A Framework for Access Control with Inference Constraints

Vasilios Katos*, Dimitrios Vrakas† and Panagiotis Katsaros†
*Department of Electrical and Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece
Email: see http://isir.ee.duth.gr
†Department of Computer Science
Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
Email: {dvrakas,katsaros}@csd.auth.gr

***Abstract*—In this paper we present an approach for investigating the feasibility of reducing inference control to access control, as the latter is a more desirable means of preventing unauthorized access to sensitive data. Access control is preferable over inference control in terms of efficiency, but it fails to offer confidentiality in the presence of inference channels. We argue that during the design phase of a data schema and the definition of user roles, inference channels should be considered. An approach is introduced that can be integrated into a risk assessment exercise to assist in determining the roles and/or attributes that lower the risks associated with information disclosure from inference. The residual risk from the remaining inference channels could be treated by well known inference control mechanisms.**

***Keywords*-access control; inference control;**

I. Introduction

An inference channel is the ability to determine sensitive data from non-sensitive data [9]. Inference control refers to the ability to prevent users from indirectly accessing data that they do not have any authorization for. Unauthorized access in this case is achieved as data themselves usually contain information other than their intended use. The "excess" information contained within a certain piece of data results to the creation of an inference channel, revealing the existence, state and value of other data. This is usually a probabilistic channel which under certain contexts is also referred as side channel and as a result the formal access control mechanisms are bypassed. Detection of such violations in the general case is considered to be hard [6] and therefore research focuses on prevention mechanisms.

Information inference is applicable to both quantitative and qualitative data. In [18] an inference control solution for quantitative data applicable to on-line analytical processing systems is presented. An attacker may derive sensitive information after performing a number of non-sensitive queries on a database and the authors address this problem by investigating the number of inference-safe queries. They also demonstrate how this upper bound depends on characteristics of the underlying datacubes and the structure of the queries. With respect to qualitative data, inference is normally feasible, due to the existence of (statistical) correlation relations between the attributes. For example, the name *Alex* is more likely to refer to a male rather than a female, as the distribution of this name over the two genders is not uniform. In the case of the name *Alexander* the inference channel would not be probabilistic, as the attribute `Name` reveals deterministically the information contained in the attribute `Gender`.

Modern information systems exhibit in many cases an unreserved collection of personal data and security and privacy policies that describe the processes of handling and processing the data are commonplace. However, the large and expanding data schemas, as well as the large volumes of the data result in an increased complexity. Considering Ashby's Law of Requisite Variety [1], we can accept that informal paths between the different type attributes are created, due to the different perturbations of the underlying system, or alternatively, inference channels. Applying Ashby's Law in the context of access control, the privacy of sensitive data cannot be protected by implementing authorization solutions on the sensitive data, when these data are part of a schema that contains unclassified data.

This paper proposes an extended access control model that depends on both access control rules and inference relationships, which are endogenous to the

data. Such model can be considered to complement well known query restriction approaches that are employed during run time. The focus of the research is on the feasibility of combining a deterministic problem (access control) with a probabilistic one (inference control) through a formal model. A metric is introduced that assists a security engineer to make database access control decisions during the design and allocation of user roles. We call upon the well known principles of separation of duties and least to know, as points of reference for developing the approach.

Throughout the paper it is assumed that a methodology for assessing the existence and magnitude of inference channels is available. It is acknowledged that the literature contains a significant amount of ongoing research in information leakage through inference channels, but such an exercise is not within the scope of this paper. However it should be noted that the applicability of our proposal depends on the reliability and accuracy of a method for assessing inference channels.

The paper is structured as follows. In Section II the inference problem is stated and the current research is presented. In Section III the extended access control approach is introduced and metrics are proposed. Section IV presents the conclusions and areas of ongoing research.

II. Access control vs. inference control

Access to sensitive data can be achieved either directly or indirectly. Preventing direct access to the data is dealt with access control mechanisms, whereas preventing indirect access is known as inference control [4]. Although these two prevention mechanisms share the same goal, they differ in several fundamental aspects.

Access control is deterministic, whereas inference control is usually related to stochastic channels. In the general case, an inference channel is ought to be stochastic as there is an associated probability of obtaining the correct value of the inferred data. Access control is static[1], whereas inference control is dynamic [2]. That is, access control rules can be well defined, and once set they apply throughout a user session. Inference on the contrary is influenced by the user's actions and more specifically by the number and structure of queries the user performs and therefore the inference control responses vary through time (see for example the work in [17]).

The differences between access and inference control call for different security mechanisms. In addition, inference control is computationally more expensive than a straightforward implementation of access control. Computational efficiency is an important requirement in a modern system and the performance, as well as accuracy of security controls is a factor taken into consideration when designing secure systems [3], [11].

In an ideal scenario, prevention of unauthorized access to sensitive data would be dealt solely by access control mechanisms. That would be the case if all attributes in the underlying schema are pairwise orthogonal, that is, knowledge of any data would not reveal any information about the values of other data. In large real life databases complete orthogonality is not a realistic assumption. In fact, orthogonality is expected to be sparse between attribute pairs as the number of attributes increases. However, access control being capable of protecting confidentiality would remain a possibility, if user access control roles grant access to attribute sets that are orthogonal (that is unrelated) to all other attributes which the role would not have access to. Traditional methodologies of role based access control design [14] do not consider role separation based on inference, but on other well established security principles such as the least to know principle [13].

In the remainder of the paper an approach based on the concept of *inference control by design* is presented. More specifically, assuming that access control solutions are more preferable than inference control solutions, we present an approach that allows one to explore the possibilities of forming access control rules that will balance trade-offs between business requirements (desirable access control rules) and inference control complexity. Reducing inference control to access control is desirable from a complexity perspective and as it is demonstrated it can be delivered with an expense to the number of roles. This would eventually lead to conflicts with business requirements and therefore the security designer and stakeholder would need to decide on the balance between number

[1]Although it is generally accepted that access control is based on static measures, in [15] the authors develop an extended access control model that includes dynamically updated rules based on user action. However the context in that paper does not consider explicitly inference control.

of user roles and inference control primitives.

III. An extended access control method

Initially we define the access control matrix in the Harrison-Ruzzo-Ullman fashion [8]. In this particular application we adopt the following convention departing from the original HRU terminology: "subjects" are represented by user "roles" and "files" correspond to data "attributes". In addition, it should be noted that we are primarily interested in the confidentiality of the data and as such the access control rules refer to read access operations.

Definition 1: Let $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ be the set of user roles and $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ be the set of different attributes. An access control matrix $\mathcal{A}$ is an $n \times m$ matrix, where for a role $r_i \in \mathcal{R}$ and data $d_j \in \mathcal{D}$ the respective element in $\mathcal{A}$ is defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } r_i \text{ has access to } d_j; \\ 0 & otherwise. \end{cases} \tag{1}$$

In addition, we need to capture the information relating to the existence of inference channels between the attributes. The inference channels exist irrespective of the shape and form of the access control matrix defined above.

Definition 2: Given a set of attributes $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$, a data disclosure matrix $\mathcal{L}$ is a $m \times m$ matrix, where element p_{ij} denotes the probability of attribute d_i revealing d_j.

This essentially means that there exists a mapping $f \subseteq \mathcal{D} \times \mathcal{D}$ where for $(i, j) \in f$, attribute d_j can be inferred from attribute d_i with probability p_{ij}. It is evident from the above definition that $\mathcal{L}$ would be of the following form:

$$\mathcal{L} = \begin{pmatrix} 1 & p_{12} & \cdots & p_{1m} \\ p_{21} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{m1} & \cdots & \cdots & 1 \end{pmatrix}$$

with $0 \leq p_{ij} \leq 1$ where $i \neq j$. From a graph theory perspective, $\mathcal{L}$ can be mapped to an adjacency matrix $A_{adj}[i, j]$ of a weighted directed graph where:

$$A_{adj}[i, j] = \begin{cases} 1 & \text{if } i = j \\ p_{ij} & \text{if } p_{ij} > 0 \\ 0 & \text{if } p_{ij} = 0 \end{cases} \tag{2}$$

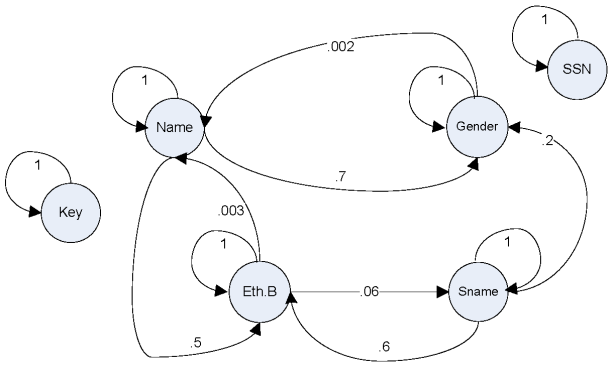

Figure 1. The weighted directed graph corresponding to $\mathcal{L}$.

with the non zero elements for $i \neq j$ representing the inference channels with probability p_{ij}.

Example. Consider for example the data and attributes in Table I and the following access control matrix on that table:

$$\mathcal{A}_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

A reasonable disclosure matrix would be the following:

$$\mathcal{L}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & .7 & .5 & 0 \\ 0 & 0 & 1 & .2 & .6 & 0 \\ 0 & .002 & 0 & 1 & 0 & 0 \\ 0 & .003 & .06 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The corresponding graph consists of three strongly connected components as shown in Fig. 1.

■

Definition 3: Let $\mathcal{L}$ be the disclosure matrix of a data scheme defined by its elements p_{ij}. This data scheme is leakage proof iff $p_{ij} = 0$ for all $i \neq j$.

Essentially a data scheme is leakage proof if knowledge of any of the attributes does not reveal any information about any other attribute, or equivalently, $\mathcal{L} = \mathbf{I}_m$, where $\mathbf{I}_m$ is the $m \times m$ identity matrix.

Definition 4: Let $\mathcal{L}$ and $\mathcal{A}$ be a disclosure and an access control matrix respectively. The access leakage matrix $\mathcal{Q}$ is defined as:

$$\mathcal{Q} = \mathcal{A} \cdot \mathcal{L}$$

Table I
AN EXAMPLE DATABASE TABLE

Key	Name	SName	Gender	Ethnic Background	SSN
101	Alex	Alisson	M	Wh.Eur.	1234567
102	Alex	Warteiner	F	Wh.Eur.	3456789
103	Muhammed	Ali	M	Asian	5678901
104	Mumba	Abonage	M	African	2231122
105	Song	Li	M	Asian	1112223

It can be trivially shown that if there are inference channels (with non zero probability) connecting attributes that can be accessed by different roles, the corresponding element in the access leakage matrix would be greater than zero. The access leakage matrix captures the *effective* access control, rather than the *desirable* access control policy.

From a practical point of view, though it is possible for a data scheme to contain inference channels, the access control policy may prohibit exploitation of them. If for example an inference channel exists between two attributes but the access control prohibits access to both attributes, then there is no inference. Conversely, if an inference channel exists between two attributes but the access control policy grants access to both attributes, then that inference channel would be redundant. This leads to the need to define the property of incidental leakage proofness:

Definition 5: Let $\mathcal{A}$ be an access control matrix and $\mathcal{L}$ the disclosure matrix with $\mathcal{L} \neq \mathbf{I}$. Then the scheme is incidentally leakage proof iff:

$$a_{ij} = \mathbf{1}_{\Re^+}(q_{ij})$$

for all a_{ij} of the access control matrix and q_{ij} the elements of the access leakage matrix $\mathcal{Q}$ where $i = 1..n, j = 1..m$ and $\mathbf{1}_{\Re^+}$ is the indicator function over $\Re^+$.

In an incidental leakage proof schema the inference channels do not influence the confidentiality of the data. Our proposal relates the problem of inference control to the problem of modifying the access control policy or adding auxiliary attributes, in order to obtain incidental leakage proofness.

A. Avoiding inference channels by improved implementation of separation of duties or least to know principles

Following the previous definitions it should be evident that inference control in this context can be expressed as an optimization problem. More specifically, we are interested in increasing the degree of separation of duties reflected in $\mathcal{A}$ or adding auxiliary attributes, which will result to a more sparse $\mathcal{L}$. In both cases the new matrices $\mathcal{A}'$ and $\mathcal{L}'$ will reduce the distance

$$d = \sum_{i=1}^{n}\sum_{j=1}^{m}(q_{ij} - a_{ij})^2$$

where q_{ij} represent the elements of the amended access leakage matrix and a_{ij} the elements of the modified access control matrix. We define four problem types that fall into two categories:

P1a. Given a $n \times m$ access control matrix $\mathcal{A}$ and a disclosure matrix $\mathcal{L}$ such that $a_{ij} \neq \mathbf{1}_{\Re^+}(q_{ij})$ for all $1 \leq i \leq n$, $1 \leq j \leq m$, develop a $(n + k \times m$ access matrix $\mathcal{A}'$ where:
- At least one row of $\mathcal{A}$ with more than one non-zero elements is partitioned into rows; elements of the same column for all resulting rows, when combined with the OR operator yield the value given in the initial matrix $\mathcal{A}$.
- The resulting scheme is *closer* to an incidentally leakage proof scheme, such that the distance d is diminished at an "acceptable" level.
- The number of inserted rows k is limited to the minimum feasible value for achieving the required d

P1b. Given a $n \times m$ access control matrix $\mathcal{A}$ and a disclosure matrix $\mathcal{L}$ such that $a_{ij} \neq \mathbf{1}_{\Re^+}(q_{ij})$ for

all $1 \leq i \leq n$, $1 \leq j \leq m$, develop a $(n + k \times m)$ access matrix $\mathcal{A}'$ where:

- at least one row of $\mathcal{A}$ with more than one non-zero elements is partitioned into rows; elements of the same column for all resulting rows, when combined with the OR operator yield the value given in the initial matrix $\mathcal{A}$.
- The number of inserted rows k is a constant number defined by the underlying business logic. Note that each additional row in the access control matrix represents an additional role in the system.
- The resulting scheme is *closer* to an incidentally leakage proof scheme, such that the distance d is the smallest possible.

These problems refer to row partitioning, which is essentially an implementation of finer grained separation of duties. Not all partitions will be practically feasible, due to restrictions imposed by the underlying business logic. For example, in a relational database model these restrictions depend on the assumed entity relationships.

P2a. Given an access control matrix $\mathcal{A}$ and a disclosure matrix $\mathcal{L}$ such that $a_{ij} \neq \mathbf{1}_{\Re^+}(q_{ij})$, for all $1 \leq i \leq n$, $1 \leq j \leq m$, develop a $(m + k) \times (m + k)$ disclosure matrix $\mathcal{L}'$ where:

- $\mathcal{L}$ is a partition of $\mathcal{L}'$
- the additional elements are all zero, except the diagonal elements which are equal to 1.
- $\mathcal{A}$ is inevitably modified resulting into a $n \times (m + k)$ matrix $\mathcal{A}'$ where the Hamming weight[2] of each row is equal to the Hamming weight of the respective row in $\mathcal{A}$
- the resulting scheme is *closer* to an incidentally leakage proof scheme, such that d is diminished at an "acceptable" level with the smallest possible k.

P2b. Given an access control matrix $\mathcal{A}$ and a disclosure matrix $\mathcal{L}$ such that $a_{ij} \neq \mathbf{1}_{\Re^+}(q_{ij})$, for all $1 \leq i \leq n$, $1 \leq j \leq m$, develop a $(m + k) \times (m + k)$ disclosure matrix $\mathcal{L}'$ where:

- $\mathcal{L}$ is a partition of $\mathcal{L}'$
- the additional elements are all zero, except the diagonal elements which are equal to 1.
- $\mathcal{A}$ is inevitably modified resulting into a $n \times (m + k)$ matrix $\mathcal{A}'$ where the Hamming weight of each row is equal to the Hamming weight of the respective row in $\mathcal{A}$
- The number of additional columns k in $\mathcal{A}'$ is a constant number defined by the underlying business logic. Note that each additional column in the access control matrix represents an additional anonymizer in the system.
- The resulting scheme is *closer* to an incidentally leakage proof scheme, such that the distance d is the smallest possible.

These problems refer to narrowing the disclosed information, thus improving implementation of the least to know security principle. The components of the weighted directed graph corresponding to A_{adj} are increased by k. This is realized by adding vertices that represent auxiliary attributes and at the same time migrating access from the attribute with the greatest sum of weights for the arcs directing from it, to the newly introduced attribute.

All problems aim to increase the sparseness of the corresponding matrices. When modifying the access control matrix the problem concerns how to break a role into two or more sub-roles, where each sub-role will have access to fewer attributes than previously, but when all sub-roles are put together will provide access equivalent to the initial role. In practice, some instances of role fragmentation may not be feasible and therefore different fragmentation alternatives have to be explored.

With respect to the disclosure matrix, no "tampering" can be made on the initial matrix itself, as the information captured refers to the inherent inference channels between the attributes. However, the added auxiliary attributes will not correspond to natural data and therefore they can be free from inference channels. A typical example of such an attribute is the pseudonym which decouples a name from other personal data as will be shown in the following example.

In both cases the method could be used as a policy planning tool to assist the policy makers in exploring role allocation alternatives. Such an exercise may augment risk assessment practices, where the policy makers and system stakeholders consent on an acceptable risk due to inference.

[2]A Hamming weight is defined as the sum of 1's in a binary vector.

B. Estimating an access control policy for reduced inference risk

P1a is a hard problem that concerns finding the minimum number of rows that have to be partitioned in the access control matrix $\mathcal{A}$, in order to drop the distance between the resulting matrix and an incidentally leakage proof construction below a given threshold. P1b is a similar problem with two minor differences: a) the number of rows that can be partitioned is fixed and b) the goal is to find a matrix with the "minimum" distance d, rather than a matrix with an "acceptable" distance.

A straightforward approach to tackle problems P1a, P1b, P2a and P2b is outlined in Algorithm 1, Algorithm 2, Algorithm 3 and Algorithm 4 respectively. Algorithm 1 follows an iterative approach, where at each step the number of additional rows added to the access control matrix is increased by 1, until the size of the new matrix is sufficient in order to find a permutation with the required distance. Note that a completely sparse matrix would be one with no more than one 1 in each row and therefore the number of additional rows that are allowed is limited by the number of 1s in the initial matrix $\mathcal{A}$ minus n. Algorithm 3 is similar to Algorithm 1 with the exception that Algorithm 3 incrementally adds new columns at the end of matrix A and respectively recalculates the corresponding disclosure matrix. Algorithms 2 and 4 do not work in a n iterative manner, since they deal with fixed size access control matrices.

At the core of all algorithms lies a function (`solve_bound`, `solve_bound2`, `solve_optimum` or `solve_optimum2`) which given a matrix with binary values, performs a systematic search in the space of possible permutations of 1s, under specific restrictions in order to come up with a new matrix that either diminishes d at a satisfactory level or minimizes it.

The `solve_bound` function (`solve_bound2` is similar with the exception that it utilizes the additional columns instead of the additional rows) can be easily implemented using a heuristic search algorithm. Algorithm 5 presents a prototype implementation of `solve_bound` that is based on the Best First Search algorithm [16]. The various matrices (permutations of the original access control matrix) are considered to be the states of the algorithm and the distance metric is used as a heuristic function. The algorithm maintains a set of states (matrices) and at each step it removes the state (matrix) with the best heuristic value (minimum distance) and finds all the children states (permutations of the matrix) that are then pushed back in the agenda. The algorithm terminates when it encounters a state with the required heuristic value (distance $\leq D_{max}$) or the Agenda becomes empty which means that the problem is unsolvable. Finally, in order to avoid endless loops the algorithm also maintains a closed set with all the states that have been examined in the past.

	Algorithm 1
Input:	A $n \times m$ access control matrix $\mathcal{A}$, a $m \times m$ disclosure matrix $\mathcal{L}$, a threshold D_{max}
Output:	A $(n+k) \times m$ access control matrix $\mathcal{A}'$
`Let T:=`	`the number of '1's in` $\mathcal{A}$
`For k = 1 to T-n`	
`Create a new (n+k)xm matrix` $\mathcal{A}'$	
`such that`	
$a'_{ij} =$	$\begin{cases} a_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 0 & \text{otherwise} \end{cases}$
$\mathcal{A}'' =$	`solve_bound`$(\mathcal{A}', \mathcal{L}, D_{max})$
`If` $\mathcal{A}'' \neq$	`NULL return` $\mathcal{A}''$
`Next k`	

	Algorithm 2
Input:	A $n \times m$ access control matrix $\mathcal{A}$, a $m \times m$ disclosure matrix $\mathcal{L}$, a threshold k
Output:	A $(n+k) \times m$ access control matrix $\mathcal{A}'$
`Create`	`a new (n+k)xm matrix` $\mathcal{A}'$
`such that`	
$a'_{ij} =$	$\begin{cases} a_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 0 & \text{otherwise} \end{cases}$
$\mathcal{A}'' =$	`solve_optimum`$(\mathcal{A}', \mathcal{L})$
`return`	$\mathcal{A}''$

The `solve_optimum` and `solve_optimum2` functions are optimization functions where there is no explicit criterion concerning the desired goal matrices, other than that they have to minimize the distance d. Therefore they can either be implemented using variations of admissible heuristic search algorithms (e.g. A* [5]), iterative optimization algorithms (e.g. Branch and Bound [10]), or even exhaustive search, where this is possible. However, since the space of possible matrices can become extremely large, it would be much more efficient to search for near optimal

solutions using stochastic optimization methods like local search algorithms [7].

Example. Returning back to the example presented earlier, it can be seen that the corresponding access leakage matrix would be:

$$\mathcal{Q}_1 = \begin{pmatrix} 1 & 1 & 1 & .9 & 1.1 & 1 \\ 1 & .002 & 0 & 1 & 0 & 0 \end{pmatrix}$$

We can verify that the access scheme is not incidentally leakage proof since:

$$\mathbf{1}_{\Re^+}(q_{ij}) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

and the distance from an incidentally leakage proof construction is $d = 2.02$. In other words, knowledge of name and surname is sufficient to gain indirect access to all attributes of the table. If the security policy does not allow such access and the business logic can be delivered with the use of pseudonyms, then an anonymizer attribute can be introduced. The enriched table is shown in Table II.

The updated access control and disclosure matrices would respectively be:

$$\mathcal{A}_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{L}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & .7 & .5 & 0 & 0 \\ 0 & 0 & 1 & .2 & .6 & 0 & 0 \\ 0 & .002 & 0 & 1 & 0 & 0 & 0 \\ 0 & .003 & .06 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and the resulting access leakage matrix would be:

$$\mathcal{Q}_2 = \begin{pmatrix} 1 & 0 & 1 & .2 & .6 & 1 & 1 \\ 1 & .002 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

with the scheme moving closer to an incidentally leakage proof scheme as the distance is now $d = 0.4$. Clearly in this example one anonymizer would not adequately solve the inference problem as there not one but two significant inference channels. Adding a second anonymizer for `SName` and repeating the above process, the resulting access leakage matrix would be:

$$\mathcal{Q}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & .002 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and $d = 0.000004$. This distance could be considered acceptably low and the residual inference channel could be addressed with inference control mechanisms.

■

Algorithm 3

Input: A $n \times m$ access control matrix $\mathcal{A}$, a $m \times m$ disclosure matrix $\mathcal{L}$, a threshold D_{max}

Output: A $n \times (m+k)$ access control matrix $\mathcal{A}'$ and a $(m+k) \times (m+k)$ disclosure matrix $\mathcal{L}'$

Create a new (m+k)x(m+k) matrix $\mathcal{L}'$ such that

$$l'_{ij} = \begin{cases} l_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 1 & m < i \leq (m+k), j = i \\ 0 & \text{otherwise} \end{cases}$$

Let T:= the number of '1's in $\mathcal{A}$

For k = 1 to T-n

Create a new (n+k)xm matrix $\mathcal{A}'$ such that

$$a'_{ij} = \begin{cases} a_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{A}'' =$ solve_bound2$(\mathcal{A}', \mathcal{L}', D_{max})$

If $\mathcal{A}'' \neq$ NULL return $(\mathcal{A}'', \mathcal{L}')$

Next k

Algorithm 4

Input: A $n \times m$ access control matrix $\mathcal{A}$, a $m \times m$ disclosure matrix $\mathcal{L}$, a threshold k

Output: A $n \times (m+k)$ access control matrix $\mathcal{A}'$ and a $(m+k) \times (m+k)$ disclosure matrix $\mathcal{L}'$

Create a new (m+k)x(m+k) matrix $\mathcal{L}'$ such that

$$l'_{ij} = \begin{cases} l_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 1 & m < i \leq (m+k), j = i \\ 0 & \text{otherwise} \end{cases}$$

Create a new nx(m+k) matrix $\mathcal{A}'$ such that

$$a'_{ij} = \begin{cases} a_{ij} & 1 \leq i \leq n, 1 \leq j \leq m \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{A}'' =$ solve_optimum2$(\mathcal{A}', \mathcal{L})$

return $\mathcal{A}''$

Table II
ADDING AN ANONYMIZER ATTRIBUTE (ANON1).

Key	Name	SName	Gender	Ethnic. Background	SSN	Anon1
101	Alex	Alisson	M	Wh.Eur.	1234567	n1
102	Alex	Warteiner	F	Wh.Eur.	3456789	n2
103	Muhammed	Ali	M	Asian	5678901	n3
104	Mumba	Abonage	M	African	2231122	n4
105	Song	Li	M	Asian	1112223	n5

Algorithm 5: Function solve_bound
Input: A $n \times m$ access control matrix $\mathcal{A}$, a $m \times m$ disclosure matrix $\mathcal{L}$, a threshold D_{max}
Output: A $n \times (m+k)$ access control matrix $\mathcal{A}'$

```
Set Agenda    =A
Set Closed    =∅
While Agenda    ≠ ∅
 Set S the matrix in Agenda with
 minimum distance d
 Set d_curr the distance of S
 Agenda = Agenda - {S}
 if d_curr ≤   D_max  return S
 if S∉     Closed
  Set Z=∅
  For each a,b:    S_{a,b} = 1
   For each c:    S_{c,b} = 0
    Create     matrix S_2 such that
              ⎧1     i = c, j = b
    s_{2_ij} = ⎨0     i = a, j = b
              ⎩s_ij   otherwise
    Z = Z  ∪{S_2}
  Set Agenda = Agenda ∪ Z
  Set Closed = Closed ∪ {S}
End while
Return NULL
```

IV. CONCLUSION AND AREAS FOR FUTURE RESEARCH

In this paper, a formal alternative to the inference problem is proposed. By reflecting upon the consensus found in the literature that access control mechanisms are more preferable than inference control mechanisms, we propose a method to deliver inference control by design. This method is concerned with investigating how instances of the (probabilistic) inference control problem can be reduced to (deterministic) access control. We propose to extend the use of matrices and matrix operations beyond access control, in order to represent the inference channels and to capture the effective access control capabilities on a given data scheme.

By acknowledging that not all inference control reductions are practically feasible, we pinpoint the need for a trade-off between meeting business requirements and maintaining low inference risks. We formally capture the trade-offs as two optimization problems. In an ideal scenario - from an inference control point of view - careful access control design could eliminate the need for inference control mechanisms whilst meeting confidentiality requirements. However as in most cases this would not be a realistic goal, we need to accommodate for the constraints that will drive the evaluation process of the different design alternatives. This aspect of the work is part of the ongoing research efforts.

The proposed distance metric d is currently an ordinal variable which meets the representation condition in accordance to a measurement theory. Although an ordinal variable is capable of showing changes (improvements or deteriorations) of the quantity measured, its actual order of magnitude has limited meaning. In order to address this, d would need to be promoted to at least a rational variable. At the time or writing this paper, it is conjectured that the proposed metric could be a rational or an absolute variable, if the leakage matrix in turn can be represented within a Markov model. This is also a research direction under investigation.

Finally, the data disclosure matrix captures pairwise dependencies between the data. Dependencies between k-tuples of data with $k > 2$, are partially accounted for through the multiplication operation between the access control and the data disclosure matrix, but this assumes a linear dependency between the higher order relationships. Consequently, the inference matrix contains the lower bounds of the inferences between the data. Such a limitation can be addressed by considering the proposed data leakage matrix to be the first order

model, whereas the k-tuples would be represented by the $(k-1)^{th}$ order model. Similar to above, a Markovian representation of the leakage matrix would support such direction.

REFERENCES

[1] R. Ashby, *An Introduction to Cybernetics.* London:Chapman & Hall,1956.

[2] J. Biskup,D. Embley, J. Lochner. Reducing inference control to access control for normalized database schemas. *Information Processing Letters* 106 812, 2008.

[3] V. Ciriani, S. Di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, P. Samarati. Fragmentation design for efficient query execution over sensitive distributed databases. *International Conference on Distributed Computing Systems*, 32-39, 2009.

[4] C. Farkas and S. Jajodia. The inference problem: a survey, *ACM SIGKDD Explorations Newsletter*, 4(2), 6-11 2002.

[5] P. E. Hart, P. E.,N. J. Nilsson, B. Raphael, B. Correction to "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". *SIGART Newsletter*, 37, 28-29, 1972.

[6] T. Hinke, H. Delugach, R. Wolf. Protecting databases from inference attacks. *Computers & Security*, 16(8), 687-708 1997.

[7] H. Hoos and T. Stutzle, T. *Stochastic Local Search: Foundations and Applications* Morgan Kaufmann, 2005.

[8] M. Harrison, W. Ruzzo, J. Ullman. Protection in operating systems. *Commun. ACM* 19, 461-471, 1976.

[9] S. Jajodia, and C. Meadows. Inference problems in multilevel secure database management systems. In: M. Abrams, S. Jajodia and H. Podell, Editors, *Information security: an integrated collection of essays*, IEEE Computer Society Press, Los Alamitos, Calif., 570584, 1995.

[10] A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. *Econometrica* 28(3), 497-520, 1960.

[11] E. Magkos, M. Maragoudakis, V. Chrissikopoulos, S. Gritzalis. Accurate and large-scale privacy-preserving data mining using the election paradigm. *Data & Knowledge Engineering*, 68(11), 1224-1236, 2009.

[12] S. Russell and P. Norvig. *Artificial Intelligence: A modern approach*. Prentice-Hall.

[13] R. Sandhu and P. Samarati. Access control: principle and practice. *IEEE Communications Magazine*, 32(9), 40-48, 1994

[14] R. Sandhu., D. Ferraiolo, R. Kuhn. The NIST Model for Role Based Access Control: Towards a Unified Standard. In *Proceedings of 5th ACM Workshop on Role-Based Access Control*, 47-63, 2000.

[15] J. Park and R. Sandhu. The UCONABC Usage Control Model. *ACM Transactions on Information and System Security*. 7(1), 128174, 2004.

[16] Pearl, J. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 48, 1984.

[17] T. Toland, C. Farkas, C. Eastman. The inference problem: Maintaining maximal availability in the presence of database updates. *Computers & Security* 29 88103, 2010.

[18] L. Wang, D. Wijesekera, S. Jajodia. Cardinality-based inference control in data cubes. *J. Comput. Secur.* 12(5), 655-692, 2004.