methods while also preserving clustering of confidential information that is required for the RFM analysis. Obviously, using existing perturbation methods directly is not feasible as it does not provide accurate clustering due to the problems identified earlier. So, I first consider partitioning the dataset for the RFM analysis before applying the perturbation. That is, I suggest creating decile cells *ex-ante*. Since perturbation methods preserve the original mean and standard deviation, creating decile cells first and then perturbing each individual cell later would provide more accurate data set that can be used for the subsequent RFM analysis.

However, slicing the whole dataset into 125 clusters (5 × 5 × 5) or 1000 clusters (10 × 10 × 10) creates another problem. Since each individual cluster is too narrow compare to the whole, perturbing the data within the cluster may not be as secure as perturbing the whole. Especially if the whole dataset is small, partial disclosure maybe possible as one can infer the original value from a very narrow cluster. Therefore, applying RFM slicing first then perturbing may not be feasible.

The strength of data perturbation methods, especially General Additive Data Perturbation method suggested by Muralidhar et al. [8], is that they maintain covariance not only among perturbed data attributes but also between perturbed data attributes and original attributes. It leads to an idea of multiple-staged data perturbation. Since the covariance is always preserved, it is possible to perturb one attribute first, then after the first perturbation, perturb another attribute while treating the previously perturbed attribute as the original attribute. It provides the possibility of slicing the whole dataset into five (quintile) or ten (decile) clusters instead of 125 or 1,000 clusters.

Data perturbation "shuffles" the data with a small random noise added to data to maintain mean and standard deviation. Since the data should maintain the mean and the standard deviation, the value added to one record should be subtracted from another record(s), and the distance between the mean should be the same to preserve same mean and standard deviation. I expect that by applying GADP method to the dataset directly, the rank order of each confidential field would be destroyed, and therefore decile membership by each confidential value would not be preserved. However, by the nature of the GADP method, I expect that mean, standard deviation and correlation would be maintained.

The GADP method of Muralidhar et al. [8] is applied to the electronic commerce merchant example. The dataset consists of randomly generated 10,000 records. Descriptive statistics of original dataset is reported in Appendix A. The original data is partitioned into 125 groups by its percentile score of RFM variables. Each group is numbered from "1" (lowest) to "5" (highest). As discussed earlier, all of three RFM variables ("recency", "frequency" and "monetary values") can be considered to be confidential. When GADP method is applied to the whole electronic commerce merchant dataset, it shows that about 86% of records are misclassified, 22% of customers are misclassified into next adjacent cell while 63.8% of customers is misclassified into even further cells. Summary of misclassification is shown on Table 4. According to the experiment, applying GADP method to the whole dataset destroys decile membership, and therefore GADP

TABLE 4
TABLE OF MISCLASSIFICATION OF GADP DATASET VS. ORIGINAL DATASET

Distance of Misclassification	Recency (%)	Frequency (%)	Monetary (%)	Average (%)
0	14.55%	13.18%	14.60%	**14.11%**
1	23.08%	21.18%	22.09%	**22.12%**
2	17.84%	14.98%	17.64%	**16.82%**
3	14.38%	13.13%	14.43%	**13.98%**
4	11.22%	12.07%	11.02%	**11.44%**
5	8.24%	8.98%	8.42%	**8.55%**
6	5.42%	7.26%	5.76%	**6.15%**
7	3.02%	5.53%	3.62%	**4.06%**
8	1.67%	3.38%	1.92%	**2.32%**
9	0.58%	0.31%	0.50%	**0.46%**

method is not directly applicable for RFM analysis.

I argue that partitioning dataset into decile or quintile first and then applying perturbation method individually to each decile or quintile would maintain the original decile or quintile membership of confidential fields. In my proposed approach, GADP method is applied three times. Each recency, frequency and monetary variables is partitioned into five (quintile, set 1) or ten (decile, set 2) groups and perturbed one at a time.

After each perturbation, original attribute is replaced with perturbed attribute and previously perturbed attributes are considered as non-confidential attributes. For instance, at the first stage, recency variable is partitioned into five groups and for each cluster, GADP method is applied. After the first perturbation, original recency variable is replaced with the perturbed recency variable. At the next stage, frequency variable is partitioned into five groups. The GADP method is applied to each group individually only to the frequency attribute, then merged to create one large set, and replaced with the original frequency attribute and so on. Since GADP method maintains covariance of perturbed and original attributes, it is acceptable to replace the original attributes with the perturbed attributes for the next perturbation. Descriptive statistics of each partitioned perturbed sets are also depicted in Appendix A.

After applying the suggested method, the percentile score is generated again according to perturbed RFM scores and the results are compared with original percentile scores to decide whether any changes have been made. Since data is pre-segmented before perturbation, I expect that the percentile rank by perturbed RFM values would be similar to the percentile rank by original RFM values. However, I also expect that the classification might not be perfectly the same, and there might be some misclassification due to the random effect of the GADP method.

Lastly, the RFM analysis is performed using three data sets (original, GADP and partitioned GADP). The RFM analysis results from GADP method (Set 2) and suggested method (Sets 3 and 4) are compared with that of original dataset (Set 1). Response rate, selected decile, and selected customers are compared. First, quintile group of "recency", "frequency" and "monetary value" variables are recorded into "R", "F" and "M" variable respectively. "RFM" variable is recorded according to the "R", "F" and "M" scores (RFM = R × 100 + F × 10 + M). Response rate for each RFM cells are then computed. By applying threshold rate of 2.5% to the dataset, the most responsible group is identified, and a list of most responsible customers is generated. All sets are compared against the original set to decide whether there is any difference between them.

5 RESULTS

Two different partitioning, quintile and decile are applied and compared against original dataset. Result of quintile partitioning (Set 3) before applying GADP method shows a little improvement in terms of fewer misclassifications; however, only 39% of records are correctly classified while 61% of records are misclassified. About 37% of records are off by one adjacent cell. This is due to random noise effect of perturbation method. Records at the border of each cell may exchange their position due to addition of noise. About 24% of records are misclassified into even further cells. This is significant improvement from conventional GADP method.

Decile classification result (Set 4) produced by performing RFM after using suggested method is compared to that performed using the original data. Results show that only 13% of records are misclassified, 10.8% of customers are misclassified off of one cell, and 1.3% of customers are misclassified into further cells. Summary of misclassification is shown in Table 5. Although some misclassifications still exist, customers who are misclassified are only off by a maximum of 3 cells. The result shows that by partitioning cells before applying GADP method considerably increases feasi-

TABLE 5
TABLE OF MISCLASSIFICATION (SET 3 VS. SET 1)

Distance of Misclassification	Recency (%)	Frequency (%)	Monetary (%)	Average (%)
0	55.78%	22.82%	54.72%	**38.77%**
1	42.03%	32.10%	42.24%	**37.17%**
2	2.08%	19.73%	2.60%	**11.17%**
3	0.11%	9.86%	0.28%	**5.07%**
4		4.97%	0.10%	**2.54%**
5		2.12%	0.03%	**1.08%**
6		2.76%	0.01%	**1.39%**
7		2.71%	0.01%	**1.36%**
8		2.01%	0.01%	**1.01%**
9		0.92%		**0.46%**

bility of the RFM analysis while benefiting from the security provided by the GADP method. Furthermore, statistics such as mean, standard deviation and covariance is retained. Based on the simulation, I can conclude that suggested method would better maintain the original decile membership of confidential field.

Although partitioning the dataset significantly reduced rate of misclassification, some of the customers are still misclassified due to the random effect of GADP method. Moreover, the ranks of some RFM cells by response rate have also changed as shown in Figure 3. However, this few misclassifications may be acceptable since RFM analysis is a "prediction" of future behavior of customer, and there is always a statistical error between the "true" behavior of customers and statistically predicted forecast. Moreover, variance of misclassification is considerably smaller than that applying only GADP method to the whole dataset. There is significantly less chance to classify "best" customers into "worst" customers and vice versa, but it is still likely to classify "best" customers into second or third "best" customers and vice versa.

Since perturbation method does not raise issues of exact disclosure, level of security in perturbation

TABLE 6
TABLE OF MISCLASSIFICATION (SET 3 VS. SET 1)

Distance of Misclassification	Recency (%)	Frequency (%)	Monetary (%)	Average (%)
0	91.12%	85.23%	90.56%	**87.90%**
1	8.81%	12.25%	9.26%	**10.76%**
2	0.04%	1.57%	0.16%	**0.87%**
3	0.03%	0.95%	0.02%	**0.49%**

method depends on possibility of partial disclosure. To measure confidentiality of the methods, Muralidhar et al. [8] used two measures of (1) variance between original attribute and the perturbed attribute and (2) maximum proportion of variance that can be explained by any linear combination of confidential attributes using a linear combination of non-confidential attributes. Variance between original attribute and the perturbed attribute can be measured by

$$S_1 = Var(X - Y)/Var(X) \qquad (1)$$

where X represents a single original attribute and Y represents a single perturbed attribute. Therefore threshold for S1 is 1 meaning that, if S1 is greater than 1, no additional information is supplied due to perturbation. For example, if S1 is 1.8, variance between original attribute and the perturbed attribute exceeds variance of original attribute alone, therefore the perturbation does not reveal any additional information. Note that for S1, 1 is a hurdle rate, and anything above 1 can be considered as secured.

Maximum proportion of variance among linear

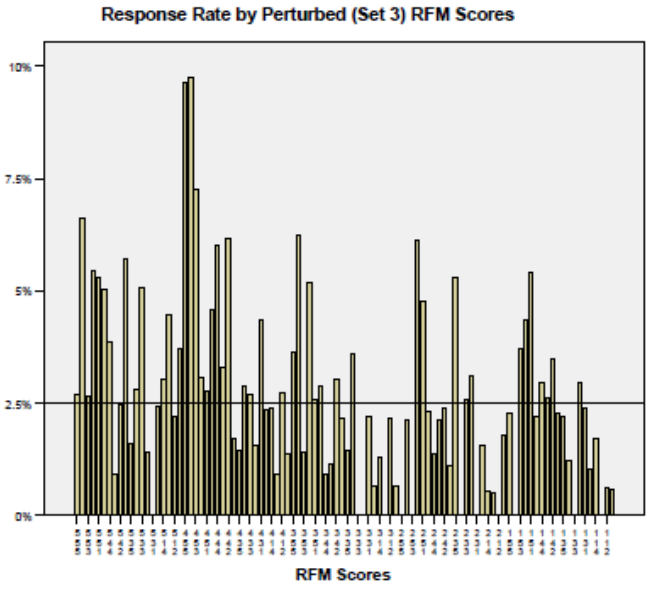

Fig. 3. Bar-chart showing response rate for each RFM segment (set 3)

combination of confidential and non-confidential attributes can be measured by

$$S_2 = 1 - \lambda \qquad (2)$$

where λ represents the largest eigenvalue resulting from the following matrix:

$$\sum_{XX}^{-1} \sum_{XV} \sum_{VV}^{-1} \sum_{VX}, \text{ where } V = \{S, Y\},$$

and

$$\sum_{VV} = \begin{bmatrix} \sum_{SS} & \sum_{SY} \\ \sum_{YS} & \sum_{YY} \end{bmatrix}.$$

Muralidhar et al. [8] suggests that linear combination (S_2) should be higher than .5. However,

GADP method predefines the linear combination S_2 and produces the perturbed dataset according to the predefined linear combination S_2. In my marketing example, S_2 is set to .65. However, it is worthwhile to report that for S_2 = .5 and S_2 = .8, results in terms of S_1 and classification result of RFM analysis were not significantly different from initial analysis S_2 = .65. I suspect that it is because each recency, frequency and monetary values by nature is not linear dependant of one another.

Security level measured by S_1 and S_2 for GADP method in my marketing example is S_1 = 1.30 and S_2 = .65. Security level measures of my suggested models are S_1 = 1.13 and S_2 = .65 (for Set 1, quintile partitioning) and S_1 = 1.06 and S_2 = .65 (for Set 2 decile partitioning). The result shows that variance between original attributes and perturbed attributes are slightly lower due to partitioning. When partitioning occurred first, perturbation was accomplished within the segment. Variance between original attributes and perturbed

attributes should be limited to the range of segment. To preserve clustering, some restriction should be relaxed. Security measures imply that more partitioning yield better clustering results with less possibility of partial disclosure. However, due to multiple staged slicing of original data, I could maintain the security level of S_1 above the threshold of 1 that Muralidhar et al suggests.

The result suggests that the proposed method in this paper actually maintains the desired security level ($S_1 > 1$ and $S_2 > .5$**)** while provide ability of performing RFM analysis. Both quintile and decile partitioning provides secure dataset after perturbation. However, decile partitioning provides much accurate classification. This is because the number of partition for perturbation was twice as many as actual RFM analysis therefore it offsets the misclassification rates. Yet, security level of S_1 is only 1.06. Although the hurdle rate for S_1 is 1 and the result is greater than 1 so that it is secure in my example, it is too close to 1. Depending on dataset size or variance of attribute, especially small number of data or small variance in one or more of RFM attributes, security level S_1 could go below 1.

DBAs should consider this possibility and must check S_1 after perturbation to ensure security level. Note that S_1 does not mean that the perturbed dataset is completely useless, but it has some risk of partial disclosure. To increase S_1, one can consider quintile partitioning or partitioning into somewhere between quintile and decile. The result shows that the security level of S_1 increases as the number of partition decreases by sacrificing accuracy. Note that the rate increase of misclassification for quintile from decile is much greater than the security gain from decile to quintile. Depending on precision of marketing campaign and desired security level, one should decide level of accuracy at the cost of security.

6. CONCLUSION

The RFM analysis is widely used for direct marketing. It is a simple and powerful tool to analyze customer's behaviors. As the Internet commerce expands, it is easy to collect customer's purchase history and track their shopping behavior. As the competition grows intense, understanding customer's behavior is a key to success in the Information Society. However, privacy issues are also becoming grave concerns for consumers. Customers are annoyed by spam mails and solicitation telephone calls, and claim that the firms should protect their privacy and analyze their private information legitimately.

Many techniques have been introduced to overcome privacy issues and they include query restriction methods and data perturbation methods. Query restriction methods are not applicable to the RFM analysis due to their limitation on accessing data and threat of exact or partial disclosure. Although data perturbation methods protect against such partial disclosure, they are of limited value for RFM applications that require preserving the clustering of data into ordered deciles. Thus, although data perturbation provides full access to data and better security and maintaining some aspects of statistical information, it is not directly applicable to the RFM analysis since it destroys the rank order of the data and therefore the segmentation would be inaccurate.

My study shows the possibility of applying a simple modification to perturbation methods in order to be able to perform the RFM analysis. My method of slicing the database into decile and perturbing each decile separately would maintain the mean and the standard deviation of each decile. Moreover, as in the General Additive Perturbation method that is intended to maintain covariance between confidential and non-confidential fields, the covariance between RFM variables and other variables is preserved, which adds accuracy of the analysis to the suggested method.

In this paper, I showed that current data security methods may not be applicable to some business analysis that deals with more than the mean, standard deviation and covariance between variables. My study suggested simple variation of existing methods that can solve the issues with the RFM analysis. However, the proposed method is only applicable to specific RFM analysis that maintains covariance between confidential and non-confidential fields and may not be applicable to other possible analyses. The proposed method is only applicable when the database needs be segmented into quintile or decile.

Since perturbation method guarantees protection against exact disclosure, there is no threat of exact disclosure even if data is partitioned into small pieces and perturbed individually. However, because partitioning limits range of shuffling effect, partial disclosure is possible. Therefore, for achieving the maximum utility while preserving maximum security level, the number of partition should be minimized.

REFERENCES

[1] Adam, N. R. and Wortmann, J. C. "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys*, vol. 21, no. 4, pp. 515-556, 1989.

[2] Bitran, G. R. and S. V. Mondschein. "Mailing Decisions in the Catalog Sales Industry," *Management Science*, vol. 42, no. 9, pp. 1364-1381, Sept. 1996.

[3] Chevalier, J. and Goolsbee, A. "Price competition online: Amazon versus Barnes and Noble," *Quantitative Marketing and Economics*, vol. 1, no. 2, pp. 203-222, 2003.

[4] Chin, R. Y. and G. Ozsoyoglu. "Auditing and Inference Control in Statistical Databases," *IEEE Transactions on Software Engineering*, vol. 8, no. 6, pp. 574-582, Nov. 1982.

[5] Desai, M. S. and Richards, T.C. and Desai, K. J. "E-commerce policies and customer privacy," *Information Management & Computer Security,* vol. 11, no. 1, pp. 19-29, 2003.

[6] Hughes, A.M. "Boosting Response with RFM," *American Demographics*, May 1996, pg. 4.

[7] Hughes, A.M. "Strategic Database Marketing, 2nd ed.," *McGraw-Hill*. New York, NY. 2000.

[8] Muralidhar, K., R. Parsa and R. Sarathy, "A General Additive Data Perturbation Method for Database Security," *Management Science*, vol. 45, no. 10, pp. 1399-1431, 1999.

[9] Gangopadhyay, A. and Ahluwalia, M. "Preserving Privacy in Mining Association Rules," *The Second Secure Knowledge Management Workshop (SKM)*, Brooklyn, New York, 2006.

[10] Gopal, R.D. and Goes, P.B. and Garfinkel, R. S. "Interval Protection of Confidential Information in a Database," *INFORMS Journal on Computing*, vol. 10, no. 3, pp. 309-322, 1999.

[11] Drozdenko, R.G. and Drake P.D. "Optimal Database Marketing," *Sage Publications, Inc.*, Thousand Oaks, CA, 2002.

[12] Wang, H., K. O. Lee and C. Wang. "Consumer Privacy Concerns about Internet Marketing," *Communications of The ACM*, vol. 41, no. 3, pp. 63-70, March 1998.

[13] Wilson, R. and Rosen, P.A. "Protecting data through 'perturbation' techniques: The impact on knowledge discovery in databases," *Journal of Database Management*, vol. 14, no. 2, pp. 14-26, 2003.

Yong Jick Lee is an assistant professor at the Department of Business Administration, Jungwon University, Goesan, South Korea. His research interest is primarily in the area of consumer privacy and security issues in businesses and e-commerce environment. His earlier works appeared in some notable journals and conferences including *Decision Support Systems* and *Journal of Information Systems*.

APPENDIX A.
DESCRIPTIVE STATISTICS FOR EACH DATASET

Set 1 - Original Dataset

	Attributes		
	Recency	Frequency	Monetary
Mean	451.5	4.4	71.9
St. Dev.	309.8	7.0	9.3

	Covariance Matrix		
	Recency	Frequency	Monetary
Recency	96003.96		
Frequency	-482.17	48.82	
Monetary	0.99	-42.38	86.38

Set 2 - GADP Dataset

	Attributes		
	Recency	Frequency	Monetary
Mean	451.0	4.5	71.9
St. Dev.	310.6	7.0	9.3

S1	1.30

	Covariance Matrix		
	Recency	Frequency	Monetary
Recency	96480.38		
Frequency	-457.49	49.55	
Monetary	0.18	-24.44	86.83

Set 3 - Partitioned GADP (quintile) Dataset

	Attributes		
	Recency	Frequency	Monetary
Mean	451.2	4.5	71.9
St. Dev.	309.6	7.0	9.3

S1	1.13

	Covariance Matrix		
	Recency	Frequency	Monetary
Recency	96031.13		
Frequency	-480.76	49.53	
Monetary	0.98	-41.44	86.08

Set 4 - Partitioned GADP (decile) Dataset

	Attributes		
	Recency	Frequency	Monetary
Mean	451.5	4.4	71.9
St. Dev.	309.8	7.0	9.3

S1	1.02

	Covariance Matrix		
	Recency	Frequency	Monetary
Recency	96604.97		
Frequency	-498.88	50.12	
Monetary	1.36	-50.84	86.05