COMPUTER STUDIES PUBLICATION No TR-A5-83

Technical Report

QUERY COMPLEXITY FOR STATISTICAL DATABASE SECURITY

M Y Chan

CENTRE OF COMPUTER STUDIES AND APPLICATIONS

UNIVERSITY OF HONG KONG

POKFULAM ROAD

HONG KONG

Mee Yee Chan

Centre of Computer Studies & Applications

University of Hong Kong

ABSTRACT

The problem of protecting statistical databases from compromise while satisfying arbitrary on-line aggregate requests is approached afresh. New perspective is gained through the introduction of the concept of query complexity for information release. Compatible with complexity a new distortive plan is proposed that provides effective protection against compromise, easy implementation, and viable statistics, and is very competitive with existent policies. Experimental results on simulated databases support the proposal.

1. INTRODUCTION

Privacy considerations are of prime importance in the dissemination of statistics for databases which contain confidential facts about individuals. With the advent of modern database management systems accommodating users with on-line aggregates for arbitrary subpopulations, the problem takes added dimensions. While this flexibility renders the statistical database an invaluable tool for research and planning, the powers of the malicious snooper are magnified. It is well-noted that given such querying freedom, even with the suppression of implicating attributes such as name, address and social security number, identification of an individual and disclosure of his information are often possible [11].

In the interest of statistical database security, past research prescribes either a curtailment of querying freedom through restrictions or a sacrifice of statistical exactness through distortions. Investigated protection mechanisms include the restrictions of query set size restriction [2], partitioning [16], statistical database design [3] and threat monitoring [11, 14] and the distortions of roundings [1, 4, 9, 12], random sampling [8] and data swapping [5, 15]. Unfortunately, with few exceptions the specific proposals meet with little success. Misfortunes encompass costly implementations, gross inflexibilities or inaccuracies, and susceptibility to attacks of rephrasing epitomized by trackers [6, 7, 13] and attacks of inconsistency exploitation [10].

With this bleak view of the state of the art of statistical inference control, we seek to remedy the situation through a new framework for information release that offers effective protection against compromise, easy implementation, and reasonable statistics.

We approach the statistical database dilemma afresh with the rudimentary question of query intention as the point of departure. Much of past literature dwells on the concept of query set size as the foremost indicator of intention. The belief seems to be that interrogations of large populations signify truly statistical intent while small signify malicious intent. Indeed there is eagerness to produce distortive devices that effect error inversely proportionate to query set size and restrictive counterparts that preclude the small-cardinality request from being answered. However, pure reliance on query set cardinality has elicited too few viable safeguards, for although small-size queries invite privacy invasion

or compromise, large do not guarantee bona fide intentions. That cardinality is misleading is demonstrated by trackers.

All statistical database security measures embody an interpretation of query intention on which basis they restrict and distort. The success and direction of a strategy depends upon this interpretation. So in search of truer perspective, we augment the cardinality concept. In particular, we propose a factor of query complexity in intention analysis. If we describe query populations by characteristic formulas [6], query complexity is essentially the complication of the characteristic formula. We stand by the principle that the request that isolates a single individual, reguardless of it wording, incorporates a more complicated query set description than the sincere statistical appeal. Thus, query interpretation which emphasizes complexity translates to the following :

(i) low-complexity, large-size inquiries are deemed innocuous;

(ii) low-complexity, small-size inquiries are thought to intend no harm but happen to assist compromise;

(iii) high-complexity, large-size inquiries are thought to result from rephrasing attempts; and

(iv) high-complexity, small-size inquiries are deemed malicious.

We adopt this viewpoint for statistical database security.

In Section 2 we shall elaborate further on our perception of the statistical query and its complexity. In Section 3 a procedure for information release based upon complexity-emphasized query interpretation will be outlined. Section 4 evaluates the procedure in terms of noncompromisability, implementation ease, accuracy and

flexibility. Section 5 concludes. Some simulation results are presented in the appendix.

2. THE STATISTICAL QUERY AND ITS COMPLEXITY

Our perception of the statistical query begins with the notion of a characteristic formula which in turn rests on the characteristic. A characteristic <A, v> is simply an association of a value v to an attribute A. <SEX, male>, <AGE, 35>, <PROFESSION, lawyer> are examples. A characteristic formula is a logical expression with operators of AND, OR, NOT and characteristics as terms. Characteristic formulas describe query populaions. Hence, <SEX, male> AND (NOT <AGE, 35>) alludes to the set of all males who are not 35 years old. Parentheses are added for clarity. A statistical query q(C) asks for the count, or average or percentile (percentiles include minimum, maximum, median) for a particular attribute (such as salary), of the population characterized by formula C.

Query complexity is a function of the characteristic formula C in a statistical query q(C). In accordance with intuition for query complexity is the following recommendation. Let

$$Q(C) = N \,/\, \prod_{x \in X(C)} w(x) \ ,$$

where N is the number of individuals represented in the database, X(C) is the set of attributes referenced in formula C, and w(x) is the number of values attribute x may assume. Since q(C) hints an apportionment of individuals into $\prod w(x)$ partitions, under the assumption that individuals take on values for attributes with equal

chance, Q(C) is a lower bound for the expected number of persons which C may address. Thus, with low Q correlating high- complexity and high Q correlating low-complexity, we find satisfaction in having high-complexity queries likely to refer to smaller populations than low-complexity and satisfaction in having the property that

(i) $Q(C) = Q(\text{NOT } C)$,

(ii) $Q(C) \geq Q(C \text{ AND } D)$, and

(iii) $Q(C) \geq Q(C \text{ OR } D)$

for all characteristic formulas C and D. Also, a query applied to a bigger database appears less malicious through Q(C)'s dependence on N, and a query which say asks for the population of all males usually arouses more suspicion than a query which asks for the population of all 35 year-old persons through Q(C)'s dependence on attribute sensitivity w(x). Furthermore, it is ensured that rephrasing a request originally in its simplest, least-complicated form ultimately involves a query of similar or worse complexity. To see this consider q(C) and its rephrasing taken to involve queries q(C OR T) and q(T) where T is a characteristic formula. We readily have the required complexity query in q(C OR T). However, it may be the case that a less-complicated-than-q(C) query exists and serves the same purpose as q(C OR T); then q(T) can be argued to be of similar complication to q(C). This reasoning expressly applies to tracker rephrasings (T becomes a tracker). As for others, analogous lines of thought apply.

Granted that our final objective were to rank a query by quality of complexity onto a scale of 1, ..., m, it would be necessary to map real-valued Q(C), $0 < Q(C) \leq N$, to integer L(C), $1 \leq L(C) \leq m$, remembering to relate low Q(C) to high L(C). Since this mapping of

Q(C) is more of an implementation issue, we postpone a recommendation for it until the next section.

3. DISTORTIVE INFORMATION RELEASE

Having explained our idea of query complexity, let us illustrate its application as part of a distortive procedure for information release. Compatible with m levels of complexity, we prescribe an error innoculation system involving m levels of pseudorandom roundings of data. Each successive level reveals a potentially more distorted version of the statistical database. The procedure is as follows :

(1) Decide on integers m, $k \geq 1$, with m being the number of levels of complexity and distortion and k as a Q(C)-mapping-to-L(C) parameter that is somewhat analogous to query set size restriction's minimum size for allowable queries.

(2) For each statistical attribute A, an attribute for which averages or percentiles may be sought, compute the standard deviation of its values SD(A). Let r(A) = SD(A) / m. r(A) i is the rounding base for the ith level, i = 1, ..., m.

(3) For each individual I associate m+1 perturbations

$p_1(I), \ldots, p_m(I)$ and $p(I)$

where each perturbation is pseudorandomly generated as -1, 0, or 1 with equal probability.

(4) In regards to each query q(C), let

$$L(C) = \begin{cases} 1 & \text{if } Q(C) > k\,2^{m-1}, \\ \max\{i \in \{1,\ldots,m\}: Q(C) \leq k\,2^{m-i}\} & \text{otherwise.} \end{cases}$$

(a) For q(C) which asks for the average or percentile of attribute A, if individual I satisfies formula C and has characteristic <A, v>, use

$$v'(L(C)) = v + r(A) \sum_{i=1}^{L(C)} p_i(I)$$

instead of v to compute the statistic. Error in response is acknowledged as $\pm$ r(A) L(C) in the worst case with error for large-size, average queries likely to be much less.

(b) For q(C) that asks for the count, if individual I satisfies C, use 1 + p(I) instead of a contribution of 1 in the enumeration. Percentage error is acknowledged to likely decrease with increasing query set size.

Note that participation of an individual is based on true values rather than distorted; in other words, the actual response set is considered whereby special properties may be preserved.

4. EVALUATION OF PROPOSAL

Firstly, we see that the proposal of Section 3 is indeed harmonious with our query complexity and cardinality interpretation of query intention :

(i) Precise averages and percentiles with likely accurate counts accompany low-complexity, large-size requests.

(ii) Precise averages and percentiles with possibly inaccurate

counts accompany low-complexity, small-size requests.

(iii) Imprecise averages and percentiles with likely accurate counts accompany high-complexity, large-size requests.

(iv) Imprecise averages and percentiles with possibly inaccurate counts accompany high-complexity, small-size requests.

Direct compromise is resisted through the workings of (ii) and (iv) wherein small counts may be grossly distorted. Additionally in (iv), a case where a snooper may venture to decide the count value by preknowledge, we use extremely imprecise data values for statistics.

Insofar as indirect compromise strategies of rephrasing and inconsistency exploitation are concerned, there is also a display of resistance. Rephrasings of count queries do not induce the inconsistency necessary to arrive at better responses. m levels of direct data distortion manage the same effect for averages and percentiles. The distortion essentially amounts to a generalization of traditional (one-level) data distortion, with reduced tendency for small perturbations to facilitate compromise and for large to undermine utility, while keeping with its immunity to rephrasing and inconsistency exploitation within each level. Consistency is also established across levels. At level i, for $i = 1, \ldots, m-1$, a perturbed value for characteristic $\langle A, v \rangle$ is acknowledged as $v'(i) \pm r(A)\, i$. Its value at level i+1 is reported as $v'(i+1) \pm r(A)\,(i+1)$. Since the range of values $[v'(i+1) - r(A)\,(i+1),\ v'(i+1) + r(A)\,(i+1)]$ is guaranteed to include $[v'(i) - r(A)\, i,\ v'(i) + r(A)\, i]$, the response of $v'(i+1)$ is entirely consistent with $v'(i)$. Also, to guard against inconsistency derived from comparing averages and counts to totals, we have deliberately disallowed explicit requests for totals. Lacking the zero-bias trait, where responses are

unbiased estimators of true answers, averaging ploys lose strength. Thus, indirect attacks are thwarted.

In terms of implementation costs, the scheme demands little effort. Larger choices for the number of perturbation levels m increase cost for implementation as the spectrum of possible data precisions widen to include higher degrees of exactness. The second parameter k correlates query complexity to the m levels of distortion. In fact, k helps to draw the subtle line between low- and high-complexity. With responses for levels i, ..., m ($1 \leq i \leq m$) deemed imprecise, a high-complexity inquiry q(C) is one whose

$$Q(C) \geq k\, 2^{m-i}.$$

5. CONCLUSION

In summary, we have presented a distortion method that is effective against compromise in the sense that answers given do not disclose confidential facts and cannot be improved via applications of indirect attacks. It is easy to implement and overhead is minimal. Complete querying liberty from arbitrary on-line aggregates is afforded. Reasonably accurate statistics are argued on the grounds of a new and practical query interpretation idea that combines query complexity and query set cardinality.

The scheme embeds a more constructive barrier against trackers than query set size restriction. The difficulty of choosing m and k does not at all compare with the difficulty of choosing appropriate

partitions in partitioning. The implementation hardships and uncertainties of threat monitoring and data swapping are not suffered. Unlike statistical database design, we do not require much support from a database management system, and unlike random sampling auxilliary protection devices against compromise are unnecessary. Because roundings of responses attempt to be unbiased estimators of true answers, they are standardly circumvented by averaging techniques. The best that can be done in these roundings is to make circumvention costly. Hence, we perceive our protection scheme to be competitive with past ideals, and finally, the concept of query complexity is seen with favor and promise for statistical database security.

ACKNOWLEDGEMENT

The author is grateful to Dr. C. K. Yuen for his interest.

APPENDIX.

The following table results from a simulation of the query complexity distortion technique on two artificial student databases of size N = 200 and N= 500 for various choices of m and k. Attributes of interest include sex, major, grade point average, and SAT scores both math and verbal :

SEX : male, female
MAJOR : Math, CS, Phys, Chem, Engl, Biol, Psych
GPA : 1.0-4.0
SATM : 40-80
SATV : 40-80

QUERY SET		true	N=200 m=6 k=1	m=3 k=1	true	N=500 m=6 k=1	m=6 k=5
1. (SATM>60)	level		3	1		2	4
	count	107	104	104	255	247	247
avg	SATM	69.7	70.2	69.8	70.1	70.3	70.8
	SATV	60.9	61.4	61.0	59.9	60.1	60.5
	GPA	2.35	2.39	2.37	2.42	2.44	2.47
max	SATM	80	83	80	80	80	80
	SATV	80	80	85	80	83	86
	GPA	4.0	4.0	4.4	4.0	3.8	4.0
2. (SATM>60) and (SATV>60)	level		6	3		6	6
	count	61	52	52	129	125	125
avg	SATM	69.8	70.5	70.3	70.3	70.2	70.2
	SATV	70.5	71.2	71.0	70.2	70.0	70.0
	GPA	2.34	2.38	2.38	2.44	2.43	2.43
max	SATM	80	83	85	80	80	80
	SATV	80	77	80	80	89	89
	GPA	4.0	4.0	3.2	4.0	4.0	4.0
3. (SATM>60) and (SATV>60) and (GPA>3.0)	level		6	3		6	6
	count	15	13	13	41	40	40
avg	SATM	69.1	69.9	67.5	69.9	69.3	69.3
	SATV	68.3	69.1	66.6	69.1	68.5	68.5
	GPA	3.58	3.63	3.45	3.57	3.53	3.53
max	SATM	78	75	78	80	80	80
	SATV	77	86	82	80	89	89
	GPA	4.0	4.0	3.2	4.0	4.0	4.0
4. (SATM>60) and (SATV>60) and (GPA>3.0) and (SEX=female)	level		6	3		6	6
	count	9	7	7	22	19	19
avg	SATM	69.4	70.8	70.6	70.3	70.5	70.5
	SATV	70.2	71.6	71.3	71.1	71.2	71.2
	GPA	3.63	3.72	3.72	3.57	3.58	3.58
max	SATM	78	75	78	79	67	67
	SATV	77	86	82	80	89	89
	GPA	4.0	3.8	4.0	4.0	3.8	3.8
5. (SATM>60) and (SATV>60) and (GPA>3.0) and (SEX=female) and (MAJOR=Math or Engl)	level		6	3		6	6
	count	1	1	1	8	6	6
avg	SATM	73.0	67.0	58.0	71.0	68.0	68.0
	SATV	77.0	71.0	62.0	72.3	69.3	69.3
	GPA	3.80	3.40	2.60	3.44	3.24	3.24
max	SATM	73	67	58	78	72	72
	SATV	77	71	62	78	81	81
	GPA	3.8	3.4	2.6	3.8	3.4	3.4

REFERENCES.

1. BECK, L. L. A security mechanism for statistical databases. ACM Trans. Database Syst. 5,3 (Sept 1980), 316-338.

2. CHIN, F. Y. Security in statistical databases for queries with small counts. ACM Trans. Database Syst. 3,1 (March 1978), 92-104.

3. CHIN, F. Y. Statistical database design. ACM Trans. Database Syst. 6,1 (March 1981), 113-139.

4. CONWAY, R., AND STRIP, D. Selective partial access to a database. Proc. 1976 ACM Ann. Conf., 85-89.

5. DALENIUS, T., AND REISS, S. P. Data-swapping — A technique for disclosure control. Comput. Sci. Tech. Rep. 39, Brown Univ., Providence, R. I., July 1978.

6. DENNING, D. E., DENNING, P. J., AND SCHWARTZ, M. D. The tracker : A threat to statistical database security. ACM Trans. Database Syst. 4,1 (March 1979), 76-96.

7. DENNING, D. E., AND SCHLORER, J. A fast procedure for finding a tracker in a statistical database. ACM Trans. Database Syst. 5,1 (March 1980), 88-102.

8. DENNING, D. E. Secure statistical databases with random sample queries. ACM Trans. Database Syst. 5,3 (Sept 1980), 291-315.

9. FELLEGI, I. P., AND PHILLIPS, J. L. Statistical confidentiality : Some theory and applications to data dissemination. Ann. Econ. Soc. Meas. 3,2 (April 1974) 399-409.

10. HAQ, M. I. On safeguarding statistical disclosure by giving approximate answers to queries. Int. Computing Symp. 1977, 491-495.

11. HOFFMAN, L. J., AND MILLER, W. F. Getting a personal dossier from a statistical data bank. Datamation 16,5 (May 1970), 74-75.

12. NARGUNDKAR, M. S., AND SAVELAND, W. Random rounding to prevent statistical disclosure. Proc. Am. Stat. Assoc., Soc. Stat. Sect.(1972), 382-385.

13. SCHLORER, J. Identification and retrieval of personal records from a statistical data bank. Methods Inform. Med. 14,1 (Jan 1975), 7-13.

14. SCHLORER, J. Confidentiality of statistical records : A threat monitoring scheme for on-line dialogue. Methods Inform. Med. 15,1 (Jan 1976), 36-42.

15. SCHLORER, J. Security of statistical databases : Multidimensional transformation. ACM Trans. Database Syst. 6,1 (March 1981), 95-112.

16. YU, C. T., AND CHIN, F. Y. A study on the protection of statistical data bases. ACM SIGMOD Int. Conf. Management of Data, 1977, 169-181.

Chan, M.Y.

Query complexity for statistical database security. 1983.