WP. 17
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Disclosure risk when responding to queries with deterministic guarantees

Prepared by Krish Muralidhar, University of Kentucky and
Rathindra Sarathy, Oklahoma State University

Disclosure risk when responding to queries with deterministic guarantees

Krish Muralidhar[*] and Rathindra Sarathy[**]

[*] University of Kentucky, Lexington, KY 40506, USA (krishm@uky.edu)
[**] Oklahoma State University, Stillwater, OK 74078 USA (rathin.sarathy@okstate.edu)

Abstract: With greater computing and communication capabilities, statistical agencies have shown an increased interest in providing users with the ability to issue ad hoc queries regarding confidential data and receive responses to these queries. Most query/response systems attempt to provide enhanced results to specific user queries. While this type of system allows greater flexibility and perhaps enhanced usefulness, they also create new disclosure risk issues. Of particular concern are query/response systems where the responses are based on a "deterministic" approach. Our definition of "deterministic" is somewhat broader than previous definitions of the term. Using the example of Confidentiality via Camouflage, a procedure intended to provide an interval response to queries, we illustrate that the disclosure risk associated with such systems can be very high. In some cases, we show that these systems could result in complete, deterministic disclosure of the entire data. Based on our results, we urge statistical agencies to carefully evaluate the disclosure risk associated when implementing query/response systems for confidential data.

1 Introduction

The computing ability and communication capabilities have increased considerably over the last few years. With this increase, there has been an increased demand from users for the ability to issue ad hoc queries to statistical databases. Statistical agencies are under increasing pressure to provide users with this ability. Statistical agencies have responded with the proposal to create remote data access with the ability to issue and receive responses to ad hoc queries.

One of the major problems with ad hoc queries is the question of analytical validity (or usefulness of the responses). With traditional methods (such as microdata access), it is possible for the data administrator to provide some general guarantees regarding the validity of the data for analytical purposes. With ad hoc queries, it is very difficult to provide such measures of analytical usefulness. In some cases, it may be possible to provide error bounds for the data and/or the statistic being released (such as a confidence interval). An alternative is to provide the responses themselves in the form of an interval. While these improve the quality of the responses, they also have the potential for increased disclosure.

The traditional methods of data access (primarily microdata release) have been well studied and understood. With the increase in remote data access and ad hoc queries, there is likely to be demand for new techniques that facilitate this process. However, it is important that any new techniques are evaluated with the same level of rigor as

microdata release techniques. Failure to do so can result in disclosure of confidential information. The objective of this study is to illustrate that failure to rigorously evaluate disclosure risk characteristics of new techniques can have serious consequences. Specifically, we illustrate the disclosure risk associated with Confidentiality via Camouflage (CVC), a new method intended to provide "correct responses to ad hoc queries to a database." (Garfinkel et al. 2002). Using the CVC method, we illustrate that a masking procedure with deterministic characteristics (either in the response and/or in the construction of the masking) will result in a much higher level of disclosure. In the worst case, we show that such methods can result in complete deterministic disclosure of the entire database.

2 CVC for Binary Data

There are few masking mechanisms that provide interval responses to queries. Of these, the CVC approach is probably the best known one. Hence, in evaluating the disclosure risk associated with interval responses, we will use the CVC approach. As the name implies, CVC attempts to "camouflage the confidential vector $\boldsymbol{a}$ in a set of vectors and then answer all queries as if they pertained to the entire set." (Garfinkel et al. 2002).

For binary CVC (BCVC), the database administrator must select the number of vectors k among which the confidential binary data $\boldsymbol{a}$ is to be hidden. As we will discuss later, larger values of k increase the security but also increases computational complexity. The authors then set about creating $\boldsymbol{V} = \{\boldsymbol{V^1}, \boldsymbol{V^2}, \dots, \boldsymbol{V^k}\}$ where $\boldsymbol{V^j} = \left(v_1^j, v_2^j, \dots, v_n^j\right)$. The vector $\boldsymbol{a}$ is then arbitrarily assigned $\boldsymbol{V^j}$. For simplicity and without loss of generality, we will set $\boldsymbol{V^k} = \boldsymbol{a}$. Thus, the individual value a_i is now hidden among the values $\left(v_1^1, v_1^2, \dots, v_1^k\right)$ (and $v_1^k = a_1$). CVC also requires that given a_i at least one of the values $\left(v_i^1, v_i^2, \dots, v_i^{k-1}\right) \neq a_i$, while the other values can be assigned randomly.

As a simple illustration consider the example provided in Garfinkel et al. (2002, Table 1, page 750) consisting of a binary confidential variable for (n =) 14 observations. In the example, the values in $\boldsymbol{a}$ represent those who tested negative (0) and positive (1) for HIV. This camouflaged data is reproduced below in Table 1.

Now consider a query of the form $COUNT(a_1, a_2)$ (which is the equivalent of the SUM query). Based on the information in Table 2, the response from BCVC for this query would be the interval $\left[MIN\left(COUNT(a_1, a_2)\right),\ MAX\left(COUNT(a_1, a_2)\right)\right]$ = [0, 2]. It is easy to verify that when responses are provided in this form, the construction of the $\boldsymbol{V}$ vectors guarantees that the resulting interval always contains the true value. For a single binary variable, since the $COUNT$ query is the only meaningful query

that can be issued, the BCVC response interval always contains the true response to the query.

Record	V^1	V^2	$V^3 = a$
1	0	1	0
2	1	1	0
3	1	0	0
4	1	0	1
5	0	1	0
6	0	1	0
7	1	0	1
8	1	0	1
9	1	1	0
10	1	0	1
11	0	0	1
12	1	0	0
13	0	0	1
14	1	0	0

Table 1. Camouflaged Binary Example

Record	V^1	V^2	$V^3 = a$	**Response**	
				Lower Limit	**Upper Limit**
1	0	1	0	0	1
2	1	1	0	0	1
$COUNT(a_1, a_2)$	1	2	0	0	2

Table 2. Response to $\boldsymbol{COUNT}(a_1, a_2)$

The advantages of the BCVC approach in a remote data access system are obvious. First and foremost, it is capable of responding to ad hoc queries. It provides an interval response which allows the intruder to directly evaluate the accuracy of the response. It handles additions and deletions to the system very easily. BCVC seems ideally suited for remote data access systems. In terms of disclosure, BCVC claims that it provides both "deterministic and stochastic protection of the confidential data" (Garfinkel et al., 2002) since every record is represented by the interval [0, 1].

3 Deterministic Components of CVC

One of the attractive features of the CVC attractive for remote data access situations is the fact that it provides the *deterministic guarantee* that the interval response for

any query will contain the true value. It is this deterministic guarantee that allows the user to evaluate the usefulness (or accuracy) of the response. In order to ensure that the true value is always contained in the response interval, it is necessary that the camouflage vectors have a *deterministic relationship* with the true value vector. In BCVC, this relationship is obvious. In order for every system response to contain the true value, *it is necessary that one of the camouflage vectors be the true value vector*. In other words, the *deterministic component (the true value vector) of the masking process is necessary in order to maintain the deterministic guarantee that the interval response contains the true value* of the BCVC procedure.

Traditionally, deterministic methods of data masking have very poor disclosure risk characteristics. However, the claim of deterministic and stochastic protection is based almost exclusively on the single characteristic that every confidential value is represented by [0, 1]. On closer examination however, it is evident that BCVC does not provide adequate security.

3.1 Search Procedure to Reconstruct V

In this section, we show that it would be very easy to reconstruct $\boldsymbol{V}$ using responses to simple queries. The compromise procedure is as follows.

(1) Select a small subset of the data $\boldsymbol{a^{sub}}$ of size m $(m \ll n)$ such that 2^m is within computational capability.
(2) Issue every possible $COUNT$ query involving the m records and store the corresponding responses. This results in a total of $(2^m - 1)$ queries and responses.
(3) Evaluate all possible combinations of values for $\boldsymbol{a^{sub}}$ and identify combination of values that satisfy all queries and responses from the previous step. The evaluation involves a total of 2^m combinations of values.

Assume that the data in Table 1 represents a subset $(m = 14)$ of a much larger database. This subset results in a total of 16383 $(2^{14} - 1)$ query/response combinations. The search procedure described above was performed on this subset. The search procedure results in identifying three possible vectors (Table 3) that satisfy all query/response combinations. It is easy to verify that the three vectors in Table 3 are identical to those in Table 1 except that they are in different order (Vector 1 = $\boldsymbol{V^3}$; Vector 2 = $\boldsymbol{V^2}$; Vector 3 = $\boldsymbol{V^1}$). Since order is irrelevant, *the search procedure successfully identifies* $\boldsymbol{V}$. In addition, by repeating this process for the remaining records by selecting additional subsets, *an intruder can easily reconstruct* $\boldsymbol{V}$ *for the entire database*.

Record	**Vector 1**	**Vector 2**	**Vector 3**
1	0	0	1
2	0	1	1
3	0	1	0
4	1	1	0
5	0	0	1
6	0	0	1
7	1	1	0
8	1	1	0
9	0	1	1
10	1	1	0
11	1	0	0
12	0	1	0
13	1	0	0
14	0	1	0

Table 3. Vectors Identified by Search Procedure

It is relatively easy to explain the reason for the ability to reconstruct $\boldsymbol{V}$ using query responses. With $m = 14$ and $k = 3$, there are a total of 42 variables (unknown values). The query/response combinations essentially result in 16383 constraints to solve for these 42 variables, leading to the solution. While it is true that the search procedure does not guarantee the candidate vectors will be the camouflage vectors, given the ratio of variables to constraints, it is likely. In addition, the addition of one more record to this data set would *increase the number of variables by 3*, but would *double the number of constraints* thereby increasing the probability of reconstructing $\boldsymbol{V}$ (albeit with increased computational ability). Finally, given that this data subset has already been compromised, in many cases, it would be possible to compromise an additional record on an individual basis (suing query/responses involving the unknown record and the already compromised records). Thus, once a subset has been compromised, the intruder's ability to compromise the remaining observations is not limited by computational ability.

Note that the release of just this information constitutes disclosure for the following reasons:

(1) Even though the intruder may not know the true value vector, the knowledge that one of these is guaranteed to be the true vector alone constitutes disclosure.
(2) Even if every data point is not disclosed by the release of this data, it is easy to identify that the values of records {1, 5, 6}, {2, 9}, {3, 12, 14}, {4, 7, 8,

10}, and {11, 13} are the same. Hence, the owner of record 4 how is HIV positive knows that records 7, 8, and 10 are also HIV positive.

(3) The owners of records {2, 9, 11, 13} can identify that Vector 1 must be the true value vector since that is the only vector for which contains the true value for this individual.

Thus, these results indicate that the BCVC method is subject to exact reconstruction by an intruder. Further, since the number of camouflage vectors k is always very small, the intruder is guaranteed that the number of potential solutions is also very small. Taken together, this implies that it is very likely that the intruder will be able to compromise the entire database.

4 CVC for Numerical Data

Gopal et al. (2002) described a CVC approach for numerical data (NCVC), that is broadly similar to the binary CVC approach. Just as with the binary CVC, for numerical data, the original confidential data $\boldsymbol{a}$ is hidden among a set of k camouflage vectors $\boldsymbol{P} = \{\boldsymbol{P^1}, \boldsymbol{P^2}, \ldots, \boldsymbol{P^k}\}$. However, unlike the binary case, the true vector is *not* assigned to one of the $\boldsymbol{P}$ vectors. Instead, for a given record i, the values $\{P_i^1, P_i^2, \ldots, P_i^k\}$ are selected such that at least one of the values is less than a_i, at least one value great than a_i, and the constraint that $\sum_{j=1\ to\ k} P_i^j \gamma_j = a_i$. The values of γ_j are randomly selected with the constraint that $\sum_{j=1\ to\ k} \gamma_j = 1$. In this implementation, the database administrator specifies k and generates the values of γ_j.

As before, the deterministic features of NCVC are necessitated by the deterministic guarantee that the response will interval will contain the true value. In this case, the true value vector $\boldsymbol{a}$ is not a part of $\boldsymbol{P}$. However, for any given record i, we can write $\sum_{j=1\ to\ k} P_i^j \gamma_j = a_i$. Hence, there is a deterministic relationship between the true value and the camouflage vectors. In order to ensure that the interval responses to all queries will contain the true value, it is necessary to maintain the deterministic relationship between $\boldsymbol{a}$ and $\boldsymbol{P}$.

For linear queries, the response from the numerical CVC approach is identical to the binary CVC approach where the responses are constructed as if they pertained to the entire $\boldsymbol{P}$ instead of $\boldsymbol{a}$. The released response would then be the maximum and minimum from these computations. Gopal et al. (2002) provide derivations detailing the computation of responses to more complicated queries such as variance, correlation, etc. Many of these responses would require the solution to non-linear programs. Hence, the computational complexity increases considerably with increase in the number of camouflage vectors k. The interested reader is referred to Gopal et al. (2002) for a comprehensive discussion.

Table 2 provides an example of the numerical CVC approach for dataset with n = 14, k = 3, and $\gamma = \{\gamma_1 = 0.20, \gamma_2 = 0.30, \gamma_3 = 0.50\}$. Note that this example is the same as the one used by Gopal et al. (2002, Table 3, page 505) to illustrate the numerical CVC approach.

Record	P^1	P^2	P^3	a
1	60	53	54.2	55
2	31	29	32.2	31
3	99	110	108.4	107
4	28	31	26.2	28
5	53	64	66.4	63
6	78	82	83.6	82
7	30	28	29.2	29
8	29	31	31.8	31
9	63	60	58.8	60
10	26	27	27.4	27
11	46	50	45.6	47
12	34	31	31.8	32
13	91	100	91.6	94
14	51	51	51	51

Table 4. Camouflaged Numerical Example

For the purposes of this study, we will limit our discussion to simple linear queries. Consider the queries $(a_1 + a_2)$ and $(a_1 - a_2)$. The response would be computed based on the information presented in Table 5. The response interval for the query $(a_1 + a_2)$ would be the minimum and maximum of the $(a_1 + a_2)$ row namely [82, 91]. Similarly, the response interval for the query $(a_1 - a_2)$ would be [22, 29].

Record	P^1	P^2	P^3	a	**Response**	
					Lower Limit	**Upper Limit**
1	60	53	54.2	55	53	60
2	31	29	32.2	31	29	32.2
$(a_1 + a_2)$	91	82	86.4	86	82	91
$(a_1 - a_2)$	29	24	22.0	24	22	29

Table 5. Computing Responses for Numerical Data

4.1 Search Procedure to Compromise Numerical Data

In performing the search, we assumed that the adversary has knowledge that the confidential attribute is integer. Since masking mechanisms are intended to prevent disclosure of confidential information to legitimate users, it is reasonable to assume that the user is provided meta knowledge regarding the attribute of which the precision with which the attribute is measured is an important one. Hence, it is reasonable to assume that the adversary/user has knowledge that the attribute is integer. Furthermore, since any attribute measured with a particular precision can always be transformed into an integer value by simple multiplication, we can use this approach for any data.

For any given record, the search space is easily identified by issuing a singleton query. For example, the query "Value of record 1" results in the interval response (53, 60) implying that the search space for record 1 are the integer values between 53 and 60 (both values inclusive). Ignoring the last record (which has not been protected), the number of potential candidate solutions is for the data in Table 4 is 2,903,040,000. Given the number of potential candidate solutions, a slightly modified, more intelligent version of the brute force search procedure employed for the binary data was used for the numerical data.

In the modified procedure, the search procedure was conducted using the first five records in the data set. Using the responses to queries involving these five records, the potential candidate solutions were identified. The process was then repeated by incrementally adding one more record to the search process. This search reduces the potential candidate solutions as we increase the number of records being considered. Reducing the number of potential candidate solutions for a small number of records means that as we add records, the number of evaluations is much lower. In the worst case, the computational requirements are no worse than before. Thus, computationally, this procedure is *always* superior to the brute force search using all records. The result of the search procedure is provided in Table 6.

Comparing Tables 4 and 6, it is easily verified that candidate solution 3 is the true value vector. As with the binary case, from Table 3, it is easy to see that knowledge of the value of a single record is adequate to compromise the entire database. For example, if the adversary is aware that the value of the first record is 55, then the only candidate that could possibly be the true solution is candidate solution 3. Thus, this knowledge completely discloses the database. This is true when the adversary has knowledge of any one of the following records {1, 3, 5, 7, 11, 12, 13}. For all other records, it would be necessary to have knowledge of the true value of two records to compromise the data. Thus, the disclosure risk characteristics observed in the binary CVC method also allows for the compromise of numerical CVC as well.

Record	Candidate Solutions		
	1	2	3
1	60	53	55
2	31	29	31
3	99	110	107
4	28	31	28
5	53	64	63
6	78	82	82
7	30	28	29
8	29	31	31
9	63	60	60
10	26	27	27
11	46	50	47
12	34	31	32
13	91	100	94
14	51	51	51

Table 6. Candidate Solutions for Numerical Data

5 Conclusions

Our objective in this paper was to illustrate the fact that data masking techniques that have a deterministic component are subject to deterministic disclosure as well. It is generally well understood that simplistic deterministic masking techniques (such as adding a constant perturbation term) are subject to high levels of disclosure. Very often however, some techniques that do not seem to be deterministic may in fact have a deterministic component in them that may result in high levels of disclosure. This is the case with the CVC approach. In binary and numerical CVC there exists a purely deterministic relationship between the camouflage vectors and the true value vector. This deterministic relationship leads to deterministic disclosure.

The deterministic masking encountered in CVC is necessary to satisfy the deterministic guarantee offered by the method. CVC offers the deterministic guarantee that the true value is contained in the response interval. Dinur and Nissim (2003) showed that with information on the overall quality of the responses, an intruder can reconstruct the original database with some accuracy. The quality of the responses is usually characterized by the statement that responses to queries are within some common overall error bound that is applicable to all queries. *The CVC provides an explicit deterministic guarantee for the quality of every ad hoc query* which goes far beyond a generic guarantee of all queries. In order to satisfy this guarantee, in the CVC method there exists a deterministic relationship between the

camouflage vectors and the true value vector, which leads to the (inevitable) deterministic disclosure of the confidential data.

References

Dinur, I. and Nissim, K. (2003) Revealing Information while Preserving Privacy. PODS 2003, San Diego, CA, 202-210.

Garfinkel, R., Gopal, R. and Goes, P. (2002) Privacy Protection of Binary Confidential Data Against Deterministic, Stochastic, and Insider Threat. *Management Science*, 48, 749-764.

Gopal, R., Ganfinkel, R. and Goes, P. (2002) Confidentiality via Camouflage: The CVC Approach to Disclosure Limitation when Answering Queries to Databases. *Operation Research*, 50, 501-516.