An Integrated Framework for Database Privacy Protection

LiWu Chang and Ira S. Moskowitz
Center for High Assurance Computer Systems, Naval Research Laboratory, Washington, DC

Key words: database inference, Bayesian network, similarity measure, information reduction, restoration

Abstract: One of the central objectives of studying database privacy protection is to protect sensitive information held in a database from being inferred by a generic database user. In this paper, we present a framework to assist in the formal analysis of the database inference problem. The framework is based on an association network which is composed of a similarity measure and a Bayesian network model.

1. INTRODUCTION

As the information explosion has grown, so has the trend of data sharing and information exchange also grown. Accordingly, privacy concerns have reached a critical level [13]. In his report [1], Anderson stated that the combination of birth date and post code (zip code) with data from a health database is sufficient to identify 98% of the UK population! It is certainly a concern for the Icelandic patients' database [11]. Many existing efforts (e.g., [10][11]) have been geared towards the hiding of stored data items and access control. It has been shown that even if the sensitive personal information is hidden, it can be derived from publicly accessible data by means

of inference [2][5][14][15][16][17][21][22]. Denning [6] categorized several different types of attacks and analyzed the protection methods where query returns are statistical quantities (e.g., mean, variance). Hinke's work on deterministically chained related attributes shows how the information can be obtained from non-obvious links [9]. Duncan [5][22] presented cell suppression techniques where the marginal probability distributions are preserved by disturbing the probability mass of component variables. Sweeney's work applies the aggregation operation to the merge of the attribute values [20].

We wish to put the inference problem upon a firm theoretical foundation. The main contribution of this paper is to categorize and discuss inference from different perspectives and represent those different views in a coherent framework. Among the above mentioned approaches, ours and [5] are similar in that both attempt to minimize the information loss for a database user. The difference is that our protection method evaluates values of each data item. At the core of our model a structured representation of probabilistic dependency among attributes is adopted.

Summarizing from previous work, we envision two perspectives of inference characterized by attribute properties. One perspective is about the probabilistic correlation among attributes. A complimentary perspective is that of individuality which emphasizes the uniqueness of each individual data item. For the former, a *Bayesian network* [18] can be used to model correlation relationships among attributes. Let attributes whose information we wish to protect be the *target attributes*. Based on this model, one can evaluate the potential impact that impinges upon a target attribute from information about other attributes, and decide the pertinent protection strategies accordingly. Although the probabilistic method is useful in describing the likelihood of the occurrence of an attribute value, it may be ineffective for identifying which attribute value is unique to a data item. This uniqueness can be deemed as the individuality of a data item. To protect such an attribute value, it is necessary to determine whether other attribute values, or their combinations, provide the

same amount of information as the special one does to the data item. Thus, the identification of individuality is separate from the probabilistic correlation analysis. The proposed framework is the first to integrate these two perspectives.

2. POLICY

We use data modification to ensure high privacy protection. Our concerns are that a user (authorized for limited data access) might be able to combine his/her information with other users, or to simply generate inferences on his/her own, to glean knowledge about data that they should not have access to. Of course we are not concerned with the data originator learning this information. Our privacy policy can be phrased as follows:

- No sensitive information can be inferred from publicly released data.
- No false information is added to the database to increase privacy protection.

Of course we are still allowing ourselves to hide data to increase privacy protection --- we are only disallowing erroneous data. Since protection always involves a certain level of modification to the data, some statistical properties of a database will inevitably be affected --- this is good for privacy concerns but bad for functionality. Our proposed model will incorporate dynamic changes as a result of new attributes being added and new data being collected.

3. INFERENCE

What information needs to be protected in a database? Consider the example medical database as shown in Table 1, where attributes “address”, “age” and “occupation” are the basic personal information, and “hepatitis”, “mental depression”, “AIDS” and “thyroid (function)” are the personal medical records. It is certain information about the unique user identification number “uid” that we wish to protect

(AIDS, suicide, etc.). Our proposed model (referred to as an *association network*) is composed of two components. One component is based on the probabilistic causal network model. The other component describes the functional dependency or the similarity relationships.

Table 1: Data set

uid	addr	age	occup	hepatitis	mental depr.	AIDS	thyroid
1	FC1	67	md	n	norm	n	n
2	Al1	83	mil	y	dep	n	l
3	TC1	43	lwy	y	dep	y	l
4	An1	19	aca	y	dep	y	l
5	WA1	54	pol	y	dep	n	n
6	Al2	28	con	n	norm	n	n
7	Re1	34	lwy	y	norm	n	n
8	An2	32	con	y	dep	y	l
9	FC1	39	aca	y	dep	y	l
10	FC2	44	pol	n	norm	n	n
11	WA2	66	mil	n	dep	n	l
12	An1	23	md	y	norm	n	n
13	TC2	34	con	n	norm	n	n
14	WA3	50	pol	y	dep	y	l
15	Re2	28	con	n	dep	n	l
16	WA4	47	lwy	n	norm	n	n
17	An3	92	aca	n	dep	n	l
18	Re2	28	lwy	y	dep	y	n
19	TC3	49	mil	n	dep	n	l
20	Al3	32	aca	y	norm	n	n

3.1 Identification of Similar Attributes

To prevent inference attacks, information such as a person's name should automatically be removed from the database. However, the removal of the name attribute is hardly adequate. Other attributes, such as a person's address, may reveal essentially the same information and thus, should also be hidden from general users. Consider two attributes in a database and the natural relation given between their attribute values. If this relation is “close” to being a bijection then we say that the attributes are *similar*. In Table 2 we see the relation between “uid” and “address”. If one “uid” corresponds to one “address” value, then “address” is congruent to

"uid", this is not the case. However, the mapping between the two is almost a bijection so they are similar (only three addresses correspond to more than one uid, and in those cases they correspond to two uids). Intuitively, the less the spread of the frequency count shown in the table, the higher the similarity between the target and the candidate attributes.

Table 2: address vs. uid

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FC1	1								1											
Al1		1																		
TC1			1																	
An1				1								1								
WA1					1															
Al2						1														
Re1							1													
An2								1												
FC2										1										
WA2											1									
TC2													1							
WA3														1						
Re2															1			1		
An3																	1			
WA4																1				
TC3																			1	
Al3																				1

The criterion of determining which attributes are similar to the target attribute is quantified in terms of our information theoretical rule.

Definition 1. (Dispersion V)

$$V_i = -\sum_{j=1}^{N} Pr(t_j|c_i)\log(Pr(t_j|c_i)); \quad V = \left(\sum_{i=1}^{M} V_i\right) / M$$

where N and M stand for the number of attribute values of the target attribute T (with values t_j) and candidate attribute C (with values c_i), respectively. V_i is the dispersion measure of the ith attribute value of C, and V gives the total dispersion measure with normalization. A low V score is the selection criteria for similar. Similar attributes are the ones that we want to modify because they give us inference about the target attribute. In terms of the

frequentist's view, we have $Pr(t_j|c_i)=n_{ij}/n_i$, where n_{ij} denotes the frequency count at the ith row and jth column, and n_i is the sum of the ith row. Note that the range of this dispersion measure is from 0 to $\log N$. The minimum occurs when only one entry in each row has a non-zero value. The maximum happens when the mass n_i is evenly distributed over ALL attribute values of T. Given that T="uid" the V-score for C="address" (Table 2) is 3/17=0.18. Note that if the V-score of a candidate attribute C is less than 1, then there exists V_i-scores of C that are equal to 0, for some i. Attribute values that correspond to low V_i-scores are subject to modification.

A candidate attribute can be a combination of several attributes. For instance, the combination of "address" and "mental depression" can uniquely identify each item in the Table 1. Figure 1 shows such a combination. The fact is that a merge of several attributes with high V-scores can yield a low V-score. Using V-scores an indicator, the proposed search evaluates possible combinations of different attributes until a bijection with the target attribute is reached, or a desired V-score is reached. Attributes or their combination with low V-scores are stored.

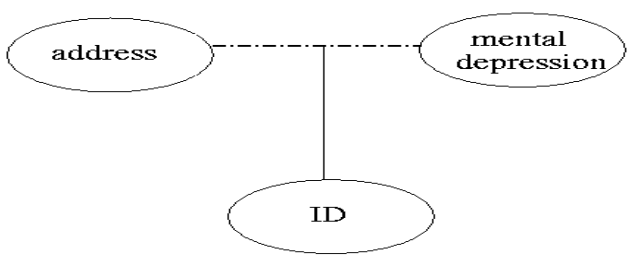

Figure 1: Example of Combination of Attributes. A node represents an attribute. The dashed line denotes the combination and the straight line denotes the similarity relationship.

3.2 Computation of Probabilistic Impact

The analysis of the probabilistic dependency is based on a Bayesian net representation ([8][18]). As shown in Figure 2, either "AIDS" or "thyroid" leads to "mental depression", while "hepatitis" and "mental

depression" support the diagnosis of "AIDS". Thus, "AIDS" can be inferred from information about "hepatitis" and "mental depression". Note that attributes about a person's background are not included in this figure because of the low statistical significance due to their large sets of attribute values.

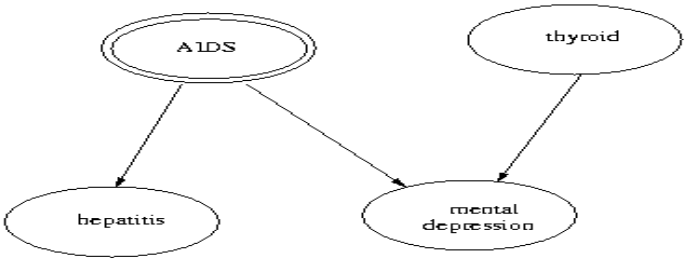

Figure 2: Architecture of a Bayesian network. An attribute is denoted by a node. An arrow indicates the probabilistic dependency between the two attributes. A double circle denotes information associated with the attribute is confidential.

As mentioned earlier, the combination of "address" and "mental depression" will lead to the identification of "uid". Thus, one may able to infer about whether a particular person contracts AIDS by joining together the information from Figure 1 and Figure 2. The joined network is shown in Figure 3. To prevent the potential association of "uid" and "AIDS", information, in particular, "mental depression" (since it contributes to both networks) must be reduced. To protect sensitive information, strategies of blocking and aggregation are used.

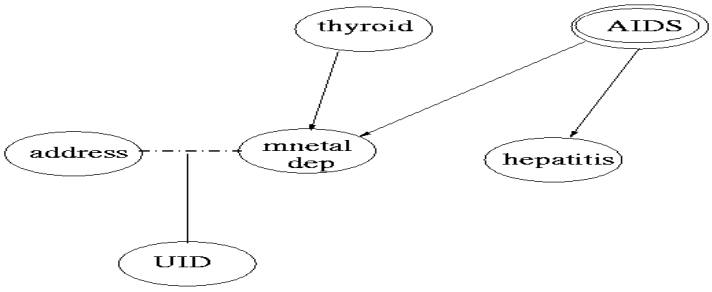

Figure 3: Architecture of a joined network

4. INFORMATION REDUCTION

In this paper, we consider the database modification strategies of blocking and merging. The purpose of modification is to mitigate database inference.

4.1 Reduction range

To give an objective quantitative description of the extent to which users are willing to tolerate the potential error induced from database modification, we invoke a quality index (QI) of a database. QI is generated during the data collection phase. It is represented as the logarithm (base 2) of the sample probability in our analysis:

Definition 2. (QI) $QI \equiv log(Pr(D|\boldsymbol{m})),$

where D denotes the data set and $\boldsymbol{m}$ denotes a model. If $\boldsymbol{m}$ is a Bayesian network model $\mathbf{B_n}$ then QI will be log Pr($D|\boldsymbol{B_n}$). QI is viewed as the lower bound of the level of tolerance, below which the validity of inference drawn from the modified database is in doubt. The operation range is defined in terms of the rate of change, γ.

Definition 3. (Ratio of Reduction)

$\gamma \equiv |QI_\{original\} - QI_\{modified\}| \ / \ |QI_\{original\}|$

For instance, if the original QI is -60 and the QI of the modified database is -63, then the allowed rate of change, γ, is 5%. Our assumption is that the estimated inherent error in the original data and the tolerance measure of how much we are allowed to perturb the data are tied together in some underlying basic manner.

4.2 Blocking

The approach of blocking is implemented by replacing certain attribute values of some data items with a question mark --- this indicates total ignorance of the preference [2]. The set of attribute values that maximally change the posterior probability of the desired target value Pr($T=t_j|D_m,\boldsymbol{B_n}$), with respect to

the modified database D_m and the given $\boldsymbol{B_n}$, are chosen for blocking. If the modification can cause drastic change to the present belief, it should be considered for hiding. The modification will stop when the change reaches beyond the specified γ.

Claim 1. The QI, $\log(\Pr(D|\boldsymbol{B_n}))$, is monotonically decreasing as more attribute values are blocked.

As an example, let the allowed rate of change γ be 3%. From Table 1, the 3% change of QI whose value changes from $\log(\Pr(D|\boldsymbol{B_n}))=-38.85$ to $\log(\Pr(D_m|\boldsymbol{B_n}))=-40$ can be best achieved by modifying
Data item 3: "hepatitis" = "y" as well as Data item 4: "mental depression" ="dep". The result of the released database is shown in Table 3. Since modification inevitably weakens the probabilistic dependency, it may lead to the change of network topology $\boldsymbol{B_n}$. Thus, the causal dependency of the target also needs to be re-evaluated.

Table 3: medical records released to generic users

hepatitis	n	y	?	y	y	n	y	y	y	n	n	y	n	y	n	n	n	y	n	y
mental	n	d	d	?	d	n	n	d	d	n	d	n	n	d	d	n	d	d	d	n
AIDS	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
thyroid	n	l	l	l	n	n	n	l	l	n	l	n	n	l	l	n	l	n	l	n

4.3 Aggregation

We apply an aggregation operation [17] for combining different values of an attribute of low V_i-score. Aggregation may be done according to the known taxonomic structure imposed on attribute values (e.g., home address with respect to zip code). One example is shown in Table 4, where home addresses of Table 1 are merged into larger districts lexicographically.

Table 4: merge of attribute values

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FC	1								1	1										
Al		1				1														1
TC			1										1						1	
An				1				1				1					1			
WA					1						1			1		1				
Re							1								1			1		

Aggregation amounts to the reduction of the complexity of a probability space spanned by attributes [7] and therefore, increases the statistical significance [4]. For the number of attribute values changing from 17 to 6, the threshold of the confidence region is given by a finite number that is 11.1 with the confidence level 0.95 based on chi-square estimation. In the absence of such structure, the concept clustering method with clustering criterion based on $\Pr(\boldsymbol{B_n}|D_m)$ will be used as the selection criterion.

5. ASSOCIATION NETWORK

As discussed, different data analysis methods are used in light of the different statistical properties of attributes. We integrate the similarity relation and its related taxonomy structure [18] with probabilistic causal (Bayesian) to form what we call an association network as in Figure 4. It provides the basis for privacy protection analysis. We envision the following steps for generation.

- Conduct the similarity selection and Bayesian network induction. Attributes with low *V*-score will have their values be either aggregated to increase significance level or replaced with pseudo-code.
- Evaluate impact on target attributes from other attributes in association networks.
- Modify attribute values according to a calculated priority.
- After modification, (randomly) check if other combinations still violate the privacy protection criterion.

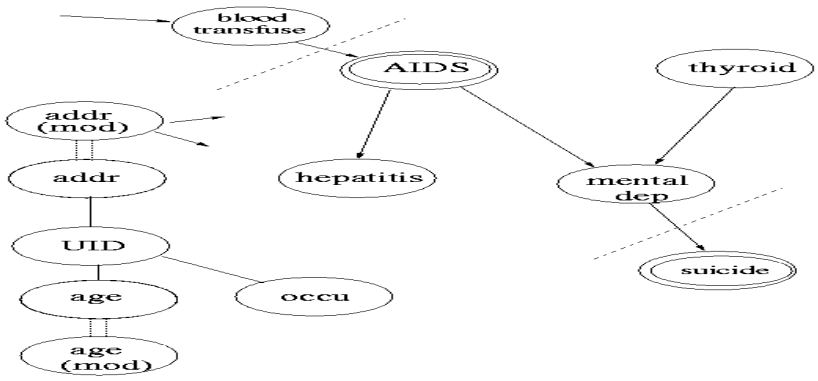

Figure 4: Association network model. The double-dashed line denotes an aggregated attribute. The aggregated attribute may have probabilistic dependency with other attributes. Attributes outside the dashed line are not included in the current database.

5.1 Restoration

It is possible to (partially) restore hidden attribute values if the information of the underlying Bayesian network structures of the database are known – this is the worst case to defend against. As in [2][8][12], the restoration approach primarily selects the set of instantiation **x** to the hidden values with respect to log Pr(*Dm*{**x**}|***Bn***) for *Dm.* With data of Table 3, one could obtain the values of "AIDS" shown in *Table 5*. Note that the two blockings (i.e., data items 3 and 4) are also correctly restored to their original states.

Table 5: restored medical records

hepatitis	n	y	y	y	y	n	y	y	y	n	n	y	n	y	n	n	n	y	n	y
mental	n	d	d	d	d	n	n	d	d	n	d	n	n	d	d	n	d	d	d	n
AIDS	n	y	y	y	n	n	n	y	y	n	y	n	n	y	y	n	n	y	n	n
thyroid	n	l	l	l	n	n	n	l	l	n	l	n	n	l	l	n	l	n	l	n

Changes of the "AIDS" values occur in three places - a reasonably good guess, but a bad outcome for privacy protection. If the number of blockings increases to 4 with "mental depression" of data items 3, 8, 14 and 18 being blocked, the restoration is disrupted. The result is shown in Table 6, where changes in the restored values of "AIDS" increase to

seven, a fairly random outcome. In general, to ensure no restoration, one needs to modify associated causes and evaluate their ramifications [3]. We will consider the combined strategy with respect to the constraint γ.

Table 6: restored from more blocking

hepatitis	n	y	y	y	y	n	y	y	y	n	n	y	n	y	n	n	n	y	n	y
mental	n	d	?	d	d	n	n	?	d	n	d	n	n	?	d	n	d	?	d	n
AIDS	n	y	n	y	n	n	n	n	y	n	y	n	n	n	y	n	n	n	n	n
thyroid	n	l	l	l	n	n	n	l	l	n	l	n	n	l	l	n	l	n	l	n

5.2 Effectiveness Evaluation

The result of blocking will push the target probability toward the uniform distribution. In fact,

Claim 3. The entropy measure of T with $\Pr(T|Dm,\boldsymbol{Bn})$ is monotonically increasing w.r.t. blockings.

This property is in tune with our intuition that uniformity gives maximal entropy, while specificity gives minimal entropy. The evaluation of the effectiveness of modification in our framework is carried out by cross-validation over Dm where effectiveness is measured in terms of the error rate $U_{cf}(e,s)$ [19], meaning the chance of having e errors with s test data at the confidence level cf. For instance, in Table 3, with 3 misclassified test data and 7 test data, the predicted error rate, $U_{cf}(3,7)$, is 0.43 at cf=10%. The result means that if the error rate is high, the network model is unreliable and thus, the inference is mitigated.

6. CONCLUSION

Our results suggest that database privacy protection requires extensive evaluation and analysis of data relationships. Our model requires two-tier processing. First, a similarity analysis is carried out for examining similar attributes. The second tier is based on the probabilistic dependency

analysis of attributes. Blocking and aggregation are used to prevent inference. Inference is analyzed with an association network, which consists of the probabilistic dependency structure, the taxonomy structure and the similarity measure. This provides a unified framework for database inference analysis.

Acknowledgements We thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] Anderson, R. (1998) "http://www.cl.cam.ac.uk/simrja14/caldicott/caldicott.html".
[2] Chang, L. & Moskowitz, I. S. (1998) "Bayesian Methods Applied to the Database Inference Problem," Database Security XII (ed. Jajodia), pp. 237-251, Kluwer.
[3] Chang, L. & Moskowitz, I. S. (2001) "Analysis of Database Inference," in preparation.
[4] Cowan, G. (1998) "Statistical Data Analysis," Clarendon Press.
[5] Duncan, G. (1995) "Restricted Data versus Restricted Access,"In Seminar on New Directions in Statistical Methodology, OMB, pp 43-56.
[6] Denning, D. & Neumann, P. (1985) "Requirements and Model for IDES-A Real-Time Intrusion-Detection Expert System," # 83F83-01-00 CS SRI International.
[7] Freidman, J. (1996) "On Bias, Variance, 0&1 - Loss, and the Curse-of-Dimensionality,"Data Mining and Knowledge Discovery, 1, 55-77.
[8] Heckerman D. (1996) "Bayesian Networks for Knowledge Discovery," Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, pp. 273-305.
[9] Hinke, T., Delugach, H. & Wolf, R. (1997) "Protecting Databases from Inference Attack," Computers & Security, Vol. 16, No. 8, pp 687-708.
[10] HIPAA (1999) The Health Insurance Portability and Accountability Act seminar, NY.
[11] Iceland Database (1999) "http://www.decode.is/ppt/protection/index.htm".
[12] Kong, A., Liu, J. & Wong, W. (1994) "Sequential Imputation and Bayesian Missing Data Problems," Journal of ASA, Vol. 89, No. 425, pp 278-288.
[13] Lewis, P. (2000) book review "Losing Privacy in the Age of the Internet" (author Garfinkle) New York Times, Feb. 10, 2000.
[14] Lin, T.Y., Hinke, T.H., Marks, D.G., & Thuraisingham, B. (1996) "Security and Data Mining," Database Security Vol. 9: Status and Prospects, IFIP.
[15] Marks, D. "Inference in MLS Database Systems," IEEE Trans. KDE, V 8, # 1, pp46-55.
[16] Moskowitz, I. S. & Chang, L. (2000) "A Computational Intelligence Approach to the Database Inference Problem," Advances in Intelligent Systems: Theory and Applications (ed M. Mohammadian) IOS Press, 2000.
[17] Office of Management and Budget (1994) "Report on Statistical Disclosure Limitation Methodology," paper 22.
[18] Pearl, J. (1989) "Probabilistic Reasoning in Intelligent Systems," Morgan Kauffman.
[19] Quinlan, R. (1992) "C4.5", Morgan Kaufmann.
[20] Sweeney, L. (1997) "Maintaining anonymity when sharing medical data," MIT working paper, AIWP-WP344.
[21] Thuraisingham, B. (1998) "Data Mining: Technologies, Tools and Trends:, CRC Press.
[22] Zayatz, L. & Rowland, S. (1999) "Disclosure Limitation for American Factfinder," Census Bureau report (manuscript).