Novel Micro-Aggregation Techniques for Secure Statistical Databases

By

Ebaa Fayyoumi

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

Ottawa-Carleton Institute for Computer Science
School of Computer Science
Carleton University
Ottawa, Ontario

September 2008

Your file Votre référence
ISBN: 978-0-494-43893-0
Our file Notre référence
ISBN: 978-0-494-43893-0

Dedicated with all my love

to

my loving husband, Ismail,

and to

my wonderful daughter, Zeina.

♡

"*To be sure, this word information in communication theory relates not so much to what you do say, as to what you could say. That is, information is a measure of one's freedom of choice when one selects a message.*"
(C. E. SHANNON and W. WEAVER [142])

Abstract

A Micro-Aggregation Technique (MAT) is a Statistical Disclosure Control scheme that is used to protect a SDB. The aim of this Doctoral Thesis and our research endeavor is to study the Micro-Aggregation Problem and to design and implement novel $MATs$ that could prevent the disclosure of confidential information, and simultaneously not significantly harm the utility of the data being provided to the user. The research undertaken enhances the general performance of $MATs$ by either minimizing the computation, or by minimizing the value of the Information Loss (IL), the Disclosure Risk (DR), or a composite measure of the latter two indices, the Scoring Index (SI).

This Thesis describes four new methodologies, which are our primary contributions:

1. We have considered an existing MAT algorithm, namely the so-called k-Ward algorithm and optimized it for large data sets. This has been done by taking advantage of the distinct properties of the distance matrix and/or utilizing the principle of recursion.

2. We have merged the rich fields of Learning Automata (LA) and $MATs$ to present a novel Fixed-Structure LA to micro-aggregate a micro-data file. The proposed algorithm has been shown to be superior to the state-of-the-art methods.

3. We have suggested how a neural network philosophy can lead to an enhanced MAT. To achieve this we have investigated the effect of replacing the Euclidean distance, which is used to measure the similarity between the individual records in the micro-data file, by the association and the interaction rules that govern the neural network.

4. We have proposed a methodology to use the theory of causal networks and dependency to improve *any* MAT. The results of such a preprocessing phase

assist in solving a very difficult problem, namely that of determining the number and identity of the variables to be used in any micro-aggregation process.

The Thesis also lists various open problems and avenues for future research.

Acknowledgements

This Thesis is the end of my long journey towards obtaining my Ph.D. degree in Computer Science. This could not have happened without the support of many people. The Thesis will not be complete without an acknowledgment of my gratitude.

First, I thank God who helped me through all the tests that came my way during the past five years. You have certainly enriched my life. May your name be exalted, honored, and glorified.

I am deeply indebted to my Supervisor, Professor John Oommen, who traveled with me on this long journey. I could not have wished for a better supervisor. It is difficult to overstate my gratitude to him. He was always there to listen to me, and to offer me sound advice. He taught me that there were different ways to approach a research problem, and instilled in me the persistence needed to accomplish the goal. He has been both a friend and mentor. He taught me how to write academic papers, had confidence in me when I doubted myself, and was able to draw out my research strengths. He was always there to meet me and to discuss my ideas. He was also very through in preparing the corresponding papers and chapters. Without his enthusiasm, his inspiration, and his tireless efforts to explain things clearly and simply, I could have lost my way, and would not have finished this dissertation.

Besides my advisor, I am also grateful to my Thesis Committee for their evaluation and for their constructive comments. I am thankful that in the midst of all their activity, they took time to provide feedback as members of the Committee. I specially thank Dr. Shirley Mills for her friendship and encouragement, and for permitting me to attend her courses on *Data Mining* and *Categorical Data Analysis*. I also thank Dr. Herna Viktor and Dr. Anil Maheshwari for their insightful comments.

I am extremely grateful to Dr. Josep Domingo-Ferrer for all his support and advice, and for providing me with the real-life data sets. I am also thankful to him, Dr. Josep Mateo-Sanz, and Dr. Francesc Sebe for responding to my e-mails and answering my queries. Without their help, I would have been hindered by many

hurdles.

I am indebted to Dr. J. P. Corriveau, the Graduate Director in the School of Computer Science, at Carleton University, for looking after the issues that relate to my financial support.

I am grateful to the staff at the School of Computer Science for helping me in numerous ways. In particular, I mention Linda Pfeiffer, Claire Ryan, Sharmila Namasivayampillai, and Gerardo Reynaga. I would also like to thank the members of Dr. Oommen's previous and present research group, with a special reference to Dragos, Khalid, Denis, and my friend, Cesar.

Where would I be without my family? It is hard for me to find words to express my gratitude to my parents, Naser Eddin Fayyoumi and Heyam Saeidan. They gave me life in the first place, raised me with both care and sincerity, and taught me the fundamental values of life itself. Without your gentle love, endless support, and encouragement, this Thesis would not have materialized. My aim is to attain the goals that you set for me: "Be a useful person and contribute to society". I would like to thank my sisters Enas, Alaa, Eman, Enal, and Atheel for encouraging me to continue my studies, believing in me, and taking care of our parents, while I was immersed in my research.

My literary skills are far too weak to adequately express my appreciation to my husband, Dr. Ismail Al-Fasfous. His dedication, love, and persistent confidence in me has taken the load off my shoulders. He has always encouraged me to accomplish my tasks in a professional and commendable manner. I appreciate his constant patience, understanding, endless support, and encouragement in every situation. I am honored to know him and humbled to share life with him. His love and kindness are foundation stones for all that I am and currently have. Now that my Thesis is completed, I am looking forward to spending more time with him.

Finally, there is one extraordinary person who deserves a very sincere acknowledgment, which is my daughter, Zeina Al-Fasfous. I owe her so much! She has been my

source of strength and love. I could not have survived the pressures of life without her. I am glad that she is growing up, and I hope that she will be proud of me when she understands what these last years have entailed. Thank you, pretty girl, for your love, and hugs, and for motivating me to keep reaching for excellence. Thank you for everything that you are, and for everything that you will become.

The financial support for my Doctoral studies was partially provided by The Hashemite University in Zarqa, Jordan. This assistance was crucial to the success of my research, and, I gratefully acknowledge it. I also thank the administration of the Hashemite University for releasing me from my academic responsibilities while I undertook this major endeavor!

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

ACA	:	Adaptive Clustering Algorithm
$ACNN$	:	Associative Cluster Neural Network
AI	:	Artificial Intelligent
ASR	:	Association Similarity Rule
DR	:	Disclosure Risk
$EMAD$	:	Enhancing Micro-Aggregation with Dependence
$EMIM$	:	Expected Mutual Information Measure
EPP	:	Equi-Partitioning Problem
$FSSA$	:	Fixed Structure Stochastic Automata
G_{DR}	:	The overall Global DR
G_{IL}	:	The overall Global IL
$IAMAT$	:	Interactive-Associative Micro-Aggregation Technique
ID	:	Interval Disclosure
IL	:	Information Loss
kW	:	k-Ward MAT
kWD	:	k-Ward Diagonal
$kWDR$	:	Recursive k-Ward Diagonal
kWR	:	Recursive k-Ward
$kWR-D$	:	Recursive k-Ward Dynamic threshold
$kWR-F$	:	Recursive k-Ward Fixed threshold
LA	:	Learning Automata

MAP	:	Micro-Aggregation Problem
MAT	:	Micro-Aggregation Technique
MD	:	Maximum-Distance
$MDAV$	:	Maximum Distance Average Vector
ML	:	Maximum Likelihood
MST	:	Maximum Spanning Tree
NN	:	Neural Network
OMA	:	Object Migrating Automaton
$OMMA$	:	Object Migrating Micro-aggregated Automaton
PCA	:	Principal Component Analysis
$PRAM$	:	Post-RAndomization Method
RLD	:	Record Linkage Disclosure
SDB	:	Statistical DataBases
SDC	:	Statistical Disclosure Control
SI	:	Score Index
SSA	:	Sum of Squares Among the groups
SSE	:	Sum of Squares Error
SST	:	Total Sum of Squares error
$V-MDAV$	:	Variable-size Maximum Distance to Average Vector
$VSSA$	:	Variable Structure Stochastic Automata
ZS	:	Z-Scores

Chapter 1

Introduction

1.1 Introduction

The development of empirical research and the growing capacity of modern computer systems have led to Statistical DataBases ($SDBs$) that contain a large amount of information, some of which may be sensitive. $SDBs$ are used to produce statistical summaries about certain groups, such as the sum, count, average, min and max [30]. However, it is essential that these statistical summaries do not disclose the content of any single individual record. This requirement is often difficult to satisfy, because the user may ask many different *"legal"* queries and deduce confidential information from the responses obtained[1]. For example, comparing the mean salary of two groups differing only by a single record will compromise the salary value of the individual whose record is in one group but not in the other [2, 4, 12, 30, 69, 80, 86, 123].

The problem of enhancing the security in $SDBs$ has received a lot of attention in

[1] $SDBs$ can be exactly or partially compromised if there is an authorized user who constructs one or more queries using *a priori* information, which, in turn, can be used to infer the value of a certain field. The most effective methods for compromising $SDBs$ are listed as follows: Generating a query set whose cardinality is unity, using an individual tracker, utilizing a general tracker, or presenting queries that possess the "overlap" property. More details of these mechanisms are given in [12, 30, 69].

recent years. This is because of the fact that government agencies and private organizations increasingly face the problem of protecting confidential information contained in their databases. Census bureaus have successfully dealt with this problem by removing information from the databases that can easily identify an individual, *i.e.* Social Insurance Number. They release statistical summaries drawn from a small sample of the entire population [35, 101, 97].

Generally speaking, the problem at hand concentrates on guaranteeing the security of $SDBs$, and on obtaining useful, unbiased statistical information. Several avenues to achieve this have been reported in the literature, each of them having their own properties.

One approach to solve this problem is to restrict access to the data. INGRES [69], for example, distinguishes between two types of statistical queries: Those that apply to the whole relation and those that apply to a subset of any relation. The first type of query is equivalent to one with a non-characteristic formula, which will be answered normally. As opposed to this, the second type of query is modified with relevant access rules in such a way that a potential intruder only gets *Read Access* to a subset of rows in the relation and, thus, obtains an answer which approximates the true one. This, of-course, prevents compromising the database, but it is very restrictive for most situations.

Another approach to resolve this is to provide the system with a wide range of powerful techniques, each of which yields inference protection against intruders. The families of these techniques are listed and classified in the literature [16, 86], and a more detailed survey is found in Chapter 2. However, it is generally assumed that they can be classified into three groups: Query set restriction, various perturbation-based approaches, and conceptual schema approaches. These are briefly detailed in the following paragraphs.

1. **Query Set Restriction Approaches**

   These techniques protect against inference by restricting and not releasing the

statistical summaries that could reveal pieces of confidential information about single individuals in the *SDB*. The set of released data is a subset of the set which satisfies the restrictions that have been applied on the size, type of the query result and/or the percentage of the overlap between the result of the current query and the previously submitted queries. Although this approach prevents an exact compromise, it eliminates the usefulness of *SDBs*. *Query set size control*, *Query set overlap control Auditing*, *Partitioning* and *Cell suppression* are examples of methods within this category [16, 69].

2. **Perturbation-Based Approaches**

   These techniques provide inference protection by introducing a modification to the information in an unpredictable way. The information can be modified before answering the queries or after processing them, but rather, before releasing the result to the user. Data perturbation is the name of methods in the first scenario where a proxy database is created from the original database and all the statistical queries are calculated against the proxy database. Methods within the second sub-class are called output perturbation methods, and they do not need a proxy database because all the statistical queries are calculated against the real database, after which noise is added to the result. Data perturbation avoids the inference and always provides consistent results. On the contrary, output perturbation could provide the user with inconsistent results. *Data swapping*, *Fixed data perturbation*, *Micro-aggregation*, *Random sample queries*, *Varying output perturbing*, and *Rounding* are examples of methods within this category [16, 69].

3. **Conceptual Approaches**

   Methods of this sub-class address the inference problem by altering the conceptual level schema on which the *SDB* is built. It concentrates on re-designing the schema of the relations. Although the benefit of such solutions are significant with respect to security, the implementation is very expensive because these approaches require major changes to the structure of the database model, as well as the additional time required for uploading sessions to populate the

tables. *The Conceptual Model* and *Lattice Model* techniques are examples of methods within this category [16, 69].

The quality of the information released to the user in the presence of an inference protection technique is measured using different criteria such as security, robustness, bias, precision, and consistency [2, 16]. Security is the fundamental measurement of performance that avoids both partial and exact compromises. Robustness concerns the assumptions made by the system about the supplementary knowledge of the $SDBs$. As opposed to these, the difference between bias and precision is that the bias represents the variance between the perturbed value returned to the user and the true value, while precision refers to the variance in a system's bias, which must be within a specified confidence interval. Finally, consistency represents the ability of a perturbation-based system to respond to the queries without contradiction[2] or paradox[3].

As previously mentioned, many researchers have proposed an ensemble of methods for preventing a breach of the security in $SDBs$, while others have constructed models for studying this problem. The research within the scope of this Doctoral Thesis deals with the Micro-Aggregation Technique (MAT) which is considered as one of the perturbative methods. In spite of a few reported weakness described later in the Thesis, $MATs$ have recently emerged as being some of the most promising protection methods [100]. It is used to protect micro-data files by storing the individual records in groups possessing a minimum size constraint. Whenever the query is submitted, it is addressed to a *group* containing the record, but never to a specific record within the group. This prevents a user from isolating a record with overlapping queries. An MAT possesses many attractive features such as its robust performance, its consistent responses, and ease of implementation. It is important to mention that the MAT offers more protection to outliers than other SDC techniques, and the choice of its unique parameter (which we shall specify later) is less influential than others. The

[2]A contradiction can occur when the query responses are perturbed. In such a case, there is a possibility that comparing responses of two queries will reveal arithmetic inconsistencies.

[3]A paradox refers to responses that violate the properties of statistical queries.

clear connection between the properties of an MAT and the anonymization of $SDBs$ increases the importance of the former [40, 53, 57, 94, 131, 133, 134].

The primary goal of this Thesis is to report a set of newly proposed strategies that enhance the general performance of $MATs$. These enhancements are obtained by either minimizing the required computational time to micro-aggregate all the micro-units in the micro-data file, or by minimizing the value of the Information Loss, (IL), the Disclosure Risk, (DR), or a criterion which is a composite index based on the IL and DR. Our research will involve concentrating on all these aspects from a variety of perspectives.

This chapter provides an introduction to this Doctoral Thesis, and illustrates its main motivations and objectives. The chapter is organized as follows: Section 1.2 presents the main motivations and objectives of the work. Section 1.3 gives the organization of the Thesis, and catalogues the main contributions of each chapter.

1.2 Motivations and Objectives

As mentioned earlier, this Thesis focuses on $MATs$ that are a family of techniques that modify the micro-data file in order to guarantee the factual anonymity of the data. At the same time, the intention is that there should not be a huge reduction in the information content of the data so that the user is able to get useful, un-biased statistical summaries. Micro-aggregation is a clustering problem with cardinality constraints that originated in the area of Statistical Disclosure Control (SDC) for micro-data. In other words, an MAT is typically implemented by clustering the micro-individual records into groups (where each group satisfies certain group-size constraints) based on the similarity between them, and then replaces the individual values by the aggregated value.

There are numerous micro-aggregation algorithms; they can be classified as follows: uni-variate *vs.* multi-variate, heuristic *vs.* optimal, and fixed-size *vs.* data-oriented. A detailed description and survey of these methods can be found in Chapter 2. In this Thesis, our aim will be to enhance existing algorithms that belong to the heuristic, uni-variate, data-oriented type, and to develop new algorithms that belong to multi-variate, heuristic, with fixed-size or data-oriented type. The reason for seeking such solutions will be explained presently, and in more detail in Chapter 2. In this regard we mention that although an optimal algorithm for the uni-variate case was proposed in [74], unfortunately, for the multi-variate case, the problem is NP-hard [102, 103].

The main advantage of having several heuristic algorithms with fixed or data-oriented group size constraints is that there are some situations in which the micro-aggregation of key attributes is insufficient to preserve confidentiality. Therefore, re-computing the micro-aggregated sets using other heuristic algorithms or changing a security parameter (such as the group size constraint), could also lead to potentially superior solutions [42].

1.2.1 Motivations

A significant amount of work has been done to devise and optimize $MATs$, most of which involve invoking clustering as a foundational tool. The methods available for clustering data are numerous and a list would probably number in the hundreds. Rather than re-visit the problem by merely considering new clustering methods, our aim is to see if we can optimize $MATs$ by utilizing some *fundamentally new* tools. In order to motivate the Thesis, we list these below:

- We would like to see if we can incorporate a recursive strategy in developing $MATs$. To the best of our knowledge, we are not aware of $MATs$ that have been developed for large data sets, and that are entirely constructed by first invoking $MATs$ on small-size data sets whose results are subsequently merged

to yield the final comprehensive result.

- A second motivation for this Thesis is to see if we can devise a set of algorithms for the Micro-Aggregation problem (MAP) that use Learning Automata, (LA). As far as we know, there is no existing MAT for which the basic computational tool is a fixed or variable structure LA. Indeed, our aim is to find a strategy by which we can merge the rich fields of MAT and LA.

- The third motivation focuses on devising multi-variate $MATs$ that are based on a neural-based clustering method. Indeed, we are not aware of any Neural Network (NN) technique that has been *explicitly* used as a fundamental tool to solve the MAP. We would like to investigate whether the concept of homogeneity (similarity) between micro-records can be measured by incorporating the phenomena of Association and the Interaction. Again, to the best of our knowledge, no MAT has been described so as to utilize these criteria.

- The last motivation of this Thesis is to investigate whether it is possible to introduce into the existing MAT solutions, schemes that determine the *dependency* between the random variables (rather than the records). As far as we know, dependence information has not been used in devising $MATs$, and it is our hope that our research will lead to methods that utilize such strategies.

This summarizes the motivation for our Thesis.

1.2.2 Objectives

With the motivations previously listed serving as a beacon to develop new $MATs$, the objectives of the Thesis are the following:

1. First of all, we would like to develop fast recursive mechanisms that invoke $MATs$ for large-size data sets. In particular, we would like to optimize the k-Ward micro-aggregation algorithm with respect to the computational time for

both uni-variate and projected multi-variate micro-data sets. Our intention is to increase the speed of the micro-aggregation without sacrificing the ultimate utility of the data. We would like to demonstrate that it is possible to take advantage of the distinct properties of the distance matrix and the principle of recursion, to substantially reduce the effective computational burden of the scheme.

2. We would like to demonstrate how the principles behind the families of fixed-structure stochastic LA can be used to enhance $MATs$. Integrating these families would lead to novel contributions in both these domains. The results we present prove that this can be achieved for uni-variate and multi-variate micro-data sets for fixed-size constraints.

3. We would like to use NN techniques, such as those previously used for associative clustering, to enhance $MATs$. In particular, our objective is to demonstrate the applicability of the association similarity rule (instead of the Euclidean distance) to yield micro-aggregation. In this regard, we have utilized the *Associative Clustering Neural Network* learning scheme to effectively solve the MAP.

4. We would like to demonstrate that while developing $MATs$, it is expedient to incorporate information about the structure and dependency between the random variables in the micro-data file. We have demonstrated that including such information will enhance the process of determining how many variables and which variables should be used in the micro-aggregation process. The question of knowing the most suitable metric (with respect to the required computational time and the accuracy) for estimating the dependence model has also been addressed.

1.3 Organization of the Thesis

In this section, we present the overall organization of the Doctoral Thesis and the content/contributions of each chapter.

- **Chapter 2: Literature Review**
  This chapter presents the background material needed for the Thesis. We first introduce the concept of SDC, and then describe its application domain and the various data types that are currently used in $SDBs$. We then proceed to focus our scope by concentrating on the family of micro-data types in $SDBs$. At this point, we emphasize the importance of the IL and DR measures, and survey the various ways of resolving the conflicting goals that they represent. Additionally, we also summarize the perturbative and non-perturbative SDC methods for micro-data protection. Since our research focuses on the MAT, we formally state the MAP and attempt to survey it in a comprehensive manner. Indeed we believe that this chapter represents a complete overview of the state-of-the-art techniques.

- **Chapter 3: Enhancing the Heuristic k-Ward MAT**
  This chapter shows how we can optimize the k-Ward method by applying some matrix-based and recursive optimization techniques, thus rendering the k-Ward method to be pragmatic. These optimizations involve (a) minimizing the computations done in evaluating the between-class *distance* matrix, and (b) recursively partitioning the data set before invoking a k-Ward strategy, thus enforcing that the latter is invoked on "primitive" small-size sub-groups that terminate the recursion. The proposed optimized versions have been rigorously tested on two reference data sets, and the results have been compared with the standard benchmark methods [45]. The main contribution of this chapter is that we have clearly demonstrated that our new optimized versions minimize the computational time required to micro-aggregate *all* the individual micro-units in the micro-data file, rendering it a viable option. The computational advantage sometimes exceeds 80% if one of the optimization approaches is applied by itself,

and more than 90% if both the above enhancements are invoked simultaneously. The various optimization strategies are applied to uni-variate data sets, and also on multi-variate data sets projected using the first principle component or the sum of Z-scores. The work done in this chapter was published in [65].

- **Chapter 4: Using Fixed Structure LA to Solve the MAP**
  This chapter presents an interesting new method for a fixed-size, multi-variate, micro-aggregation in the Euclidean space. As mentioned earlier, the very concept of using LA for micro-aggregation is a novel contribution to the SDC field. The method is based on the concept of specifying various levels of uncertainties for pairs of individuals belonging to the same group. To accomplish this, our new LA, the Object Migrating Micro-aggregation Automaton ($OMMA$), increases or decreases these uncertainty levels depending on whether a known similar pair is "correctly" or "incorrectly" clustered by the algorithm. The experimental results presented in this chapter show that this technique is a promising scheme for the MAP, since the improvement in the data utility is as high as 10% and 8% when compared to the Maximum Distance Average Vector ($MDAV$) and Maximum Spanning Tree (MST) schemes, respectively. Besides this, our solution also offers a minimum score value when compared to the $MDAV$ method by using a score index (SI), which is a composite measure incorporating both the IL and the DR. The work done in this chapter was published in [64].

- **Chapter 5: Using NN Methods to Solve the MAP**
  This chapter presents a new practical method, based on the theory of NNs, to solve the MAP. The novelty of this chapter is its concentration on introducing a new metric for measuring the similarity. This new metric preserves the meaning of the closeness between the records so as to coalesce them into groups, and to also maximize the actual Interaction among the records inside each group. In general, incorporating the concepts of the Interaction among the records and their mutual Association, has been shown to be advantageous to solve the MAP because they have the effect of minimizing the IL. The experimental results presented in this chapter show that this technique is a promising scheme for the

MAP, since the improvement in the data utility compared to the $MDAV$ is as high as 8% and 14% for real and simulated data sets, respectively. Again, the method also yields a superior SI value when compared to the $MDAV$. The work done in this chapter was published in [107].

- **Chapter 6: Utilizing Dependence-Based Information for MATs**
  This chapter shows how the presence and structure of dependency between a set of random variables in the micro-data file is valuable information that can be used as a fundamental indicator before invoking any MAT. We present a new automated scheme as a pre-processing stage to determine the number and the identity of the variables that are to be used in the micro-aggregation process. This is achieved by constructing a completely connected undirected graph whose nodes represent the random variables in the micro-data file, whose edges represent the statistically dependencies, and whose edge weights are computed by using either an information theoretic, χ^2, or correlation-based measured. The advantages and disadvantages of these respective measures have been investigated. The experimental results presented in this chapter show that such a methodology involving projecting the multi-variate data sets reduces the solution space, which further directly reduces the computation time required to search the entire space combinatorially. Additionally, this methodology leads to a solution whose IL values are close to the minimum value of the IL obtained by *exhaustively* searching over the entire search space. The work done in this chapter was published in [106, 108].

- **Chapter 7: Conclusion and Future Research**
  This chapter summarizes the work done in the Thesis, gives the final conclusions, and presents suggestions for future research that can be pursued in the area of SDC, and in particular, to the MAP.

Chapter 2

Literature Review

2.1 Introduction

A lot of attention has recently been dedicated to the problem of maintaining the confidentiality of statistical databases through the application of statistical tools, so as to limit the identification of information on individuals and enterprises and simultaneously maximize the data utility. *Statistical Disclosure Control/Limitation*, (*SDC*), is a field in statistics and computer science (and in particular, in data mining) that has attracted much attention in recent years. Decision-makers are increasingly demanding more detailed statistical information. Indeed, investigators at universities and research centers have the capacity to perform complex statistical analysis with information obtained from various sources of micro-data. Clearly, the requirement that statistical offices publish more detailed information is growing.

There is, however, another side to this extensive use of data. Inasmuch as more confidential data is at risk, statistical offices are required by law or by established policies to protect the confidentiality of information provided to them by the respondents. This confidentiality is also vital when it concerns guaranteeing the future co-operation of respondents.

Generally speaking, two broad approaches are used to preserve the confidentiality of statistical information [2, 12, 36, 37, 69, 30, 142]:

- Access Control. Disclosure can be controlled by restricting the access to data in different ways. In this case, access may be granted only to well−defined group members, subject to well−defined conditions and/or in well−defined secure locations.
- *SDC* Techniques. The released data may be modified to reduce the risk of disclosure.

It is important to mention here that these two approaches are complementary, and are often utilized simultaneously. An agency might choose to release the data either with little modification and under very strict access arrangements, or with considerable modification and subject to much looser arrangements.

The main responsibility of statistical offices is to produce and disseminate pieces of statistical information in a form that is suitable to the policy makers, researchers and the general public. The information released using an *SDC* scheme may have some degree of data modification, where the modifications fall within a range of two extreme models [20, 53]:

- Data Encryption. If the released data is encrypted, it can be protected very securely. However, if the user is not able to decrypt it, the released data is, obviously, useless [37, 69, 81, 89, 30].
- No Modification. If the original data is released without modification, it is clear that its usefulness/accuracy is maximal. But in this case, the confidentiality is not protected against any type of disclosure [2, 37, 69, 89].

The important question that thus arises is one of knowing how the information available can be perturbed or modified in such a way that the released data can be

statistically useful, and simultaneously not risk the privacy of the entities involved. Thus, the main goal of SDC is to modify the original data so that it can be "published" without compromising the inherent confidentiality contained in the information. The challenge for an SDC scheme is to achieve this modification with a minimum loss of the details and accuracy sought by the user of the database. Generally speaking, every SDC method has two conflicting goals [2, 102, 141, 142]:

1. Minimize Disclosure Risk (DR): This index is the risk to the confidentiality of the respondents that the data releaser (typically a statistical agency) would experience as a consequence of releasing the data.

2. Minimize Information Loss (IL): This index quantifies the *value* of the released data to legitimate data users.

Clearly, the ultimate goal of SDC techniques lies not only in reducing DR, the disclosure risk, but also in increasing the utility of the data to the user. Thus, although deleting (or hiding) all the data and releasing no information would eliminate any statistical disclosure, clearly, such a solution is inadequate, because the analytic value of the data will be absent, and vice versa. The conflict between these two criteria is obvious, and our goal is to optimize them, so as to reach a happy medium or an equilibrium point.

A formal optimization specification for the SDC can be stated in terms of Problem SDC as follows [81, 142].

Problem SDC: An optimal SDC solution aims to:

$$\text{Minimize IL}[\int(\text{Data})], \text{ subject to DR}[\int(\text{Data})] < \epsilon,$$

where:

Data	:	The original data.
$\int$(Data)	:	The released data after the use of the SDC techniques represented by $\int$.

IL[$\int$(Data)] : Information Loss in the released data.
DR[$\int$(Data)] : Disclosure Risk arising from the release of $\int$(Data).
ϵ : A threshold value representing the acceptable level of DR. □

It should be emphasized that, in many applications, it is very difficult (and sometimes impossible) to obtain suitable overall measures for DR and IL that are useful to achieve a formal optimization. This arises not only from the complexity of the problem itself, but also because of the difficulty in making assumptions about the nature of the disclosure threats and the user's information requirements. Additionally, it is clear that the decision about what data can be released depends also on the *subjective* experience and personal judgments related to the *nature* of the data [142].

2.1.1 Application Domain

There are several areas of application for SDC techniques. These include (but are not limited to) the following [5, 20]:

- Official Statistics
  Most countries are legislated by law to compel statistical agencies to guarantee the confidentiality of statistical information whenever they release data collected from citizens or companies. This has been the rationale for enhanced research in SDC in many countries, including those from the European Union (i.e., the $CASC$ project) and the USA.

- Health Information
  This is one of the most sensitive and important areas demanding privacy. For example, in the USA and in most western countries, privacy restrictions of health-related organizations require the strict and direct regulation of protected health information for use in medical research.

- E-commerce

  Electronic commerce transactions automatically lead to the automated collection of large amounts of consumer data. This information, which can give a competitive edge, is very useful to companies who receive it. Again, it is imperative that such information transfer should not result in the public profiling of individuals. It is, thus, subject to regulation.

2.2 Types of Statistical Data

The state-of-the-art in *SDC* has evolved to such a level that, at the present, it has spawned studies in involving at least three sub-disciplines[1]:

- Tabular Data

  Historically, tabular data are the most common types of data sorted in official databases. Tabular data can be obtained from micro-data by a process called static "*aggregation*" [28, 102, 142]. The two formats commonly used to display aggregated data [5, 20, 102] are:

  - *Frequency tables.* These count the number of respondents with specific characteristics.
  - *Magnitude tables.* These are analogous to frequency tables in that they are defined by cross-classification of categorical variables. However, the cells contain aggregate values, of some quantity of interest, for the corresponding respondents.

  Tabular data protection is the oldest and most established field within *SDC*. The goal here is to publish tables in such a way that no confidential information on specific individuals (whose information is present in the tables) can be

[1]Distinctions are often made between the ways in which data is released and the "technical" details of the data types. Thus, the proposed methodological approaches differ based on the type of the data involved [5, 89].

inferred. In other words, the tables have to be protected against the exact disclosure of an individual attribute[2], especially when the cells of the table contain information on only one or a very few respondents. A relevant software package in this case is τ-ARGUS [78], and more specific information on the latter can be found in [101].

- Dynamic Databases

  In the case of dynamic databases, the user submits statistical queries (*i.e.*, queries about the average, median) to the database. The most challenging issue for dynamic statistical databases is that the aggregated information obtained by an attacker as a result of successive continuous queries should not permit him to infer any confidential information on specific individuals [69, 102, 30]. One possible strategy is to perturb the answers to the queries [35, 80]. However, if the perturbed results are not acceptable and exact answers are required, the system refuses to respond to these queries. A second solution, based on query restriction, can be found in [2]. Finally, a third strategy is to provide the user with an accurate interval answer [5, 20].

- Micro-data

  Data stored and provided in tables do not provide a particularly wide scope for further statistical analysis and investigation. Formally, it is only possible to perform analysis for summaries containing the pertinent information. Consequently, there is considerable demand from researchers for the release of individual data records. It is, thus, only a more recent phenomenon that data collectors have been persuaded to publish micro-data. Therefore, the protection of micro-data is the "youngest" subdiscipline, and this field is experiencing a continuous evolution [5, 20, 28, 102].

  Micro-data consist of a series of records, each containing information on an individual unit such as a person, a firm, or an institution. Micro-data in their

[2]Three common strategies used to protect against the exact disclosure for tabular data are the so-called "table redesign" scheme [141], the scheme which involves the suppression of individual values [141], and the computationally efficient method involving resampling [38, 142].

simplest form may be represented as a single data matrix, where a row corresponds to the units and the columns to the variables [142].

When it concerns the protection of micro-data files, one has to take into consideration the fact that the information is typically obtained from surveys involving social and business data [67, 72]. In both cases, micro-data are released based on the appropriate rules of research, which is, in turn, based on the re-identification risk for the particular micro-data set. In summary, these rules determine how much the information content of a micro-data file has to be reduced or modified so as to be considered safe [101]. For social surveys, these kinds of modifications belong to the class known as "non-perturbative" *DSC* measures; for the business micro-data they belong to "perturbative" *DSC* measures[3] [72].

Generally speaking, there is a dependency between the variables or records. Such dependencies can be of a logical nature or a statistical nature. In this research, we will not take into account such dependency issues, even though sometimes ignoring such dependencies may lead to an underestimation of the disclosure risk. For example, in a record with "Disease = Cancer of the Womb" and "Sex = missing ", it is clear that the attribute for gender satisfies "Sex = Female".

The software package μ-ARGUS can be used to protect micro-data by offering several approaches, as given in [77]. This research document, in subsequent section, will cover the current state-of the art when it concerns the use of *SDC* methods for micro-data.

2.3 Variables Types

As mentioned in Section 2.2, this study focuses on processing micro-data. A micro-data set can be viewed as a file with n records or data vectors, each containing

[3]Applying non-perturbative techniques on business data tends to yield either an unacceptably high disclosure risk associated with the released data, or a large reduction in the information that renders the analysis meaningless or impossible [72].

data about individuals or respondents, who, in turn, can be persons, companies, etc. Each individual is assigned a data vector formed by a number of variables/attributes [45, 60, 142].

A variable is an object that is associated with a set of values called a domain. Sometimes, a variable in a data set corresponds to the response to a question in a survey. Some variables are, in turn, derived from other variables through a computational mechanism. Variables play an important role in *SDC* analysis and in the kind of disclosure control techniques that they admit [142]. Variables in the original unprotected micro-data file can be classified in six categories (which are not necessarily disjoint) [45, 60, 142]:

- Categorical Variable
  A categorical variable is a variable which can take values over a finite set, and for which standard arithmetic operations are undefined. Two main types of categorical attributes can be distinguished [60]:
  - *Ordinal*
    An ordinal variable takes values in an *ordered* range of categories. Thus, global relational operators, such as *max* and *min* operations are meaningful, and can be used by *SDC* techniques. `Weekdays`, `Months`, etc, are examples of ordinal variables.
  - *Nominal*
    A nominal variable takes values in an *unordered* range of categories. The only meaningful operator in this case is the comparison of *Equality*. The variables `Eye-Color` and `Sex` are examples of nominal attributes.

  In some cases, an additional structure on the domain of the categorical variable can be imposed. Categories sometimes have a hierarchical structure, thus leading to so-called hierarchical variables. Such a hierarchical variable has several partitions defined in its domain. Usually, the least detailed partition is at the highest level, and the most detailed one is at the lowest level. All the remaining

partitions are at levels between these two extremes [142].

Generally speaking, categorical variables play an important role in *identifying* the corresponding respondent in the micro-data files [142].

- Continuous Variable

  A variable is considered continuous if it is numerical (and real-valued) and if it permits arithmetic operations. When we design methods to protect continuous data, it is imperative that we optimize the use of these arithmetic operations. But, simultaneously, we have to be aware of every combination of numerical values in the original data set to preserve its uniqueness, *i.e.*, if it, indeed, is to be unique. `Salary` and `Age` are examples of continuous variables [60].

- Identifying or Key Variable

  An identifying or key variable is an attribute that unambiguously identifies the respondent. Examples of such variables are the `Social_Security_Number`, `Passport_Number`, `Employee_Number`, etc. Clearly, all these key variables should be excluded from the micro-data file before being released to prevent confidential information from being linked to specific respondents. Generally speaking, variable of this type can lead to a *direct* disclosure [45, 60, 142].

- Quasi-identifying Variables

  Quasi-identifiers constitute a set of variables in the micro-data file that, in combination, can be linked with any external information. These permit the attacker to re-identify the respondent to whom the records in the micro-data file refer to [142]. Variables of this kind could lead to *indirect* disclosures [45]. The problem here is that it is not possible to eliminate quasi-identifiers because any variable in the micro-data file can potentially be a quasi-identifier. Indeed, there is no "God-given" procedure which can assist in deciding whether a variable belongs in this class or not.

- Sensitive Variable

  A sensitive variable is an attribute that is found in the released data, but which is, typically, unknown to an intruder before the data is released. This attribute

contains non-confidential information that could enable the accurate estimation of confidential information. Examples of such sensitive variables are `Religion`, `Health_Condition`, etc [20, 60, 89, 142].

- Safe Variable
  Attributes of this class contain non-confidential information pertaining to the respondent. Additionally, such a variable does not add any *"critical"* information to the confidential data. Examples of such a variable could be `Town`, `Country`, `Nationality`, etc. In protecting a data set, we should also consider the fact that an attribute of this kind could, without our knowledge, also be a quasi-identifier [20, 60, 89, 142].

2.4 Micro-data Performance Measures

As mentioned in Section 2.1, the purpose of *SDC* can be given quite informally by stating that, given the original micro-data set, V, the goal is to release a protected micro-data set, V', in such a way that [102, 60]:

- Disclosure Risk, (*i.e.,* the risk that a user or an intruder can use V' to determine confidential sensitive attributes on a specific individual among those in V), is low.
- User Analysis on V' and on V yield the same or at least similar results.

2.4.1 Micro-data Disclosure Risk

Micro-data *SDC* attempts to protect the privacy of individual data in the released data set V'. The issue of disclosure is far from being trivial. It would occur if the released data enabled the intruder to identify or determine the value of the sensitive confidential information with a fair degree of certainty [89, 142]. We refer to this

kind of disclosure as *"Predictive Disclosure"*, because it concerns the prediction of the unknown (or unspecified) value of a variable. If the predictive disclosure can be made with certainty, it is referred to as being a *"deterministic"* predictive disclosure. This is contrasted with a *"probabilistic"* predictive disclosure, which involves knowing the unknown value within a probabilistic framework. The deterministic type of predictive disclosure can further be classified as being either *"exact"* or *"approximate"*. In the exact case, the intruder predicts the exact value of the confidential information, while in the approximate case the intruder is able to deduce an interval for the confidential value [123, 142]. Other common types of disclosure are listed below [89, 117]:

- True Disclosure. This occurs when the exact or approximate value of the sensitive data represents the real status of the attribute at the time of dissemination.
- Apparent Disclosure. This occurs when the exact or approximate value of the sensitive data represents a reasonable, but not the exact, value of the attribute at the time of dissemination.
- Positive Disclosure. This occurs when the information learned pertains directly to the identified unit.
- Negative Disclosure. This occurs when the information learned does not pertain directly to the identified unit but to its compliments.

Since DR can be defined as the risk caused by using the protected data set V' to deduce the confidential information on an individual found in the original data set V, it can be viewed from two different perspectives [23]:

- Attribute Disclosure
  This occurs when the attribute of an individual can be determined more accurately when the user has access to a related statistic, as opposed to the case when the user does not have this access. In other words, attribute disclosure occurs when the respondent can be associated with either the exact attribute value

in the disseminated data or an estimated (approximated) attribute value based on the disseminated data [23]. In this case, a rank-based interval technique can be used to measure the attribute disclosure. Each attribute in the protected data set, V', is ranked independently. A rank interval is defined around the value that the attribute takes in each record, r. The ranks of values within the interval for an attribute for records around r should differ by less than a certain percentage of the total number of records, and the rank in the center of the interval should correspond to the value of the attribute in record r. If this is true, the proportion of original values that fall into the interval centered around their corresponding protected values is a measure of DR. For more information about this metric, the reader is referred to [23, 44, 48, 53]. Further details about this technique will be included in Chapters 4 and 5.

- Identity Disclosure

  This occurs when a record in the protected data set can be linked with a respondent's identity. In other words, the identity disclosure permits a respondent's identity to be associated with the disseminated data records. This association is assumed to be exact [23, 44, 48]. Identity disclosure can be further classified as into two categories.

  - *Uniqueness*

    The risk of identity disclosure can be measured as the probability that rare combinations of attribute values in the set V' are also rare in the set V. Such a quantifying approach is used in the context of non-perturbative SDC methods [23, 44, 48]. Here, we further distinguish two kind of measures: File-level risk measures and record-level risk measures. In the case of file-level risk measures, DR is defined as the probability that a sample that is unique is also unique to the entire population. Generally, the size of the sample is much smaller than the size of the population. If this was not the case, an intruder would be able to locate the unique value in the released sample, which, in turn, would lead to the identification of the individual respondent in the original population [53]. A record-level risk

measure is defined as the probability that a particular sample record is re-identified. The author of [63] integrates two methods for measuring DR at the record-level with file level risk measures. More details of these DR measures can be found in [23, 44, 48, 72, 77].

- *Re-identification*

  The concept of re-identification quantifies the estimate to the number of re-identifications that can be obtained by an experienced intruder. Re-identification methods are schemes used to measure the DR, and they provide more unified approaches to this problem than the uniqueness methods. This is because a re-identification method can be applied to both perturbative and non-perturbative SDC. In this case, a record linkage technique is used to measure the identity disclosure in SDC. It consists of linking each record, a, in the set V', to a record, b, in the set V. The pair (a, b) is a match if the record b turns out to be the record a. It is important to mention that measuring the risk of identity disclosure also assumes that the intruder has an external data set that shares some attributes with the protected data set and that also contains some identifier attributes. Thus, the intruder has to link the protected data set with the external data set using the shared attributes. The number of matches will then represent the number of protected records whose corresponding respondent values can be re-identified by the intruder. The main types of record linkage, which are typically used to measure the identity disclosure in SDC, are discussed below [23, 44, 48].

  * *Distance-based Records Linkage*

    Distance-based record linkage measures the scenario when each record a in the set V' is linked to its nearest record b in the set V. In such a case we require a definition of a distance function for expressing nearness between records, and there also is a need to standardize attributes (so as to avoid scaling problems) and to assign each attribute a weight on the record-level distance. The linkage computation then proceeds by evaluating the distances between records in the sets V

and V'. To do this, the distance to every record in the set V is first computed. Thereafter, the *"nearest"* and *"second nearest"* records in V are considered. A record in the set V' is labeled as being "linked" when the nearest record in the set V has the same record number (or index), while a record in V' is labeled "linked to second nearest" when the second nearest record in the set V has the same record number (or index). In all other cases, the record in V' is labeled as being "not linked". The main advantages of using a distance measure for record linkages are simplicity for the implementer and intuitiveness for the user. The main difficulty appears in determining an appropriate distance for categorical attributes, and for masking methods such as those which use local recording when the masked protected file contains new labels with respect to V. Finally, this technique is used to measure the DR of the micro-aggregation in the context of the Euclidean distance [23, 44, 48, 53]. The authors of [132] explore the use of the Mahalanobis distance, which leads to a better estimation of the true dissimilarity, and thus to a higher level of protection against attackers. This techniques will also be discussed in more details in Chapters 4 and 5.

* *Probabilistic Record Linkage*

  Probabilistic record linkage aims at linking pairs of records (a, b) in the sets $V'$and V, respectively. The user needs to provide as input two probability values: The maximum acceptable probability of linking a pair that is an unmatched pair, (LT), and the maximum acceptable probability of not linking a pair that is a match, (NLT). For each pair, the user computes the corresponding indices. Then, LT and NLT in the index range are used to label the pair as being either *linked*, *clerical*, or *non-linked* as follows: If the index is above LT, the pair is *linked*; if it is below NLT, the pair is *non-linked*; if it is between LT and NLT, the pair is *clerical*, which is the case when the pair cannot be automatically classified but requires human inspection.

The measure for DR is then defined as the number of matches over the number of records in the set V'. Although the probabilistic record linkage scheme is more sophisticated than the distance-based one, it does not, on the other hand, require a re-scaling or a weighting of the attributes [23, 44, 48, 53, 146], and is, thus, often a preferable option.

* *Other Record Linkage Methods*
  Recently, the use of other record linkage methods has also been considered for assessing DR. While, previously, the record linkage methods assumed that the two files contained shared variables, other methods that relaxed this condition have been recently developed. Some methods use the correlations between the variables, or auxiliary population files, to improve the linkage accuracy. Alternative methods use Markov Random Files, or graph partitioning algorithms so as to increase the likelihood of a set of linkages between corresponding records in a group of files. It should be mentioned that these most recent methods are often used for extracting and linking information from a group of webpages [144].

2.4.2 Micro-data Information Loss

The degradation of the data, of course, reduces the ability of data users to conduct the analysis they need for legitimate purposes. These effects fall into two categories [21, 72]:

- Reduction of analytical completeness
  Some control methods, (typically those involving the recoding of taxonomic schemes into coarser categorizations), that could have been conducted with non-recoded data are not achievable.

- Loss of analytical validity
  The loss of analytical validity is harder to define, but in some ways more critical

because of its insidious nature. Technically, loss of validity can be said to occur when a disclosure control method has changed a data set to the point where a user arrives at a conclusion different from what a similar analysis on the original data set would have led to.

The loss in the content of information depends on the usage of data. Data is usually used in various areas, and it is very difficult to identify all its uses at the time of dissemination. Additionally, assuming all data uses can be identified, issuing several versions of the same original micro-data may result in unexpected disclosure. Thus, it is more reasonable to measure the amount of the IL in a generic way, which estimates how much the data is harmed by using any SDC method [21, 55, 93]. Generally speaking, the IL is very small, if the analytical structure of the micro-data file, after the masking is very similar (nearly identical) to its structure before the masking. In fact, preserving the structure of the micro-data file is very important to ensure that the new masked file will be analytically valid and interesting [44, 48, 93].

It is important to give a clear definition for both the terms *"analytically valid"* and *"analytically interesting"* [21, 55, 93].

- Analytically Valid. The micro-data file is analytically valid if, with respect to the original file, it approximately preserves the following:

  1. The means and covariances on a small set of sub-domains.
  2. The marginal values for a few tabulations of the data.
  3. At least one distributional characteristic.

- Analytically Interesting. The micro-data file is analytically valid if at least six attributes on the important sub-domains, which can be validly analyzed, are provided.

It is difficult to come up with a precise description of analytical validity and analytical interest without determining where the data will be used. But, several

complementary ways are highlighted to ensure the preservation of the structure of the original data [44, 48, 55, 93]:

1. Compare the original micro-data with the masked micro-data. The more similar the SDC method is to the "identity" function, the less impact there is on the information content.

2. Compare some statistics computed on the original and masked micro-data. Small differences between the statistics yield to little loss in the information content.

3. Analyze the behavior of the particular SDC method used to measure its influence on the structure of the original micro-data.

An automatic data modification procedure can be used to measure the amount of the IL. Theoretically, if the original micro-data file is considered as a message to be transmitted, and the masked micro-data file as the transmitted message, and the modification that separates them as a noisy channel, then the idea is to use information theoretic concepts to quantify the IL quickly follows. Although this idea is possible to some extent, it is not practical because it is time-consuming [142].

Generally speaking, a data protector who protects the micro-data file usually knows how to protect the data and how to optimize the loss in the information content. Thus, a formal measure for the IL is used in order to compare the different possibilities of getting a safe micro-data file [142]. It is important to highlight the formal measures used to estimate the IL, which can be classified into two categories based on the type of the used data [21, 44, 48, 55, 93]:

1. ***Information Loss Measure for Continuous Data***
   There is no single quantitative measure that reflects the difference between the original and the masked continuous micro-data file. Therefore, the following tools are used to characterize the information contained in the micro-data file

[21, 44, 48, 55, 93]. It is most likely that these various tools are used to measure the loss in the data utility, (*i.e.,* the loss of the accuracy sought by database users), through the discrepancies between the matrices X, V, R, RF, C, and F obtained from the original data and the corresponding X', V', R', RF', C', and F' obtained from the protected micro-data file, where:

X and X'	:	are the matrix representing the original and protected micro-file, respectively, which consists of n records and p attributes.
V and V'	:	are the covariance matrices of X and X', respectively.
R and R'	:	are the correlation matrices of X and X', respectively.
RF and RF'	:	are the correlation matrices between the p factors $PC_1, \ldots, PC_p$ obtained through principle components analysis from X and X', respectively.
C and C'	:	are the commonalities between each of the p variables and the first principle component PC for X and X', respectively.
F and F'	:	are the factor score coefficient matrices, which contain the factors that should multiply each variable in X and X', respectively, to obtain its projection on each. principle component.

Matrix discrepancy can be measured in at least three ways [21, 44, 48, 55, 93].

(a) *Mean square error*

Sum of squared component differences between pairs of matrices, divided by the number of cells in either matrix.

(b) *Mean absolute error*

Sum of absolute component differences between pairs of matrices, divided by the number of cells in either matrix.

(c) *Mean variation*

Sum of absolute percent variation of components in the matrix computed

on protected micro-data with respect to components in the matrix of the original micro-data, divided by the number of cells in either matrix. This approach is not affected by scale changes of the attributes. In other words, in this case, there is no need to standardize the attributes.

Measuring the IL using any (or all) of the various tools, which have been discussed above, has a major disadvantage: The measure is only defined as a function of the protected masked micro-data, which does not depend on the original micro-data. The authors of [55] measure the information loss for a variable, V, as a function of three elements: The condition probability, the original micro-data i, and the masked category j. Thus, the per-record IL when $V = i$ is masked as $V' = j$ can be defined as:

$$P(P_{V,V'}, i, j) = -\log P(V' = j|V = i).$$

Measuring the IL for the whole micro-data file is then defined as:

$$P(P_{V,V'}, O, M) = \sum_{r \epsilon M} P(P_{V,V'}, i_r, j_r),$$

where i_r is the value taken by the variable V in record r of the original micro-data file, O, and j_r is the value taken by the variable V' in record r of the masked micro-data file, M [93].

Finally, it is worth mentioning that the quantity IL is to be traded-off with the DR. Since DR is bounded, there is no risk higher than 100 percent. On the other hand, the upper bound does not practically exist, but it should be enforced for IL measures [21]. The authors of [93] have shown a way to obtain probabilistic information loss measures for assessing the impact of SDC methods on continuous micro-data sets. Being probabilistic, a value between 0 and 1 makes the protection easier, by finding an optimal balance between IL and DR.

For more information about the various tools that are used to estimate the amount of the IL, and on measuring the matrix discrepancy, the reader is

referred to [21, 44, 48, 55, 93]. The question of estimating the general IL will be explained in more detail in Chapters 4 and 5.

2. ***Information Loss Measure for Categorical Data.***
   Applying the above measures on categorical data is not possible. For categorical data, three kinds of IL measures have been proposed in the literature [21, 44, 48, 55]:

   (a) *Direct Comparison of Categorical values*
   The comparison of the original and the protected matrices requires a clear definition of the distance for categorical attributes. The distance is measured by considering the distance between pairs of categorical values, when comparing a record in the original micro-data file and its corresponding record in the masked version [21, 55].
   Measuring the distance for nominal attributes only uses the comparison of equality. But for the ordinal attributes, the distance is measured by the total number of categories between them, measured over the cardinality of the range of the ordinal attribute. More information on this issue can be found in [21, 55].

   (b) *Comparison of Contingency Tables*
   An alternative strategy to the direct comparison of the values in the case of categorical variables is to compare their contingency tables. Contingency tables are compared for a file before and after applying the masking method. The number of differences between both contingency tables quantifies the amount of the IL. Usually, the number of cells in the contingency table depends on the number of categories in the variable; thus, a normalized version exists, which is exactly the number of differences between the two contingency tables divided by the number of cells in all the tables considered. For more information about such a metric, the reader is referred to [21, 44, 55].

   (c) *Entropy-Based Measures*
   Entropy is essentially an information theoretic measure, but it can be used

in *SDC* if the masking process is modeled as the noise that would be added to the original micro-data in the event of its being transmitted over a noise channel [21, 83]. The use of entropy-based methods is limited to *PRAM*, because of the issues involving generalized noise addition, local suppression, and global recoding methods [48].

Let V be a variable in the original micro-data file and V' be the corresponding variable in the masked file. The condition uncertainty of V given V' can be deduced from the Markov matrix in *PRAM* as follows [21, 48]:

$$H(V|V'=j) = -\sum_{i=1}^{n} P(v=i|V'=j)\log P(V=i|V'=j).$$

Thus, the loss in the information content can be calculated by accumulating the above expression for all individuals in the masked micro-data file, M as [21, 48]:

$$IL(P_{V,V'}, M) = \sum_{r\epsilon M} H(V|V'=j_r).$$

2.4.3 Trade-off between DR and IL

Government agencies collect various types of data, but due to access restrictions, micro-data is usually accessed using a secure computer system. Of course, this access restriction affects public policy decisions made by any agency that has access to the non-confidential statistical summaries. This requirement to access the original micro-data should not only satisfy the confidentiality requirements, but also the sufficient utility [82].

There is a very broad choice of methods for continuous micro-data protection (see Section 2.1). Most of these methods are parametric (*i.e.*, in the micro-aggregation technique one parameter is the minimal number of records per group). Therefore, the user has to go through two choices rather than a single choice: A primary choice to select the masked method and a secondary choice to set an appropriate value for the parameters of the selected method [93].

The optimal way to choose the appropriate method and parameter should yield to an optimal trade-off between IL and DR. This brings us to the need to combine the measures of IL with those related to DR. Generally speaking, there are two approaches to do this [93]:

1. Explicit
   This method adapts a score for the method performance rating. Typically, this method combines the average values of the IL and DR. Using a score defines a new optimization problem by selecting a masking method and setting the appropriate values for its parameters [93].

2. Implicit
   Methods of this family assume that there is no specific score that can do justice to all methods, for all data uses, and for all disclosure scenarios. Thus, it is the responsibility of the data protector to separately compute several IL and DR measures, and to choose the most appropriate method based on a combination of the most relevant measures for specific data use/disclosure cases [93].

It is important to present a general principle that shows how to balance the trade-off between IL and DR. There are various strategies to obtain and reach the equilibrium point between them [27]. If R represents the disclosure risk and U the utility, one strategy involves computing $R-U$ for each method, S, and parameter, P. The score is computed as follows: $Score(O,M) = \frac{IL(O,M)+DR(O,M)}{2}$, where M is the protected file obtained after applying method S with parameterization P to the original file. Otherwise, the $U - R$ map, which is a graphical representation tool, is used. According to [27] "in its most basic form, $R-U$ map is the set of paired values (R,U) of DR and data utility that correspond to many strategies for data release". Such (U,R) pairs are plotted on a two dimensional graph, so that the user can easily understand the influence of a masked method and/or the parameter choice [27].

A strategy that compromises between preserving the confidentiality requirements and maximizing the data utility was introduced by Kim *et al.* in [82]. The basic idea

here is to mask the multi-variate normal quantitative data using an additive noise approach and then, when it is necessary, use a re-identification/swapping approach to assure confidentiality. The most important advantage of this strategy is that of obtaining exact sub-population estimates, because this involves applying a controlled distortion that renders the estimates from the masked file nearly identical to those from the original file. This method has been extended to mask non-normal multi-variate data by involving transformations that preserve the essential characteristics of the data such as its mean, variance, and correlations [129].

The authors of [124] state that they can improve the performance of any specific SDC method by involving a post-masking optimization technique. The latter strategy preserves the moments of the first and second order by minimizing the IL between the original and the masked micro-data file. Simultaneously, it does not increase the DR because it always prevents the optimized masked file from being very close to the original file.

In conclusion, consideration must also be given to the conceptual framework, since one must decide whether to publish the information after using one of the SDC techniques or not. This decision is a crucial one, because we have to estimate the loss associated with the possibility of disclosure and, simultaneously, the loss attained by not publishing the data. A Bayesian model can be used to make this critical decision. The final decision of whether to publish the data or not is determined by evaluating whether the risk of publication exceeds the loss of non-publication [150].

2.5 Micro-data Protection

Micro-data protection methods usually generate a new released version of the statistical data by either of the following schemes [20]:

- Generating synthetic data

  Using synthetic (or simulated) data was proposed decades ago as a method to

guard against *DR* [26]. Instead of releasing the real data, the scheme suggests generating a set of synthetic micro-data from a model fitted to the real data. The idea is to randomly generate data with the constraint that certain statistical or internal relationships of the original micro-data are preserved [142]. It would appear as if the synthetic data would have the advantage of circumventing the re-identification problem, because the published records are *"invented"* and are not derived from the real data. This is, however, not the case, because the synthetic data usually *over-fits* the original data, which may, in turn, be disclosed in the same way as the original data [20, 29, 144]. On the other hand, an obvious problem associated with using synthetic data is its data utility [47]. The only preserved statistical properties are those that are explicitly selected by the data protector [20, 47, 142]. More information about these methods can be found in [26].

- Masking original data
  Masking methods modify the original data in a special way. These modifications seek to preserve the confidential information, while providing valid and interesting analytical statistical information. *SDC* methods for micro-data are usually known as *"masking"* methods. Generally speaking, masking methods can be classified based on two different points of view [20, 55, 142]:

  - *Operational principle*
    Masking methods that work on the basis of an operational principle can, in turn, be further classified into two categories depending on their effect on the original data [20, 142, 53, 55]:

    * Perturbative
      The confidentiality of the statistical data is preserved by distorting the micro-data. This distortion may add new and unique combinations to the released version of the data, or it could delete some unique combinations of scores. All perturbative methods should provide the data set with analytical results similar to the results that would be obtained from the original data set [20, 25, 55].

* Non-perturbative

  These methods produce partial reduction of the details in the original data, instead of harming and altering the original data [20, 24, 55]. This can be achieved by making sure that every unsafe combination must not have a frequency below a predefined threshold value [142].

– *The data type on which they are used*

  This is an alternate classification of the masking methods based on the type of data that are used. This classification divides the masking methods into two parts: (i) Continuous, in which case the SDC methods are applied on continuous variables, and (ii) Categorical, which is the case when the SDC methods are applied on categorical variables [20, 55].

- Generating hybrid micro-data

  The third possibility to build a hybrid data set as a mixture of the masked original values and synthetic data set is to combine a masking method with Cholesky's decompositions [47]. A hybrid micro-data file offers numerous advantages when compared to the traditional micro-data files: (i) It overcomes the lack of analytical validity by preserving both the uni-variate and multivariate statistical properties of the original micro-data file, such as their means, covariances, and correlations. Besides this, it also preserves a fair amount of record-level similarity between the released and original files, which, in turn, allows sub-domain analysis [47]. (ii) The method is very fast, because its running time is linear in the number of records. (iii) It overcomes the identification problem by publishing the simulated data and not the original real data.

In this research, we will focus on masking methods, and an overview of the state-of-the-art of these methods will be covered in detail in Sections 2.5.1 and 2.5.2.

2.5.1 Perturbative Masking Methods

Perturbative methods release the entire original data. But, instead of releasing the exact value of the attributes, a perturbed value is released for each attribute. It is important to mention that not all the perturbative methods are designed for continuous data. Table 2.1 indicates whether the perturbative methods are suitable for continuous and/or categorical data [25, 55, 76].

Table 2.1: Perturbative methods for various data types.

Perturbative Methods	Data types	
	Continuous	Categorical
Additive Noise	X	
Micro-Aggregation	X	X
PRAM		X
Data Swapping	X	X
Rounding	X	
Resampling	X	
Lossy Compression	X	

The techniques which are known as additive noise, micro-aggregation, data swapping, and post-randomization are special cases of matrix masking. The released set V' can be computed from the set V as follows:

$$V' = A\,V\,B + C \tag{2.1}$$

where A is a record-transforming mask, B is a variable-transforming mask, and C is a displacing mask (noise) [25, 55, 102]. We detail these methods below.

- Additive Noise

  Masking with additive noise is one of the data perturbative techniques that statistically seeks to protect confidential micro-data by adding random noise to confidential numerical attributes, where the noise possesses the same correlation structure as the original data [13, 55, 102]. Note that this technique

does not belong to one of the encryption techniques, where the data is initially modified by a secure encryption, then transmitted, and finally the received data is "decrypted" by the user.

The main additive noise algorithms reported in the literature are: Masking by adding uncorrelated noise, masking by adding correlated noise, masking by adding noise involving linear transformations, and masking by adding noise involving non-linear transformations. For more details on the specific algorithms for each of these topics, the reader is referred to [13, 19, 76].

Most masking methods combine transformations by adding noise [50, 70]. White noise is most frequently used, even though it may be subject to the *bias problem.* Despite this bias problem, white noise is still built into the μ-ARGUS software package [48, 55, 77, 102]. Studies demonstrate that additive noise perturbative techniques exhibit bias when the results of a database query on perturbed data produces a significantly different "result" than the same query executed on the original data [72]. Five types of biases have been identified as [143]:

- Type A bias occurs when the perturbation of a given attribute causes summary measures (i.e., mean value) of that individual attribute to change due to a change in variance.
- Type B bias occurs when the perturbation changes the relationships (e.g., correlations) between the confidential attributes.
- Type C bias occurs when the perturbation changes the relationship (again, e.g., correlations) between confidential and nonconfidential attributes.
- Type D bias deals with the underlying distribution of the data in a database, specifically on whether or not the data has a multi-variate normal distribution.
- Type DM bias could severely impact the ability of an organization to gain significant benefits of knowledge management/discovery from the disclosed data.

Finally, the nature of additive noise makes it unsuitable for categorical data.

On the other hand, it is well suited for continuous data for the following reasons [25, 55, 102]:

- There are no assumptions on the range of possible values for the attributes.
- The random noise, which is added to the original data, is typically continuous and has a mean of zero.
- Approximate matching is possible with external files, because it depends on the amount of noise that is added.

- Micro-Aggregation

  Micro-aggregation is a perturbative technique applicable to quantitative variables [142]. The rationale behind this technique is that confidentiality rules allow publication of the original data, if records correspond to groups of k or more individuals, where no individual dominates the group and k is a threshold value [55]. The basic principle of micro-aggregation is as follows: Records are clustered into small aggregates or groups of size at least k. Rather than publishing the original variable for a given individual, the average of the values over the group to which the individual belongs is published [34, 45, 55, 102]. Micro-aggregation can be viewed as a technique like rounding or noise addition, although it has the property of preserving the grand total of the data involved [142]. In spite of keeping the total for a variable intact, it has a tendency to reduce the variances of the variables to which it is applied [142].

  Generally speaking, generalizing uni-variate micro-aggregation to yield a multivariate micro-aggregation scheme can be done using several variants such as *fixed vs. variable group size* [45, 92, 122], *exact optimal vs. heuristic micro-aggregation* [45, 92, 77, 103], and *categorical micro-aggregation* [131, 56, 60].

  Finally, individual micro-aggregation, micro-aggregation using z-scores projection and principle component projection, and micro-aggregation on unprojected multi-variate data considering two variables at a time, three at a time, four at a time, or all variables simultaneously are built into the μ-ARGUS Software [48, 76, 77, 102].

In this research, we will focus on techniques within this family. They will, thus, be covered in greater detail in a subsequent section. Our primary results in this research endeavor deal with novel micro-aggregation strategies.

- PRAM

  The Post-RAndomization Method (*PRAM*) is a probabilistic, perturbative method that has been inspired by the so-called randomized response technique [142]. It is used to protect the categorical attributes in the micro-data file [25].

  PRAM produces a new masked file by changing the scores of some variables for certain records in the original file, according to a prescribed probability mechanism (typically, a Markov matrix) [83]. The random procedure in *PRAM* methods is based on matching a record in the masked file with a known individual in the original file (which occurs with a very low probability) [73]. The Markov approach makes *PRAM* very general, because it consists of noise addition, data suppression, and data recording [25].

  Generally speaking, although the IL and DR issues associated with a *PRAM* technique largely depend on the choice of the Markov matrix, (which is still an open research topic [145]), it has been implemented in the μ-ARGUS Software package [48, 76, 77, 102]. Finally, it is worth mentioning that the application of this method on continuous data is still open, since to-date, to the best of our knowledge, *PRAM* has only been applied on categorical data [25].

- Data Swapping and Rank Swapping

  Data swapping is considered a special case of *PRAM* [142]. The basic idea behind this method is to apply a sequence of elementary swaps of confidential attributes between individual records in the micro-data files. Records are exchanged in such a way that lower-order frequency counts (or marginals) are maintained [25, 142]. An elementary swap consists of two actions:

  1. A random selection of two records, say i and j, from the micro-data set.
  2. An interchange of the values of the variable being swapped for records i and j.

The basic rationale for data swapping has also left a marked impact on subsequent methods, as reported in [25].

Another important variant of the data swapping strategy is *Rank Swapping* [44]. The latter method was originally proposed for ordinal categorical variables. But since then, it has also been used for other numerical variables [96]. To achieve this, first of all, values of an attribute are ranked in an ascending order, and then each ranked value is swapped with another ranked value that is randomly chosen within a restricted range (*i.e.,* the rank of two swapped values cannot differ by more than $P\%$ of the total number of records, where P is an input parameter). This algorithm is independently used for each original attribute in the original data set [44, 48, 55]. The authors [53] argue that a rank swapping method can be identified as being a method that yields a good trade-off between DR and IL. Consequently, this scheme has also been implemented in the μ-ARGUS Software package [48, 76, 77, 102].

- Rounding

  The rounding method is used only for continuous data. It replaces the original value of a variable by the rounded value[4] [55]. For a given variable, rounded values are chosen from a set of rounding points defining a *rounding set.* This value can either be chosen deterministically or randomly. An example of a deterministic rounding procedure is to round a value to the nearest integer multiple of the rounding base. A stochastic random rounding procedure could round each value, with certain probability, to an integer multiple of the rounding base [142]. In the case of multi-variate micro-data, rounding is usually applied to one variable at a time (uni-variate rounding), although examples of multivariate rounding have been reported [55].

- Resampling

  The resampling method was originally proposed to protect tabular data, but resampling can also be used to protect numerical micro-data [43, 27]. The

[4]The rounded values are integer multiples based on a well-chosen rounding base [142].

basic idea is taking t independent samples $S_1, S_2, \ldots, S_t$ of size n of the values of a variable X. Then, the values are independently ranked using the same ranking criterion. Finally, the masked variable X' as $x'_1, x'_2, \ldots, x'_n$, (where n is the number of records) is computed for each value, where x'_i is obtained by evaluating the average value of the i^{th} ranked value in each sample $S_1, S_2, \ldots, S_t$ [25, 44, 55, 102]. Resampling has been implemented in the μ-ARGUS Software package, and it has been tested for $t = 1$ and $t = 3$ [48, 77, 102].

- Lossy Compression

  Lossy compression is a new method that can be applied to continuous data[102]. The basic idea is to consider the micro-data file as an image (where records are rows, variables are columns, and the values themselves are pixels[5]). A lossy compression method is used on the micro-data file as follows: First of all, the original micro-data values have to be scaled to be integers into the interval of the gray-scale values, in order to produce a gray-scale image. Secondly, a compression algorithm (*i.e.,* JPEG) is applied to yield a certain percentage, $P\%$, quality on the image. Finally, a masked file is generated by unscaling (using the inverse of the scale transformation used in the first step) the image [25, 44, 48]. Here, it is important to mention that in spite of requiring an appropriate mapping between the attribute ranges and the color scales, (which depends on the compression algorithm used [55]), lossy compression methods have been implemented in the μ-ARGUS Software package. More information of such methods can be found in [48, 77, 102].

2.5.2 Non-Perturbative Masking Methods

As mentioned in Section 2.5, non-perturbative methods do not destroy the original data, but reduce the detail in the original file. Some similar methods have been applied for categorical and continuous data, but others have only been suitable for categorical data. Table 2.2 lists the non-perturbative methods, and also indicates the

[5]Typically, for a black and white image, the value of the pixel is its gray-scale level [55].

data for which each method is suitable [24]

Table 2.2: Non-Perturbative methods for various data types.

Non-Perturbative Methods	Data types	
	Continuous	Categorical
Sampling		X
Global Recoding	X	X
Top and Bottom Coding	X	X
Local Suppression		X

- Sampling

  Masking with a sampling method requires publishing a sample (or a subset) of the original records in the micro-data file, instead of publishing the entire records in the original file [55]. A sampling method is suitable for categorical micro-data, but its adequacy for continuous data is less clear. The main reason behind this is that the sampling method leaves the continuous attribute unperturbed for all records in the sample [24, 55]. This, of-course, can be easily identified, especially if the intruder matches the value of the attribute in the masked file with another value in an external administrative public file [144]. More information about the technical specifications of a real-world application of sampling can be found in [24].

- Global recoding

  Global recoding schemes, which are intensively used by statistical offices, are more appropriate for categorical micro-data, because they help disguise records with unusual combinations of categorical attributes [24, 55]. In such a scheme, for a categorical attribute, several categories are combined to form a new category, and thus the resultant category should satisfy $|D(V_i')| < |D(V_i)|$, where $|.|$ is the cardinality operator [24, 55, 102]. In other words, two or more categories of a variable in a micro-data file are combined into a single one, and the same coding is used for all the units. Of-course, reducing the number of

categories of a variable implies a reduction of the information content of the variable concerned [142].

Global recoding has been used for continuous variables by replacing the continuous variable, V_i, with another discretized version of V_i. In other words, a potentially infinite range $D(V_i)$ is mapped into a finite range $D(V_i')$ [24, 55]. The reader has to remember that by discretization, one inherently obtains an unaffordable loss of information. Also, arithmetic operations that were straightforward on the original V_i may be easily applied on the discretized version [102].

Implementing the process of global recording in computer software may either take place in an *automatic way*, where the choice of categories that are to be combined is determined according to an objective criterion, or in an *interactive way*, where the choice of categories to be combined is left to the user of the software [142]. This technique has also been implemented in the μ-ARGUS package [48, 76, 77, 102].

- Top and bottom coding

  Top-down coding is a special case of the global recoding technique. It is used on variables that can be ranked, such as continuous and ordinal categorical data [24, 44, 48, 102]. While a top category covers the values of the variable above a certain upper threshold, θ_U, a bottom category covers the values below another specified lower threshold, θ_L. The critical practical problem encountered in this technique is that of determining the appropriate values for θ_U and θ_L [142]. This method has also been implemented in the μ-ARGUS software package [77]

- Local suppression

  Local suppression is a non-perturbative technique that involves deleting a value in a record and replacing it by a missing indicator [24, 55]. This procedure is applied to a single record at a time, and it does not require the deletion of all the values of the same variable. Thus, local suppression is not harmless when it comes to imputing values for missing variables [142].

  In practice, local suppression is combined with global recoding, (which is the

primary *SDC* technique for a micro-data file), so as to remove a few remaining instances of unsafe records [142].

Generally speaking, this techniques is more applicable for categorical data than for continuous data. This is because, sometimes, the continuous attribute is a component of a set of key attributes, which renders each combination of key values unique. Since the suppression of the values of a continuous attribute is meaningless, it is generally assumed that local suppression cannot be used for continuous variables [24].

2.6 Micro-Aggregation Techniques (*MATs*)

2.6.1 Introduction

At present, a large number of agencies have collected useful data suitable for research purposes. A decision to prohibit these useful pieces of information from being disseminated has sometimes been made in an attempt to preserve the confidential nature of the data. Therefore, as explained earlier, there is a tacit conflict between the principle of privacy and the need to discover new characteristics found in the data. As, in practice, neither principle can be totaly preserved, often, a trade-off between them is acceptable, and in order to access the confidential micro-data at lower costs, we have seen that it is expedient to apply an *SDC* technique.

As explained earlier, various methods have been proposed and applied to ensure a good measure of confidentiality and data utility. Each method either provides a modified version of the entire data set, or the original exact data set where sensitive values of the variables are either erased or re-coded [115].

One of the most recent techniques proposed is the strategy called "Micro-aggregation". The latter comprises of a family of statistical disclosure limitation techniques used to protect micro-data files, each of which contains records on individual data subjects.

These belong to the family of substitution/perturbation approaches [10, 28, 45, 76, 92, 115], where the individual values are replaced by values computed on small aggregates *prior* to publication. In other words, instead of releasing the actual values of the individual records, the system releases the mean of the "group" (or any other measure of central tendency, *i.e.,* median, mode, weighted average [115]) to which the observation belongs. The confidentiality of the individual data subjects is protected by ensuring that each group has at least a minimum number of observations, k [54]. If this were not the case, the average would not avoid disclosure because an individual contributing to the partition/ group, or an external individual, can guess the value of another respondent [22, 131]. The objective of this technique is to group similar records together, so that the replacement of actual values by the means of *their* associated groups, results in minimizing the IL [38, 74, 84].

Basically, $MATs$ can be operationally defined with respect to two dominant successive phases [54, 131]:

1. *Partitioning*
   The original micro-data file is partitioned into several disjointed clusters/groups so that all records in the same group are similar to each other and, simultaneously, dissimilar to the records in other groups. Additionally, each group is forced to contain at least k records.

2. *Aggregation*
   This phase computes a certain kind of prototype for each cluster/ group, and it replaces the original values in the micro-units by the computed prototype. This phase usually depends on the type of the variable concerned.

The actual implementation of the MAT requires a clustering method and an aggregation method. Such methods were originally used for numerical data types. In this Thesis, we will concentrate on classes of numerical $MATs$, and will define the Micro-Aggregation Problem, based on such numerical variables.

2.6.2 Micro-Aggregation Problem (MAP)

The MAP, as formulated in [22, 45, 74, 84, 92], can be stated as follows: A micro-data set $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ is specified in terms of the n "individuals", namely the $U_i's$, each representing a data vector whose components are p continuous variables. Each data vector can be viewed as $U_i = [u_{i1}, u_{i2}, \ldots, u_{ip}]^T$, where u_{ij} specifies the value of the j^{th} variable in the i^{th} data vector, and it represents the j^{th} dimension in the space. Associated with the problem is a positive integer, k, referred to as the *security parameter*. Micro-aggregation involves partitioning the n data vectors into m groups so as to obtain a k-partition $\mathbb{P}_k = \{G_i \mid 1 \leq i \leq m\}$, such that each group, G_i, of size, n_i, contains either k data vectors (fixed-size case), or between k and $2k-1$ data vectors (data-oriented case). Each data vector is contained in exactly one group, implying that $\mathbb{U} = \bigcup_{i=1}^{m} G_i$ and $G_i \bigcap G_j = \emptyset$. The j^{th} data vector in the i^{th} group is denoted by X_{ij} (where each X_{ij} is an element of $\mathcal{U}$), while $\bar{X}_i$ is the average of the data vectors over the i^{th} group, and $\bar{X}$ is the average of the data vectors over the entire set of all the n elements of $\mathcal{U}$. Thus, $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$, and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

The optimal k-partition, $\mathbb{P}_k^*$, is defined to be the one that maximizes the within-group homogeneity[6], since a larger value of this index implies a smaller information loss. The within-group similarity is defined as the *Sum of Squares Error*, (SSE), computed on the basis of the Euclidean distances of each individual data vector X_{ij} to the centroid $\bar{X}_i$ of the group to which it belongs. It is given by:

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^T (X_{ij} - \bar{X}_i). \tag{2.2}$$

Analogously, the between-groups similarity is defined as the *Sum of Squares Among* the groups,(SSA), which is the squared deviations of the means from the total mean.

[6] A large number of measures which quantify the "group homogeneity" have been reported in the literature. There are usually based on several distance definitions, such as the Euclidean distance, the Minkowski distance, and the Chebyshev distance. The most common homogeneity measure for clustering is the within-group sum of squares, the SSE [52].

It is given as:

$$SSA = \sum_{i=1}^{m} n_i(\bar{X}_i - \bar{X})^T(\bar{X}_i - \bar{X}). \tag{2.3}$$

The *Total Sum of Squares* is denoted by $SST = SSA + SSE$, or explicitly

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^T(X_{ij} - \bar{X}). \tag{2.4}$$

The *Information Loss* is quantified by a measure, IL standardized in $[0, 1]$ as:

$$IL = \frac{SSE}{SST}. \tag{2.5}$$

The *Information contained* in the micro-aggregated data is given by the ratio:

$$\frac{SSA}{SST} = 1 - IL. \tag{2.6}$$

It is important to highlight that the analysis of variance methods can be used as alternative methods to investigate the degree of information that is retained, or, equivalently, that is lost, by using the MAT. More details regarding this metric can be found in [115].

2.6.3 State-of-Art: MATs

The principle of the MAT is based on the concepts related to the confidentiality of tabular data, where any cell of a table that contains data relative to less than three units, or which is dominated by a single unit, should not be disseminated so as to be consistent with a confidentiality rule. This principle is applied to a micro-data file by aggregating a fixed number of micro-units to avoid being dominated by any one of

them. Therefore, an *MAT* proposes that individual micro-units are replaced by the average of a subset of them [31, 32, 142].

Basically, an *MAT* relies on a clustering technique and an aggregation technique: *MATs* were originally used for numerical data [22, 131], and they can be further classified as *Uni-variate methods* or *Multi-variate methods* [28, 91].

2.6.3.1 Uni-Variate MATs

Practical heuristic-based *MATs* have been proposed in [8, 22, 31, 33, 34]. The partitioning mechanism advocated (for uni-dimensional data) in all these papers is the same: First of all, the elements are ranked in an ascending or descending order. Subsequently, groups of k consecutive values assumed by the variable in question are replaced by their average. If the total number of elements n is not a multiple of k, the last group will contain more than k elements [89, 115]. As pointed out in [45], one method to substantially reduce the information loss with such a philosophy is to assign the additional elements to the group containing the modal (or median value [46]) value of the data (rather than including them in the extreme sets).

Two main approaches are used to sort one-dimensional data [31, 45, 89, 91, 115]:

1. *Single axis*
   The micro-data set being micro-aggregated consists of a single variable or a number of variables. It is called a uni-variate micro-aggregation because the ranking of the micro-data file is based on its projection along a single chosen axis. This approach is recommended when there is a high correlation between the variables, because any single variable will reflect the characteristics of the data vector [92, 115]. If a particular variable is chosen to rank all the micro-units, this variable must somehow reflect the size of the data vector [91]. The result usually depends on the variable chosen, because it typifies the underlying structure of the population, and, therefore, it could render any statistical inference to be ambiguous [115].

Since there is no guarantee that proximity on one variable implies proximity on others, a principle component has been used as a natural alternative to sort the data vector [31, 34, 89, 91]. Generally speaking, a principle component analysis reduces the dimensionality of the problem from P to usually 1, 2, or 3 dimensions, by taking into account all the variables of the initial set, and, simultaneously, preserving the underlying multi-dimensional correlation structure. Although it is true that such a scheme is sub-optimal, it can be seen that the resulting solution will not be far from the optimal one because it seeks to minimize the differences between the distances of the elements before and after the transformation process [115]. The rationale for this process is to preserve, as much as possible, the total variance of the original variables in the projected data [100].

Another alternative strategy, like the principle components analysis, takes all the variables into account, and is based on the sum of z-scores. Such a criterion gives equal importance to the variables by adding (across all variables) the standardized values for each observation [31, 34, 89, 91]. In other words, the rationale for this process is to store the records so as to take into account the variances of all the variables [100]. Finally, it is important to highlight that there is no guarantee that the axis chosen will provide an optimal micro-aggregation, since these methods are very sensitive to the set of variables chosen [115]. Therefore, a single axis MAT suffers from two sources of IL: The one-dimensional data sorting loss and the associated micro-aggregation loss [91, 92].

The authors of [100] have recently replaced the use of projection methods in micro-aggregation by the Sugeno integral aggregation function to calculate the projected axis. They have reported that using any aggregation function reduces the execution time, and, in some sense, increases the data protection.

2. *Individual Ranking*

   There are many cases where the variables are not highly correlated. This forces heterogeneous values to become members of the same group if any method involving a single axis is used [115]. In order to avoid the IL caused by sorting

the micro-data file onto a single axis, the authors of [31, 32] propose that various uni-dimensional variables should be aggregated separately. This approach, which is known as blurring [32, 33, 34], deals with multi-variate data sets by micro-aggregating the variables one variable at a time. Indeed, the variables are sequentially and independently micro-aggregated, and every aggregation involves a fixed size of contiguous values. Therefore, this method produces better outlying observations, while leaving the majority of the data structure intact [31, 33]. The authors of [10] give a quantitative theoretical measure, which estimates the loss of the variance due to individual ranking micro-aggregation methods. Their results show that the expected total variance loss decreases with respect to the increase in the size of the sample under three different distributions using the moments and spacing approaches [115]. Generally speaking, although individual ranking owes its popularity to its simplicity and to the fact that it leads to a low information loss [46], it also leads to a high disclosure risk[7]. It has a conceptual drawback, namely, that it does not partition the n data vectors in the micro-data set on a *data-vector* basis. Rather, the partition is obtained for each variable *in turn*, in the micro-data set [46, 89, 92].

The authors of [31, 46, 85] highlight an avenue by which the individual ranking MAT can be extended to permit the simultaneous micro-aggregation of several variables, so that a single partition for the entire data set is obtained. An alternative to the individual ranking scheme is the weighted moving average method that requires choosing a number of weights. For example, if we assume that a triple weight (a, b, c) has been chosen, this means that the observations are first ranked independently, and instead of replacing the individual units by the average value, each value for the ranked variables in the observation is replaced by a new computed value comprising of $a\%$ of the previous value, $b\%$ of the current corresponding value, and $c\%$ of the next value in the observation sequence [115]. While this method does not lead to any aggregate being repeated

[7]In individual sorting, an intruder knows that the real value of a variable in a data vector in the i^{th} group is between the average values of the $i - 1^{th}$ group and $i + 1^{th}$ group. Consequently, if these two averages are close to each other, the intruder can determine a narrow real-valued interval to enable him to compromise the security of the system [46, 89, 91].

k times in each group, it simultaneously suffers from the increment in the IL due to the two missing weighted moving averages at the beginning and at the end. This is because the replacement of these observations, using repetitions of existing values, may lead to values which are far from the original data set [115].

Hanani, a formal algorithm to find the optimal solution for the k-partition problem to minimize the IL, was proposed by Defays and Nanopoulos in [34]. The paper incorporated the concept of determining a suitable set of hyperplanes separating the n data vectors into a number of homogenous groups. It did this by enforcing the fixed size group and by deriving each partition from a simple ranking of units in terms of a uni-dimensional variable that groups contiguous units [115]. As pointed out in [34], the problem is quite complex, and its practical implementation is both difficult and complicated. For example, it is highly recommended to start the algorithm from different initial partitions in order to avoid falling into local minima [115].

Initial research in the field proposed "fixed" $MATs$, which required that the size of each partition group was a fixed constant, k. These, in turn, led to the *Fixed-Size Micro-aggregation* or the *Classical Micro-aggregation* [33, 34, 115] algorithm. Recent developments [45, 89, 91, 92, 126] have concentrated on further reducing the information loss by using variable-sized data-dependent groups, which leads to families of *Data-Oriented Micro-aggregation* algorithms. The philosophy utilized is that groups need not consist of exactly k data vectors, but of *at least* k data vectors. They also preserve the natural data aggregate by allowing the group size to be between k and $2k-1$, depending on the structure of the data, so as to lead to more homogenous groups and to minimize the ultimate information loss [28, 45, 92]. Although methods yielding variable-sized groups are marginally more complex than those involving a fixed-size micro-aggregation, they are less likely to compromise the *"privacy"* of the micro-data sets as shown in [85].

Figure 2.1 illustrates the difference between fixed-size and variable-size $MATs$,

where Figure(a) shows nine micro-data units of two variables. If a fixed-size micro-aggregation with $k = 3$ is used, as in Figure 2.1(b), a partition of three groups will be obtained, which seems to be unnatural to the distribution of the data. On the other hand, if a variable-size micro-aggregation is used, as in Figure 2.1(c), a partition of two homogeneous groups will be obtained. This definitely achieves a smaller IL compared to the fixed-size micro-aggregation [22, 46, 54, 89, 91, 92, 115].

Two alternative heuristic approaches that incorporate variable-size micro-aggregation have been presented in [43, 45, 91]:

- Micro-aggregation based on a genetic algorithm.
- Micro-aggregation based on the Ward's algorithm.

The authors of [45] presented a genetic algorithm that appears as an alternative linear heuristic. It presents the k-partitions as a binary string (the "chromosomes"), and combines directed and random search strategies to attain a global optimum [89, 91]. It starts from a random population of strings, evaluation, selection, crossover, and mutation operators are successively and iteratively applied to obtain a new superior generation that satisfies the convergence criterion. This method offers a good trade-off between the required speed and the information loss when the data set is very large or when the MAT has to be invoked online. But there does not appear to be a theoretical framework that allows optimal values to be found for the five input parameters: The probability of crossover, the probability of mutation, the population size, the maximum number of generations, and the convergence criterion used as a stopping iteration. Another weak point is that since this scheme is based on a heuristic random search, it provides solutions that are not reproducible. In other words, running a micro-aggregation based on a genetic algorithm twice on the same data set will probably yield two different micro-aggregated files [45, 91, 92]. Unfortunately, the technique presented in [45, 91, 92] is not easily adapted for the the multi-variate case. The main problem comes from the fact that a multi-dimensional space is only partially ordered. This renders the solution to the problem of properly representing

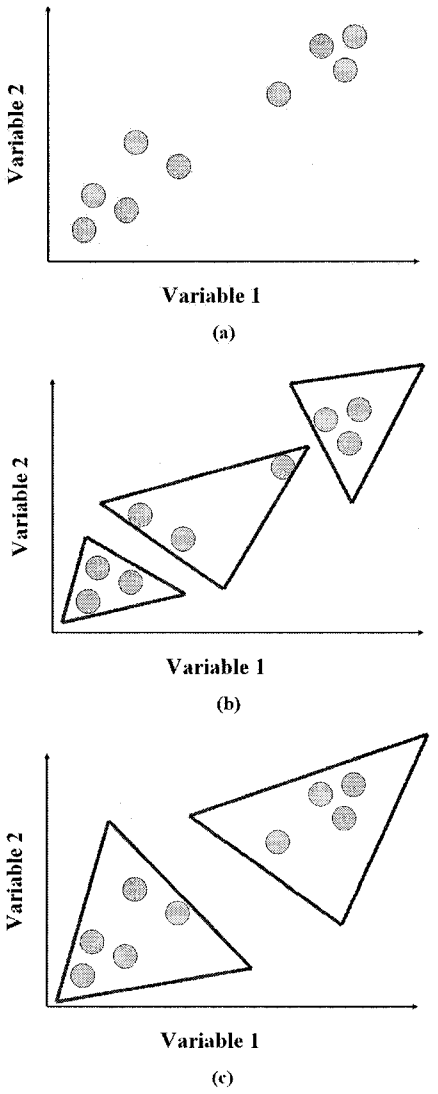


Figure 2.1: Fixed-size groups versus variable-sized groups.

multi-variate k-partitions as binary strings, to be far from obvious [45, 89].

A hierarchical classification method can be used to obtain building blocks for the heuristic MAT so as to yield variable-sized groups. Ward's method, proposed in [139], is attractive because it is stepwise optimal: Two groups or data elements coalesced at each step are chosen so that the increase in the within-groups index, the SSE, caused by their union is minimal. However, Ward's method had to be adapted into the so-called k-Ward's method to render it applicable for micro-aggregation [43, 91]. The standard method merely involved building up a grouping hierarchy, whereas the ultimate goal is a k-partition of the initial data set. As well, the authors of [43] showed that a k-partition algorithm can be naturally turned into a truly multi-variate MAT. The main drawback of this method is its storage and running time complexities. Extending this result, a *Secure-k Ward* scheme was proposed in [85] so as to enhance the individual's privacy. This was done by employing two phases, an intra-group optimization and an inter-group optimization, with the intention of minimizing the information loss and the overall mean deviation. Finally, it should also be mentioned that the scheme proposed in [85] is compatible with the previous k-Ward's method proposed in the literature.

The author of [121] showed that the width of the cluster is a suitable measure, especially for the highly skewed data. The total within-cluster width is measured by summing the gap between adjacent members of the clusters, which implies that the size of the cluster is controlled by a minimum number of adjacent gaps. Two direct methods based on the width of the cluster were proposed in [121, 122] to micro-aggregate uni-variate data. First of all, after the data is sorted, all the gaps between the sorted data are computed, as well as the midpoints of the gaps. After this, a binary sorting tree is constructed in which the gap midpoints serve as the internal nodes of the tree and the data values are the external leaves of the tree. Finally, both methods attempt to achieve the micro-aggregation. The first proposed method uses the modified *quick sort* algorithm, which operates by recursively partitioning its input, while the second method treats the micro-aggregation problem as an *optimization problem*, which requires an objective function to choose between various sets of

micro-aggregations and a model with constraints that are included so as to define these possible sets of micro-aggregation. It is important to note that the adjacency concept is simple, natural and easy to implement in the uni-variate data. However, for bi-variate or multi-variate data, the adjacency is natural, the simplicity, and ease of implementation are lost due to not having well-defined multi-variate sorting techniques. Standard techniques from computational geometry have been adapted to the *MAP* [122], but we consider this to be outside the scope of this Thesis.

The exact solution for the uni-variate *MAP* that appears in both the individual ranking and projected data approaches has recently been shown to be polynomially solvable as a shortest path problem [74]. Here, optimal partitions were shown to correspond to the shortest path on a graph. The strategy works as follows: First of all, the uni-variate data set is sorted in an ascending order. Thereafter, a graph is constructed in which each arc corresponds to a possible group that may be part of an optimal partition. Each arc is assigned a weight that is the error that would result if that group were to be a part of the ultimate partition. The micro-aggregated solution involves determining the set of arcs in the shortest paths of the graph. The complexity of the resulting algorithm is $O(n \log n)$, where n is the number of records in the database. In spite of having a polynomial algorithm for the optimal uni-variate *MAP* [74], heuristic algorithms are needed, because attaining the minimum information loss for the *MAP* is NP-hard for the multi-variate case [84, 103].

Finally, Table 2.3 summarizes the main numerical uni-variate micro-aggregation methods and their classification.

2.6.3.2 Multi-Variate MATs

A large number of heuristic methods have been proposed for solving the multi-variate *MAP*. These methods can be further summarized as follows [28, 46]:

1. *Projecting multi-variate data onto a single axis*

Table 2.3: The classification of the numerical uni-variate micro-aggregation methods.

Classification	Micro-Aggregation Methods
Single axis	Particular variable
	Principal component
	Sum z score
	Aggregate function
Individual methods	Individual ranking
	Weighted moving average
Heuristic partitioning technique	Hanani's algorithm -(variance)
	k Wards -(distance)
	Genetic algorithm
	Minimize the total within cluster width (Quick Sorting or optimization approach)
	Shortest path -(Optimal)

Instead of using a multi-dimensional distance to sort the data vectors, all practical methods perform a straightforward uni-dimensional sorting by projecting the former onto a single axis (*i.e.,* using the first principle component, the sum of z-scores or even a particular variable) [91, 92]. It is necessary to stress that all the above described numerical uni-variate $MATs$ can easily be extended to multi-variate $MATs$ using any projection method.

2. *Unprojected multi-variate data*

   A natural improvement of the above is to deal directly with multi-variate data without projection. Either the data vectors can be sorted based on comparing all the variables simultaneously or they can be independently sorted them using more than one (for example, two or three) variable at a time. A number of heuristic methods to micro-aggregate unprojected raw data utilize search techniques including hierarchical clustering, genetic algorithms, and tabu search [45, 92].

In spite of the simplicity of the uni-variate micro-aggregation and its low complexity, it is not very attractive because it either suffers from a high IL due to the data transformation (projection on to single axis) or from high DR, (especially, for

the individual ranking) [46, 45, 89]. The empirical results reported in most of the papers suggest that the multi-variate micro-aggregation on unprojected data offers a superior trade-off between the DR and IL, especially when groups of three or four variables are micro-aggregated at a time, rather than all variables at the same time [44, 45, 46].

The first algorithm to accomplish this was invented in 1998 by Domingo-Ferrer *et al.* [91, 45] and called the "Multi-variate fixed-size micro-aggregation" method. It micro-aggregates the multi-variate micro-data file based on the concept of the diameter distance of the data set as follows: It first searches for the two most distant records from the data set, say records r and s. Using these, two clusters are created; the first one comprises of r and its (k-1) nearest records, while the second comprises of s and its (k-1) nearest records. If there are at least $2k$ records that do not belong to the two clusters formed, the aggregated records are then computed for these two clusters, and they are removed from the original data set. This process is iteratively repeated until all the records in the original data set have been used in generating the new micro-aggregated file or, in other words, until the original data set is empty. But if there are between k and $2k-1$ records, which do not belong to the two clusters formed, we form a third cluster with those records and stop the process. In a case when we have less than k records that do not belong to the two formed clusters r and s, it is mandatory to add these remaining records to the closest cluster [45, 52, 91].

The problem encountered is one of determining how to decide which are the most distant records. A truly multi-variate criterion has to be followed instead of projecting the data onto a single axis[8]. The most distant records are defined according to a truly multi-variate sorting criterion, which is the distance matrix, and this is referred to as the *"Maximum-distance"* (MD) [45, 89, 91, 92, 115] scheme. One of the disadvantages of using this sorting criterion is related to the storage capacity, since it requires quadratic time (quadratic in the size of the data set) [89]. Solanas *et al.* have

[8]If a single axis is used to make a decision of the two most distant records in the multi-variate fixed-size micro-aggregation algorithm, then this algorithm is equivalent to performing a uni-variate fixed micro-aggregation.

recently shown that when the number of records is very large, the distance matrix can be efficiently stored by applying the blocking technique [127]. Another small disadvantage is that the grouping resulting from the MD usually depends on the extreme data vector with which one starts. In other words, the efficiency of the solution depends on which extreme data vector is taken as the first data vector and which is taken as the last [45, 89, 91, 92, 115]. This concern is clearly illustrated in Figure 2.2. If k is set to equal 3, Figure 2.2(a) represents a micro-data file consisting of 6 points with 2 variables. In this case it is obvious that the most distant records are a and e. Starting from vector a we will have two clusters, shown in Figure 2.2(b). But if we choose to start from vector e, we will have another different two clusters, as shown in Figure 2.2(c). Typically, the differences in the IL that result from choosing either extreme vectors as the first or last are small [45, 89, 91, 92, 115].

A natural way to obtain multi-variate data-oriented methods is to generalize some of the uni-variate data-oriented methods described in Section 2.6.3.1. The strongest advantage of a k-Ward MAT is that it is easily adapted to the case of multi-variate unprojected data. The reason for this is that the k-Ward method was originally designed as a multi-variate clustering algorithm [62, 79, 91, 139]. It is important to remind the reader that projecting the multi-variate data in order to know the first-k and the last-k elements in the k-Wards algorithm, (as will be described in greater detail in Chapter 3), is not equivalent to a data-oriented uni-variate micro-aggregation scheme on projected data [45, 89, 91, 92, 115].

In 2005, an enhanced version of the "multi-variate fixed-size micro-aggregation" method, called the Maximum Distance to Average Vector ($MDAV$) scheme, appeared in [60] and was implemented as a built-in technique in the μ-ARGUS Software tool version 4.0 [77]. This modification is based on the concepts used to compute the micro-aggregated record, namely, it depends on computing the centroid of the data set, instead of computing its diameter. After this, a quick search for the most distant record from the centroid, say r, is done. Subsequently, a new search for the most distant record from the record r, say s, is accomplished. The next step consists of creating two clusters, the first one comprising of r and *its* $k-1$ nearest records, while

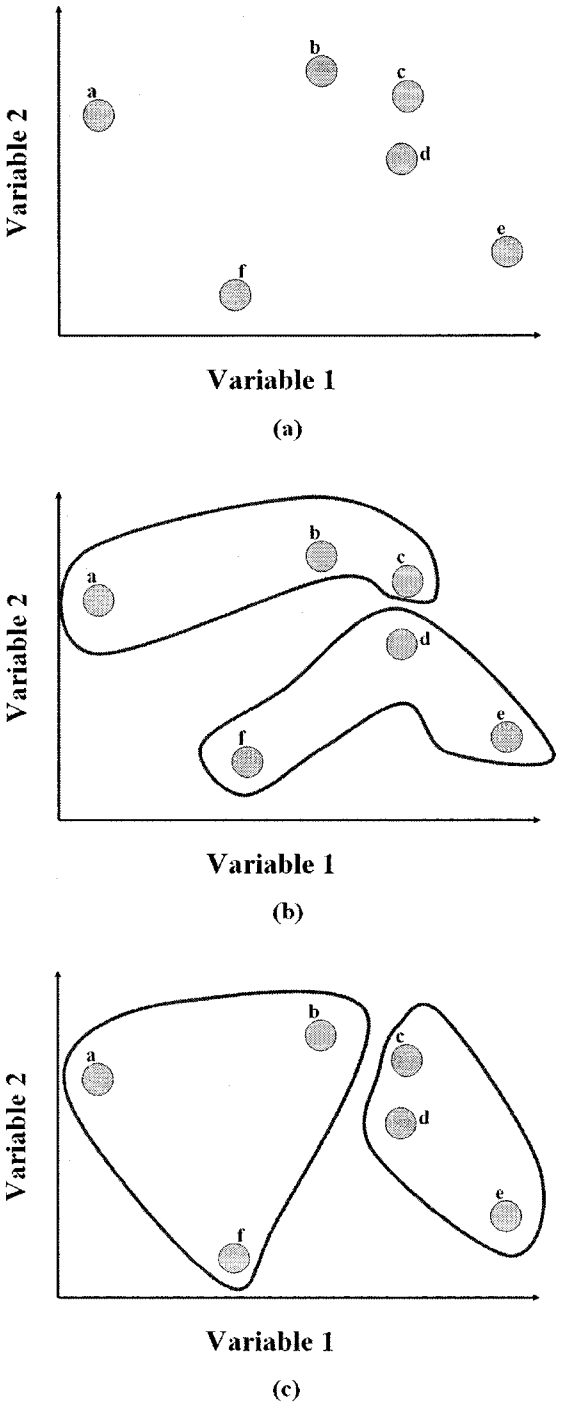


Figure 2.2: Grouping with the maximum distance criterion.

the second comprises of s with *its* nearest $k-1$ records. At the end of this stage, the two clusters are micro-aggregated and removed from the original data set. The latter steps are iteratively repeated until no more records remain in the original data set. The advantages of this new modified version of the $MDAV$ are the increased speed of the micro-aggregation and the resulting reduction of the IL [52].

$V-MDAV$, which stands for Variable-size Maximum Distance to Average Vector, is the first heuristic method applicable for a multi-variate variable-size micro-aggregation [126]. It improves the well-known $MDAV$ method in terms of lowering the SSE and, simultaneously, maintains an equivalent computational cost. $V-MDAV$ follows a strategy similar to the $MDAV$ method, by building a distance matrix and computing the global centroid. Thereafter, once the most distant record from the global centroid is found, a group of k records is formed by selecting the $k-1$ records closest to the initial one. The next step is the most important one, where the $V-MDAV$ overcomes the fixed-size constraint of the previous heuristics by adapting the data set distribution and generating variable-size groups. This step is achieved by searching for the closest unsigned record in the last formed cluster, e_{min}, and computing its distance, d_{in}. After this, we compute the minimum distance from the selected record e_{min} to any of the remaining unsigned records, d_{out}. A decision of the inclusion of e_{min} into the cluster is made if, and only if, $d_{in} < \gamma d_{out}$, where γ is a gain factor; otherwise e_{min} will not join the cluster. The extension step is terminated when either the group size is equal to $2k-1$ or when a decision of inclusion is not satisfied. The entire procedure is iteratively repeated until all the records in the original data set have been used in generating the new micro-aggregated file [52, 126].

The computation of the Minimum Spanning Tree (MST) leads to another multivariate MAT [84]. This yields a new clustering algorithm obtained by splitting the minimum spanning tree using a constraint on the minimum group size. The MST can be considered as a potential strategy for any practical application, since it exploits the natural clustering effects in the data that are consistent with the objective of reducing the IL.

Domingo-Ferrer *et al.* have recently reported a multi-variate micro-aggregation heuristic referred to as the "μ-Approximate" algorithm, which involves an approximation bound for the SSE [39, 51]. The SSE of the micro-aggregated data is no longer greater than a multiple of the minimum SSE_{opt}, which corresponds to the optimal multi-variate MAP [102, 103]. Bounds have been proven for two different optimality criteria: An $\bigcirc(k^2)$- approximation bound for minimum within-groups sums of Euclidean distances, and an $\bigcirc(k^3)$- approximation bound for minimum within-groups sums of squares [51]. Their philosophy involves creating a directed forest where all the original records are vertices. Each vertex has at most one outgoing edge, i.e., (u, v) is an edge between the two vertices v and u if v is one of the $k-1$ nearest neighbors of u. The size of every tree in the forest should be between k and $2k-1$. When the algorithm terminates, the vertices in each tree are replaced by their aggregation [39]. The authors of [51] also explain how the value of k can be determined.

As opposed to the above, the authors of [49] present a tighter approximation for the special case when $k = 2$, which is called the "Polynomial-time 2-approximation" for multi-variate micro-aggregation. Their solution represents an interesting option to implement 2-anonymity, because it results in low IL and thus high data utility in most scenarios except when internal intruders are likely, and when outliers are exist.

A more recent hybrid technique for multi-variate micro-aggregation combines a multi- variate heuristic yielding fixed-size groups and a genetic algorithm yielding variable-sized group [88]. It is well-known that fixed-size heuristics are fast and able to deal with large data sets. On the other hand, the genetic algorithm obtains very good results (*i.e.*, optimal or near-optimal), although it can only cope with very small data sets. Consequently, this new mixture blends the advantages of both types of heuristics and avoids their shortcoming. The main idea behind this hybrid algorithm is to partition the data set into a number of groups by invoking any fixed-size multi-variate micro-aggregation such as the $MDAV$, and then optimizing these partitions by means of the genetic algorithm [88].

An adaptive fuzzy c-means based micro-aggregation technique has recently been proposed to build a set of clusters in which elements can belong to more than one cluster at the same time [133]. In other words, the fuzzy c-mean leads to a set of non-disjoint clusters where each element has a partial membership index that expresses the degree of its membership to a particular cluster. Generally speaking, the membership degree is specified in the interval $[0, 1]$, where 0 implies "no membership" and 1 means a "full membership". The most interesting part of this algorithm involves replacing the individual values by the centroid value of each cluster. This is achieved by selecting a cluster to which certain records belong and by further selecting the values of different variables in the same records. There are, thus, four different alternative methods for the aggregation stage [57, 134]:

- *Selecting a cluster for a given record*
  The selection of a cluster is based on either a uniform probability distribution over the clusters with a non-zero membership or a probability distribution proportional to the membership degree.

- *Selecting the values for different variables in the same records*
  Once the cluster is selected, either all variables are aggregated using the same cluster centroid or a cluster selection technique is applied for each variable of the same record.

The authors of [57, 134] report that the fuzzy c-means method performs better than an additive noise, data distortion scheme that uses a probability distribution followed by resampling. However, it performs worse than the classical micro-aggregation, although with respect to the DR criterion, it is better. The reason behind this is that this method does not yield any clue about the masking method used nor about the clusters obtained by the method [57, 133]. Recently, an alternative fuzzy clustering method was proposed in [134], referred to as an entropy-based fuzzy c-means method. The only difference between the latter version and the previous one is the technique by which fuzziness is introduced. The fuzzy c-means strategy produces a fuzzy solution by involving the *means*, while the latter uses a term based on the entropy to force a

fuzzy solution. It appears as if an entropy-based clustering algorithm is significantly better than the standard fuzzy c-mean based clustering algorithm, additionally it is more stable [134].

It is important to remark that the exact solution to the MAP in the multi-variate case, (without projecting the data onto an "Euclidean" space of dimension of two or greater), is shown to be NP-hard [102, 103]. This opens an avenue of research, namely to produce an approximation polynomial time technique that can out-perform all the other techniques. This problem remains open.

We conclude this section by presenting in Table 2.4 a summary of the main numerical multi-variate micro-aggregation methods and their classification.

Table 2.4: The classification of the numerical multi-variate micro-aggregation methods.

Classification	Micro-Aggregation Methods
Projection onto single axis by using principle component, sum of z-scores, or even particular variable	Individual ranking
	Weighted moving average
	Hanani's algorithm
	Genetic algorithm
	k Wards
	Shortest path
Unprojected methods	Multi-variate fixed-size micro-aggregation
	k ward multi-variate micro-aggregation
	Minimum Spanning Tree (MST)
	Maximum Distance to Average Vector ($MDAV$)
	Variable-size Maximum Distance to Average Vector ($V-MDAV$)
	Fuzzy C-mean
	μ-Approximate Algorithm
	Hybrid Algorithm

2.6.4 Categorical MATs

Quantitative micro-aggregation has been a research topic for more than ten years, while qualitative micro-aggregation is quite new [54]. Although the actual implementation of micro-aggregation requires only a clustering method and an aggregation procedure, the reason behind this relative lag in the latter field is the difficulty of involving both these processes in the case of categorical data [60, 131]. A fair bit of attention has, however, recently been given to extending the MAT to categorical data types [56, 58, 115, 131].

The aggregation function for categorical data can be distinguished on the basis of the categorical data type. In other words, the main operator in this case is the plurality rule (mode, or the voting procedure) for the nominal scales, while for the ordinal scale the median is used for the aggregation step [22, 54, 56, 60].

The author of [33] was the first one to suggest a generic method based on the individual ranking technique for multi-variate data. The subsequent clustering method for categorical data appeared in [131], and this was, in turn, based on the k-modes algorithm inspired by the k-means algorithm for numerical data. More recently, the $MDAV$ scheme has also been extended for categorical data [60].

As we will not be discussing categorical data in this Thesis, our review of this sub-area terminates here.

2.6.5 Additional Applications

Apart from being methods in their own right that permit micro-aggregation to be used as an SDC technique to yield a good trade-off between the IL and DR, the MAT can also be used in other applications, such as in *linear regression* and *k-Anonymity.*

The authors of [66] enhanced the classical multiple linear regression model, which obtains the best linear unbiased estimates of the regression coefficients, to render it

more secure. This was achieved by deriving the estimator from the grouped aggregated data, although partial aggregation usually results in the loss of efficiency for estimates of the regression coefficients and their residual error variance. But this loss can be minimized by generating groups that are homogeneous with respect to the independent variables being analyzed. In other words, this can be done by minimizing the within-group variation and by maximizing the between-group variation.

Micro-aggregation has also been found to be an interesting alternative to generalization or suppression for k-anonymity[9] [5, 39, 60]. A solution to k-anonymity is based on generalization and suppression techniques. The latter perform poorly in terms of data quality because they are not suitable for all types of attributes [39]. The authors of [39, 60] report that using micro-aggregation for k-anonymity circumvents most of the problems of generalization/suppression techniques because micro-aggregation is a unified approach, unlike the dual method that combines generalization and suppression. Although the optimal MAP is NP-hard, many heuristic approaches lead to a solution that is close to the optimal one. Micro-aggregation does not complicate the data analysis by introducing new categories, or by suppressing certain data fields. Finally, micro-aggregation is also useful in protecting fields containing continuous data because it does not harm or destroy the underlying numerical semantics [60]. The author of [41] extends the use of micro-aggregation of all attributes for k-anonymity to implement the property of P-sensitive k-anonymity in a more unified and less disruptive way.

Micro-aggregation has also been reported to be relevant to the field of Artificial Intelligence (AI), and has been used in some applications that require increasing the knowledge of a system for decision making and domain representation [59]. It can also be used in data mining or to compress the data set while minimizes the IL [52].

[9] k-anonymity is a useful concept to solve the tension between the data utility and respondent privacy in individual data. An anonymized data set is said to satisfy k-anonymity for $k > 1$, if, for each combination of the values of a key identifier, at least k records exist in the data set sharing that combination [39, 60].

2.6.6 Properties and Characteristics of the MAT

2.6.6.1 Variants of the Methods

Micro-aggregation performs quite well with respect to the different existing criteria for IL and DR. Moreover, the authors of [53, 57, 131, 133, 134] show that it is the second-best method for numerical data, after the rank swapping method[10]. The authors of [94] show that the choice of parameter in the MAT has less influential than in rank swapping. Therefore, the behavior for micro-aggregation is more robust, even through medium and high values. It is important to mention that both MAT and rank swapping offer more protection to outliers than other SDC techniques. It is necessary to note that there are some important issues that should be taken into consideration when micro-aggregating social or business data. These issues are discussed below.

- Selecting a segmentation of the set of the variables to apply the micro-aggregation It is important to determine the number of variables, (the best combination that leads to the minimum value of the IL [33, 34, 90]), that is to be used in the micro-aggregation process. Moreover, the initial vector of variables can be segmented into a number of variables, multi-variate or uni-variate, which we called segments. [31, 32, 66, 115, 119, 120]. A set of P variables can be treated as a single multi-variate variable, resulting in a single grouping, or as P separate variables, resulting in p groupings for each. It is well-known that dealing with all the variables simultaneously minimizes the loss in the information gained. It is reported in [54] that multi-variate micro-aggregation on unprojected data taking two or three variables at a time (rather than all variables) is the micro-aggregation algorithm offering the best trade-off between IL and DR.

- Selecting the number of steps to be applied in the micro-aggregation

[10]Authors of [32, 33] show that some people reject data swapping on the basis of protecting individual data without preserving the moments of the initial distribution.

It is important to determine the number of steps that are required to micro-aggregate all micro-units in the original file [66, 115]. For example, let us assume that 7 variables are used to micro-aggregate a micro-data file with 13 variables. The micro-aggregation process can be done either in a single step, by using all the variables together, in two steps, by using 3 variables in the first step and 4 variables in the second step, or in three steps, by using 3 variables in the first step and 2 variables in the second and the third steps. The authors of [119, 120] show that the micro-aggregation in more than one step leads to a lower IL, and, simultaneously, a higher DR than the usual single-step micro-aggregation technique.

- Selecting the type of the variable to be applied in the micro-aggregation
  It is usually recommended that independent variables are chosen to micro-aggregate the micro-units, and it is preferable that the variable chosen is either discrete, real, or ordinal [32, 33, 66]. These types of variables do not take values over a very restricted set, and so they will minimize the probability of being compromised by intruders [14]. Generally speaking, the decision of determining which variables have to be involved in the micro-aggregation process is based on the previous knowledge about the characteristics of each variable [119, 120].

- Selecting the degree of aggregation
  The number of records per group, which is usually represented by k, a pre-defined threshold, plays an important role in determining the loss in the data utility, as well as the loss in the confidentiality. In fact, increasing the degree of the aggregation will decrease the IL and, simultaneously increase DR [33, 66]. The minimum size of such groups usually depend on several factors: The rules adapted in some countries, the degree of confidentiality of a segment of the variable, and the values taken by the variables in a group [32]. It is worth mentioning that the decision of imposing a fixed size or a minimal size constraint has a clear effect on the homogeneity of the groups. Generally speaking, a minimal size constraint usually leads to more homogeneous groups than the fixed size, and it is more difficult to be compromised [28, 45, 85, 92, 115]. In practice,

the statistical law that defines the requirements in order to make the data set confidential, usually sets the value of k to be either 3 or 4 [46, 115]. If k is greater than 4, it can be argued that basic analytic properties, such as these required for regression, can be seriously affected [140].

- Selecting the clustering procedure
  The methods are mainly characterized by the way in which the homogeneity of the groups is measured [11, 32, 33]. The main difference between the *MATs* lies in the way clusters are built (by modifications of a standard technique, novel approaches using genetic algorithms with an appropriate fitness function, etc.) [131]. The authors of [66] have suggested having a built-in function of the researcher's cardinal preference with respect to the grouping criterion. It is worth mentioning that applying the grouping criterion on the projected data will increase IL caused by the micro-aggregation itself [11, 45].

- Selecting the aggregation statistical procedure
  The *MATs* aim to partition the whole micro-units into groups of either k, or at least k micro-units group size. These micro-units are replaced by certain aggregation statistics, which are either the mean of the group or another measure of central tendency (*i.e.* median, mode, weighted average) [31, 32, 115]. This usually depends on the type of the variables whether it is quantitative or qualitative [131].

- Selecting the sorting procedures
  The combination of fixed-size *MATs* with sorting algorithms could reveal some confidential information, although the sorting methods have to set up groups that are as homogeneous as possible in order not to lose too much information. Special attention should be paid to ranking the individuals in either an ascending or descending order [115]. This is especially important if some information about an individual is available through a non-confidential source (prior knowledge), because it is easy to gain added confidential information by using successive sorting instructions [66]. A general rule applicable for the sorting strategy used for single axis methods typically depends on the *position* of the

majority of the outliers [115]. If the outliers of the variables are to the left of the distribution curve, an ascending order has to be used in ranking the values of the variable. But if they lie to the right side of the distribution curve, a descending order is used to rank the values of the variables.

- Measuring the homogeneity of a group
  Since the SST is fixed for a given data set, one should attempt to find a grouping that minimizes the SSE [31, 32, 33, 91]. Moreover, the concept of the underlying similarity in the definition of the groups can lead to various interesting variants of the basic methodology. The quantification of homogeneity in the multivariate data set is not easily defined, because it can be measured in various ways, for example, using the within-group variance, the entropy, or a measure based on any type of distance [31]. For the quantitative variables, the Euclidean distance or the variance will determine the formation of groups, while for the ordinal variable this information will be determined by the absolute value of the difference in rank or by the entropy concept [32].

At this juncture, we mention some of the drawbacks of $MATs$, which can be listed as follows:

1. The information present in the groups obtained by an MAT may sometimes obscure useful statistical information found in the database.
2. Adding or deleting certain records from different groups is expensive.
3. The bias introduced by using this technique decreases the variance with an amount equal to $\frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{k_j}\delta_{ij}^2$, where m represents the number of groups, k_j represents the number of records in groups j, δ_{ij} is the distance between the centroid of the j^{th} group and the i^{th} record which belongs to that group, and N represents the number of records in the micro-data file [104].

Finally, we conclude that there is no single method that out-performs the others [81], since there are no descriptive criteria defined to judge which is the best method

with respect to the complementary objective (IL and DR), the accuracy, and the execution time. Moreover, the choice of an optimal method usually depends on the nature of the data (*i.e.,* the type of the variables and the statistical relation between them) [11].

2.6.6.2 Inference Control in Data Mining Versus *MATs*

While *MATs* introduce intentional distortion to protect micro-data sets against disclosure, *AI* techniques try to overcome the data distortion. Generally speaking, the goals of *AI* and *MAT* seem to be contradictory. This is because while *MATs* are used to hide information, *AI* techniques are used to infer information. The power of any *AI* technique is in its ability to reconstruct the original data from several distorted versions [137]. This can be achieved either by [59]:

- Building a model for a given variable so as to discover a relationship between a sensitive attribute and others. Thus, if the attribute is disclosure protected, the aim is to infer the original value.
- Releasing multiple protected versions of the same original data file will definitely increase the disclosure risk. In this case the users will try to reconstruct and find the original information out of n different distorted versions of the same file [137], either by using re-identification procedures based on sharing a set of common variables or by processing non-common variables.

It should be mentioned that *AI* techniques can be beneficial if the data mining tools are combined with the micro-aggregation software packages. The protection can be achieved in two stages: First of all, the original data is protected by using one of the *SDC* techniques offered by the package. Secondly, the data mining tool will be used on both the original and the protected data sets to find models that lead to the original sensitive data. If the data mining scheme yields a good approximation, the data protector should return to the first stage and be enhanced.

Privacy preserving data mining is a related field that possesses similar goals [131]. It involves a set of two or more parties that want to compute joint shared functions on their private inputs. These computations should be done in a secure way, so as to preserve both correctness and privacy [81]. From this perspective, it can be seen that while the MAP is oriented to statistical databases, privacy preserving data mining is oriented to company proprietary information [131]. However, there seems to be large rewards in combining both of these strategies, and this is also another open area for research.

2.6.6.3 Clustering Techniques Versus $MATs$

The clustering problem and the MAP are similar when it concerns finding an optimal partition of elements for homogeneous groups. But, the clustering problem differs from the classical MAP in two aspects: First of all, in a clustering problem, the number of groups is given *a priori*, whereas for micro-aggregation it is not. Secondly, and more importantly, in a clustering problem there are no constraints on the cardinalities of the groups, whereas in the MAP, each group must contain no fewer than k elements [45, 28, 91, 84, 121]. Many data clustering algorithms have been proposed in literature [15, 79, 130, 138]. Clustering algorithms may be good candidates for MAT if they can be adapted by efficiently enforcing the group size constraint [84].

Chapter 3

Enhancing the Heuristic k-Ward MAT

3.1 Introduction

We consider the problem of securing a statistical database by utilizing the well-known micro-aggregation strategy, and in particular, the k-Ward strategy introduced in [139] and utilized in [45]. The latter scheme, which demonstrates a good trade-off between IL, and DR, coalesces the sorted data attribute values into groups, and on being queried, reports the means of the corresponding groups. In this chapter[1] We demonstrate that such a scheme, as reported in the literature, can be *significantly* improved on two fronts. First of all, we minimize the computations done in evaluating the between-class *distance* matrix, to require only a constant number of updating distance computations. Secondly, and more importantly, we propose that the data set be partitioned recursively before a k-Ward strategy is invoked, and that the latter

[1]A preliminary version of some of the results from this chapter appear in the *Proceedings of ACISP'06, the Eleventh Australasian Conference on Information Security and Privacy*, in Melbourne, Australia, in July 2006 [65]. The journal version of these results is currently under review.

be invoked on the "primitive" sub-groups which terminate the recursion. Our experimental results, done on uni-variate and projected multi-variate data using two benchmark data sets, demonstrate a marked improvement. While the IL is comparable to the k-Ward micro-aggregation technique proposed by Domingo-Ferrer *et.al.* [45], the computations required to achieve this loss is a fraction of the computations required in the latter - providing a computational advantage which sometimes exceeds 80% if one method is used by itself, and more than 90% if both enhancements are invoked simultaneously.

This chapter is organized as follows: The Ward algorithm and k-Ward_MAT are described comprehensively in Section 3.2 and Section 3.3, respectively. Section 3.4 presents the improved versions of the k-Ward MAT, which is followed by an algorithmic description of the two new modifications. Experimental results using the two usual reference benchmark data sets and the related discussions are provided in Section 3.5. Finally, Section 3.6 presents the overall conclusions of this work.

3.1.1 Contribution of the Chapter

As mentioned earlier the challenge in micro-aggregation is to modify the original data so that both the risk of disclosing confidential information and the loss of the data utility should be below a certain threshold determined by the data protector [42, 142]. This means that the data protector is the one who determines the MAT to be used based on the type of the data and its application [142].

The advantage of the shortest-path technique is that it leads to the optimal solution for uni-variate $MATs$ from the perspective of the IL [74]. However, unfortunately, it is often not the optimal solution from the perspective of the DR – since the IL and DR criteria are often conflicting. This is because the shortest-path method provides similar statistics to the original data, which, in turn, increases the likelihood of information being disclosed. The literature [39, 60] reports that this problem can be overcome by solely determining the appropriate value of the minimum group size

parameter. This is achieved by ensuring k-anonymity, where each data subject is distinguishable from at least $k-1$ other data subjects. But the reader should observe that approximate heuristic uni-variate $MATs$ which use a smaller value of k usually provide more protection, in spite of forfeiting some information when compared to the optimal solution. This motivates the present research which seeks for a "fast" scheme which can be effective for both the IL and the DR criteria.

The k-Ward is one of the heuristic $MATs$, which demonstrates a good trade-off between the IL and the DR, but it suffers from a fairly high computational complexity, as will be explained presently. The main contribution of this chapter is to render the k-Ward MAT a pragmatically applicable solution, and this is done by increasing the speed of the micro-aggregation process by incorporating two considerations, namely that of invoking recursive computations, (which is crucial for large-sized data sets), and by utilizing only critical values of the so-called distance matrix. This, in turn, is achieved by two enhancements to the k-Ward algorithm, namely those which we have called Recursive k-Ward (kWR), and the k-Ward Diagonal (kWD), respectively. This chapter demonstrates the power of these modifications in increasing the speed of the strategy while almost preserving the IL.

It should be mentioned that the strategy of using our recursive method can actually be applied to *any* known MAT. As a proof of concept, we initiate this research by demonstrating its power on uni-variate data. However, the power of such a method can be appeared also for multi-variate data sets by projecting them onto a specific axis in a transformed domain using, for example, a Principle Component Analysis, (PCA), the sum of Z-Scores, (ZS), or a particular variable[2]. To be objective and unbiased, though, it is fair to point out that these mechanisms cannot be applied without projecting the multi-variate data set onto a single axis. This is because there is no straightforward formal algorithm to *sort* multi-variate data without performing an appropriate projection [45, 46, 89]. Furthermore, these schemes critically depend

[2]This has also already been done for both the PCA and ZS projection methods and will be reported presently.

on the fact that the data points have to be ordered[3].

The advantage of these improved versions will be marked especially for "large" data sets. Since the distance matrix involving the set of data vectors cannot be computed and stored *a priori*[4], these distances and the resulting matrix must be computed "on demand"- *i.e.,* as and when they are needed. This, of course, demands a considerable computational overhead [91], which is significantly reduced by using the improved versions of the k-Ward's method.

The combination of invoking a recursive computation and requiring a small subset of the distance computations is novel - we are not aware of any comparable results. However, the remarkable reduction in computation (sometimes as much as 80%) renders the contribution of the chapter significant.

3.2 Ward's Method

The Ward method is an agglomerative clustering method that creates hierarchical groups of mutually exclusive partitions. Each partition attempts to attract members that are maximally similar with respect to certain specified characteristics. Given n vectors, the first partition consists of $m = n$ single-record groups. It then permits their reduction to $m - 1$ disjoint partitions by considering the union of all possible $m(m-1)/2$ pairs and selecting a union having a maximal value for the functional relation that reflects the criteria chosen by the user. This process can be repeated until the last partition consists of a set of groups with an acceptable criterion value

[3]The authors of [45] have presented a "multi-variate k-Ward" heuristic that does not rely on such a uni-variate projection. With little work, we believe that our recursive strategy can be rendered applicable to such a methodology. Although we are currently considering how our matrix optimizing operations can be applied to the scheme proposed in [45], it does not seem to be straightforward. This work is still open.

[4]Solanas *et al.* have recently proposed a strategy by which the distance matrix can be effectively stored when the number of records is very large, by applying the so-called blocking technique [127]. We are currently investigating how our methods can be further enhanced by incorporating such blocking techniques.

[139]. The Ward method is formally described below.

Algorithm 1 Ward Method [139]

Input: n data vectors, and a stopping criterion specified either in terms of the required number of groups or the minimum inter-group distances.
Output: The required partitions.
Method:

1: Create n groups each involving a single vector, and set this partition to be $\{G_1, G_2, \ldots, G_m\}$ with $m = n$.
2: Compute all the distance values above the main (zero) diagonal of the distance matrix.
3: **repeat**
4: Find the nearest pair of distinct groups, G_i and G_j. Merge them into a single group.
5: Decrement the number of groups by unity.
6: Update the distance matrix after the merge process.
7: **until** criterion is satisfied
8: **End Algorithm Wards Method**

A crucial component of the Ward algorithm is the computation of the so-called distance matrix involving all the groups. Essentially, the latter matrix is a symmetric $m \times m$ matrix, where m is the number of groups at any particular step. The intention is that by computing this matrix we will be able to coalesce two groups, G_i and G_j, resulting in a partitioning with a smaller number of groups, without forfeiting the overall similarity index excessively. Unfortunately, the k-Ward algorithm, as it has been used and reported, computes the entire matrix involving *all* the inter-group distances, which renders the computation excessive. Our goal is to minimize these computations by utilizing the ordered nature of the uni-variate data when data is, in itself, uni-dimensional, and in the projected case when the data is multi-variate, as will be explained presently.

3.3 k-Ward Micro-Aggregation Technique

Ward's agglomerative hierarchical clustering method, described above [15, 139], has been modified to provide an optimal k-partition solution by enforcing the group-size constraint [43, 45, 91]. Such a k-Ward_MAT can be applicable for quantitative data, and for qualitative data when an appropriate distance is defined. The authors of [45] proved that the k-Ward's algorithm terminates after a finite number of steps, and that the computational complexity is quadratic (i.e., $O(n^2)$). The following definitions and results are needed to understand the concept of an optimal k-partition; the proofs of the respective assertions can be found in [45, 91, 92].

Proposition 1. *An optimal solution to the k-partition problem of a data set exists such that each of its groups has a size greater than or equal to k and less than 2k [45].*

Definition 1. *For a given data set, a k-partition P is said to be finer than another k-partition P′ if every group in P is contained by a group in P′ [45, 91].*

Definition 2. *For a given data set, a k-partition P is said to be minimal with respect to the relationship "finer than" if there is no k-partition $P' \neq P$ such that P′ is finer than P [45, 91].*

Proposition 2. *For a given data set, a k-partition P is minimal with respect to the relationship "finer than" if, and only if, it consists of groups with sizes $\geq k$ and $< 2k$ [45].*

Based on the Ward algorithm and the above propositions, the authors of [45, 91] have developed the k-Ward_MAT algorithm for micro-aggregation. The essential qualifiers for this are the facts that the distance criterion is the SSE and that each group should contain between k and $2k - 1$ elements. The k-Ward_MAT algorithm is given below.

The formal proof of the convergence of the k-Ward_MAT algorithm can be found in [45]. The experimental results related to k-Ward_MAT, found in [45], are included in Section 3.5 of this chapter.

Algorithm 2 k-Ward_MAT [45]

Input: A set of *sorted* data records
Output: The groups of micro-aggregated records
Method:

1: Form a group from the first (smallest) k elements of the data set and another group with the last (largest) k elements of the data set. Initialize the intermediate elements so as to constitute single-element groups.
2: Use Ward's method until all elements in the data set belong to a group containing k or more data elements. In the process of forming groups by Ward's method, the criterion used is the SSE, and it involves computing the entire inter-group distance matrix. Also, in the process, never merge two groups *both* of which have a size greater than or equal to k.
3: **for** each group in the final partition that contains $2k$ or more data elements **do**
4: Apply this algorithm again to particular group containing $2k$ or more elements.
5: **end for**
6: **return** the set of groups and report the mean of the group (on being queried).
7: **End Algorithm** k-**Ward**_MAT

3.4 Optimized k-Ward Micro-aggregation Technique

The authors of [45] claim that the performance of MAT depends on the distribution of the variables. However, it has been observed that no single method outperforms all other methods for all variables involving real-life data. This conclusion was made after testing all the competing methods for each variable and using the method which yielded the lowest IL[5]. In this chapter, we restrict our study to the k-Ward scheme, because it demonstrates a good trade-off between the IL and the DR. In addition to this, we also propose an extension of the scheme to multi-variate micro-aggregation after a projection-based preprocessing phase. This has already been achieved, and the results will be included in a later section. In this regard, we observe that there

[5]It is well known that minimizing the two criteria, namely the IL and the DR can have contradictory implications. As researchers, we encounter a real paradox here. Such a study must assess how much information must be hidden to preserve privacy while, at the same time, how much must be retained to make the data usable. Such a discussion is, unfortunately, a study in its own right, and it is not included in this chapter. However, preliminary discussions can be found in [27, 102, 124, 120].

is much room for improvement when it concerns the speed of micro-aggregation. Indeed, the current k-Ward's method suffers from two major disadvantages: First, that of the excessive computational burden encountered by processing *all* the data elements and, additionally, that of computing the "entire" distance matrix that contains the distances between every single pair of groups. Therefore, we propose two modifications by which we can enhance this method and reduce the required time needed to micro-aggregate the data set. These optimizations also lead to enhancements in the multi-variate case.

3.4.1 Optimizing Distance-Based Computations

The first modification we achieve is in the Ward method itself [139] and, more specifically, in the phase that computes the distance matrix. The latter is an $m \times m$ matrix, where m represents the number of groups included at each step. It contains the distance values between the groups (recorded as D_{ij}) that, in turn, represents the SSE obtained by potentially merging G_i with G_j. This matrix is symmetric and has a zero diagonal.

The k-Ward's method requires a number of basic steps to generate a near optimal k-partition, where the best-case number of steps is $n(1 - 1/k)$, and the worst-case number is $(n/k - 1)(n/2 + k - 2)$ [45]. At each basic step, the number of groups is reduced by unity. Currently, there are two different approaches to compute the value of the distance matrix, namely using either a stored matrix approach, or invoking a stored data approach. In the k-Ward method, which uses a stored matrix approach[6], $(m^2 - m)/2$ values are computed at the initialization step, and $m - 1$ values are recomputed at each basic step. But using our newly introduced enhanced version of k-Ward_MAT, the so-called k-Ward Diagonal, kWD, only $m - 1$ values are computed in the initial step and at most 2 values are recomputed during a basic step. On the other hand, using a stored data approach[7] for a k-Ward scheme requires no

[6]The distance matrix is computed, stored, and retrieved from storage as needed.

[7]The distance matrix is computed when needed rather than retrieved from storage.

initialization step, and it computes $(m^2 - m)/2$ values at each basic step. As opposed to this, kWD computes only $m - 1$ values, which lie on the diagonal above the main zero diagonal.

In principle, kWD behaves just like k-Ward_MAT, but the primary difference involves the way the Ward algorithm is invoked. While k-Ward_MAT invokes the Ward algorithm as in Section 3.2, kWD avoids computing *all* the values above the main diagonal in order to find the nearest pair of distinct groups. Rather, it computes only the SSE values that lie on the principal diagonal above the main zero diagonal. Consequently, the only change required lies in Step 2 of the Ward algorithm, as shown below.

Algorithm 3 kWD

Input: A set of *sorted* data records
Output: The groups of micro-aggregated records
Method:

1: Same as in k-Ward_MAT.
2: Same as in k-Ward_MAT except that when we invoke Ward's method, rather than computing all the values above the main diagonal we compute only the values that lie on the diagonal *above* the main diagonal. But, when the value represents the operation of merging two groups both of which have a size greater than or equal k, it would further require a search for the nearest group of size less than k so as to effectively compute the minimum distance.
3: Same as Step 3-5 in k-Ward_MAT.
4: **return** the set of groups and report the mean of the group (on being queried).
5: **End Algorithm** kWD

The rationale for optimizing the computation, as in the kWD, is based on the mathematical results given presently. Figure 3.1 visualizes the computation of the distance matrix using both the k-Ward and the kWD schemes.

Since we are grouping an ordered data set, we present a result that states that the distance value of merging G_i with G_{i+j} is less than the distance value of merging G_i with G_{i+k}, when $k > j$ with a very high probability. This is shown in Theorem 1, which needs the result of Lemma 1 below.

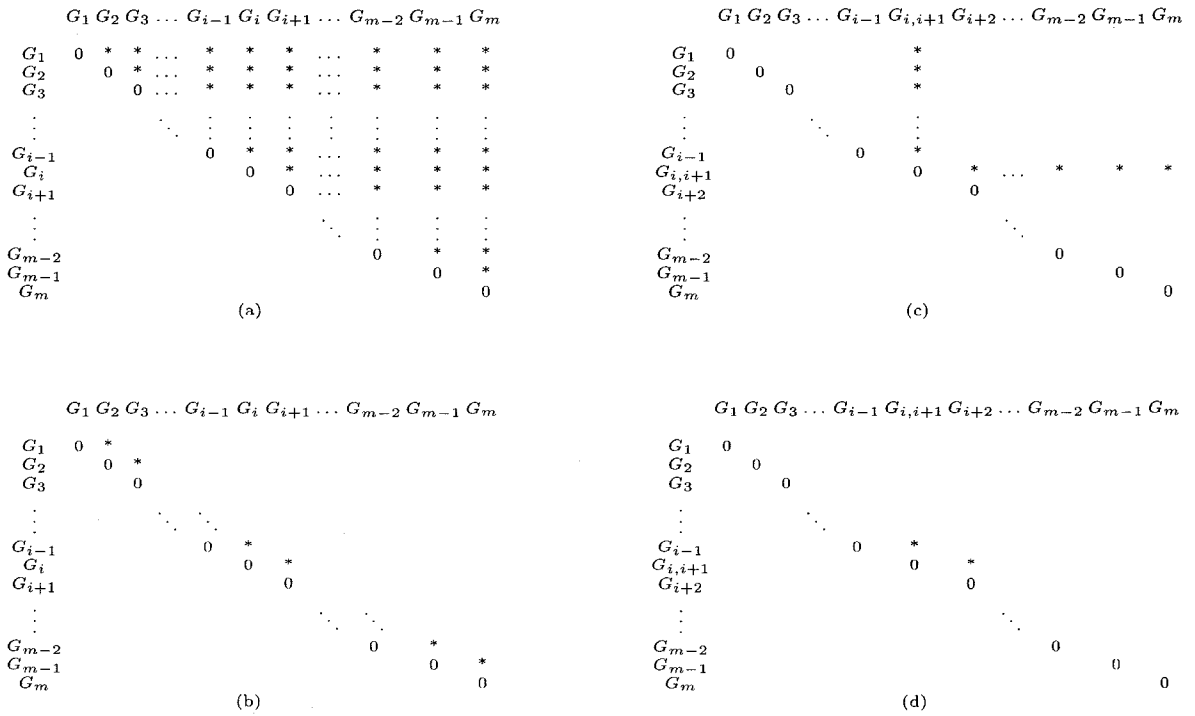


Figure 3.1: Comparison between k-Ward_MAT and the Optimized version kWD in computing the distance matrix using the stored matrix approach. In the initialization step, (a) k-Ward_MAT computes all values that lie above the diagonal, while (b) kWD computes only the diagonal which is above the main diagonal. At each basic step, (while merging groups G_i with G_{i+1}, (c) k-Ward_MAT computes all values that lie on the column corresponding to $G_{i,i+1}$ and the row $G_{i,i+1}$ above the diagonal, while (d) kWD computes at most two values ($G_{i,i+1}$,G_{i-1}) and (G_{i+2},$G_{i,i+1}$).

Lemma 1. *Let $\mathbb{P}_k^* = \{G_1, G_2, G_3, \ldots, G_m\}$ be an optimal partition. Then each partition $G_i = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ satisfies this property $x_{in_i} \leq x_{j1}$, where $i < j$, $1 \leq i < m$ and $1 < j \leq m$.*

Proof.

Let us assume that $x_{in_i} > x_{j1}$ in $\mathbb{P}_k^*$ the optimal partition. Observe that having the optimal partition implies a minimum value of the IL, which is directly proportional to the SSE. If $x_{in_i} > x_{j1}$ for any $j > i$, the SSE can be shown to have not attained the minimum value. This is due to the fact that if $\sum_{l=1}^{n_i}(x_{il} - \bar{x}_i)^2$ is the

contribution for G_i and $\sum_{l=1}^{n_j}(x_{jl}-\bar{x}_j)^2$ is the contribution for G_j, these quantities are not minimized because $x_{j1}-\bar{x}_i < x_{j1}-\bar{x}_j$ as per our assumptions. Consequently, to further minimize the SSE, x_{j1} should be partitioned into G_i and not into G_j. Thus, the IL obtained if $x_{in_i} > x_{j1}$ is not the minimum value, which means that $\mathbb{P}_k^*$ is not an optimal partition. Hence the result. □

Theorem 1. *Consider the quantity $D(G_i, G_j)$, defined as $\frac{n_i n_j}{n_i+n_j}(\bar{x}_i - \bar{x}_j)^2$. Then, if the index $k > j$, and the size of the groups G_{i+k} and G_{i+j} satisfy $n_{i+k} \leq n$ and $n_{i+j} \leq n$ respectively, then*

$$\begin{cases} D(G_i, G_{i+k}) \geq D(G_i, G_{i+j}), & \textit{Whenever } n_{i+k} \geq n_{i+j} \\ D(G_i, G_{i+k}) \geq D(G_i, G_{i+j}), & \textit{Whenever } n_{i+k} < n_{i+j} \textit{ and if the groups} \\ & \quad G_{i+j} \textit{ and } G_{i+k} \textit{ have different elements} \\ D(G_i, G_{i+k}) < D(G_i, G_{i+j}), & \textit{Otherwise.} \end{cases}$$

Proof.

First of all, observe that by definition, $\forall G_i$, $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$

Since all the elements in the data set are ranked in an ascending order, it implies that we have the ordering: $\bar{x}_i \leq \bar{x}_{i+j} \leq \bar{x}_{i+k}$, where $i = 1, 2, \ldots$, with $i+j$ and $i+k \leq n$. The distance between groups G_i and G_j is computed as $D(G_i, G_j) = \frac{n_i n_j}{n_i+n_j}(\bar{x}_i - \bar{x}_j)^2$ [45].

The latter is nothing more than the squared difference between the means of the groups weighted by a scalar factor, namely, $\frac{n_i n_j}{n_i+n_j}$. For $i = 1, 2, \ldots$, with $i+j \leq n$ and $i+k \leq n$, we now consider the various cases when (3.1) is true, knowing that (3.2) is always true.

$$(\bar{x}_i - \bar{x}_{i+j})^2 \leq (\bar{x}_i - \bar{x}_{i+k})^2 \tag{3.1}$$

$$\bar{x}_i \leq \bar{x}_{i+j} \leq \bar{x}_{i+k} \tag{3.2}$$

which is equivalent to determining when (3.3) holds:

$$\frac{n_i\ n_{i+j}}{n_i\ +\ n_{i+j}}\ (\bar{x}_i\ -\ \bar{x}_{i+j})^2\ \ \leq\ \ \frac{n_i\ n_{i+k}}{n_i\ +\ n_{i+k}}\ (\bar{x}_i\ -\ \bar{x}_{i+k})^2. \tag{3.3}$$

We consider the three pertinent mutually exclusive and exhaustive cases.

- Case $(n_{i+j} = n_{i+k})$

  If $n_{i+j}\ =\ n_{i+k}$, the multiplying terms $\frac{n_i n_{i+j}}{n_i+n_{i+j}}\ =\ \frac{n_i n_{i+k}}{n_i+n_{i+k}}$. Since the $\bar{x}_i$s are sorted by virtue of (3.2), a simple algebraic manipulation shows that the RHS of (3.3) is greater than its LHS, and (3.3) follows.

- Case $(n_{i+j} < n_{i+k})$

  The difference between n_{i+k} and n_{i+j} is bounded by $[1, k-1]$ based on Proposition 1 in [45], and so, in the worst case, the difference cannot exceed $k-1$. Thus (3.3) implies:

  $\frac{n_i\ n_{i+j}}{n_i\ +\ n_{i+j}}\ (\bar{x}_i - \bar{x}_{i+j})^2\ \ \leq\ \ \frac{n_i\ n_{i+k}}{n_i\ +\ n_{i+k}}\ (\bar{x}_i - \bar{x}_{i+k})^2$

  $\Rightarrow \frac{n_i\ n_{i+j}}{n_i\ +\ n_{i+j}}\ (\bar{x}_i - \bar{x}_{i+j})^2\ \ \leq\ \ \frac{n_i\ (n_{i+j}+k-1)}{n_i\ +\ (n_{i+j}+k-1)}\ (\bar{x}_i - \bar{x}_{i+k})^2.$

  But $\frac{(\bar{x}_i-\bar{x}_{i+k})^2}{(\bar{x}_i-\bar{x}_{i+j})^2}\ \ \geq\ \ 1$, and

  $\frac{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}}{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}\ +\ (k-1)n_i}\ \ <\ \ 1 \qquad\qquad \forall k > 1$

  Thus $\frac{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}}{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}\ +\ (k-1)n_i}\ \ <\ \ \frac{(\bar{x}_i\ -\ \bar{x}_{i+k})^2}{(\bar{x}_i\ -\ \bar{x}_{i+j})^2}$, and so (3.3) follows.

- Case $(n_{i+j} > n_{i+k})$

  As in the above case, we know that $\frac{(\bar{x}_i\ -\ \bar{x}_{i+k})^2}{(\bar{x}_i\ -\ \bar{x}_{i+j})^2}\ \ \geq\ \ 1$. Again, the maximum difference between n_{i+j} and n_{i+k} is $k-1$. Since we are grouping data which is ordered, we have:

  $\frac{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}\ +\ (k-1)n_i}{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}}\ \ >\ \ 1$

  The value of $\frac{(\bar{x}_i\ -\ \bar{x}_{i+k})^2}{(\bar{x}_i\ -\ \bar{x}_{i+j})^2}$ is almost always greater than $\frac{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}\ +\ (k-1)n_i}{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}}$ except in the single case when $\bar{x}_{i+j}\ =\ \bar{x}_{i+k}$ as shown in Lemma 1. This case occurs only when both G_{i+j} and G_{i+k} have elements with a single repetitive entry (with different frequencies). More precisely G_{i+j} and G_{i+k} contain the same quantity, v, with group sizes equal to n_{i+j} and n_{i+k}, respectively. Thus $\bar{x}_{i+j}\ =\ \bar{x}_{i+k}$, and the magnitude of the quantity $\frac{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}\ +\ (k-1)n_i}{(n_i\ +\ n_{i+j}+\ k\ -1)\ n_{i+j}}$

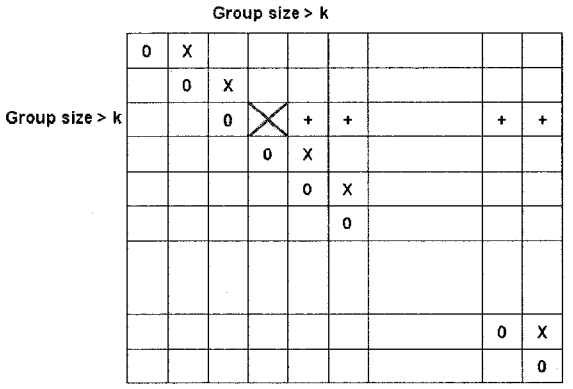


Figure 3.2: Components of the similarity distance matrix for a micro-data file that have to be computed for algorithm kWD

will play the predominant role in determining the value of the minimum distance in the distance matrix. In this case $D(G_i, G_{i+k}) \leq D(G_i, G_{i+j})$, which specifies the only case where (3.3) does not follow.

The theorem follows. □

Theorem 1 can be used to reduce the computational time required for determining the nearest pair of distinct groups to be merged in Step 2 in the k-Ward_MAT algorithm. As shown above, the condition $D(G_i, G_{i+k}) \geq D(G_i, G_{i+j})$ is almost always true, except in a single case where both groups G_{i+j} and G_{i+k} have elements with a single repetitive entry - which is an event occurring with a very small probability. Thus, if this event is ignored, it is obvious that computing the distances between G_i and G_{i+1} for all $i \leq n-1$ is sufficient to determine the minimum distance value in the distance matrix. It turns out that, due to the ordering of the elements (and the sets), the values of these distances lie on the principal diagonal above the main (zero) diagonal. This is, indeed, what the kWD algorithm achieves at Step 2, thus significantly reducing the computational time.

Theorem 2. *Algorithm KWD produces the same output as the k-Ward strategy.*

Proof.

Before proceed, we observe that by virtue of the previous result, algorithm kWD will be considered correct if it correctly achieves the following steps:

1. Computing the values, which lie on the diagonal above the main zero diagonal. As argued above, this is adequate when it concerns determining the minimum distance between groups in the similarity distance matrix.

2. Ensuring that none of the values which lie on the diagonal above the main zero diagonal, represent the merging of two groups both of which have a size greater than or equal to k.

In the case when both the merged groups have a cardinality less than k, the first point above follows directly as a consequence of Lemma 1 and Theorem 1. This is because the values which lie on the diagonal above the principal diagonal, truly evaluate the minimum distance values among the entire set of values above the main zero diagonal. More formally , this is true as a consequence of the fact that $D(G_i, G_{i+k}) \geq D(G_i, G_{i+j})$ for $i + j \leq m$, $i + k \leq m$, and $k > j$.

The argument becomes marginally more complex when the groups involved have cardinalities greater than k, which represents the second condition. In this case, if the distance value represents the event associated with such a merge, the k-Ward$_MAT$ will search for the next (or second) minimum distance between two groups along the same row, and will then merge them. The next minimum distance value is the nearest (nonadjacent) neighbor on the same row, whose cardinality is less than k. Observe that this is analogous to Step 2 of algorithm kWD, where for every group (say, with row-index i) which has a cardinality greater than or equal to k, the latter algorithm seeks out a group (with column index, j) whose cardinality is less than k. Thus, to determine the group with the associated column, all we need to do is to traverse the row from index $i+1$ till we find such a group. This is, of-course, true by virtue of the non-increasing sequence of distances along the row, (namely, those indicated by the '+' symbol in the associated row as in Figure 3.2). During the traversal, we either

encounter a group of size less than k, or result in the conclusion that no such group exists. In the first scenario, the computed distance value will be the entry used for the minimum distance candidates. As opposed to this, in the second scenario, the minimum value will be one of the remaining values which lie on the diagonal above the zero diagonal (see Figure 3.2). Since these distances are computed in the body of the loop, the theorem follows. □

3.4.2 Recursive k-Ward Optimization

Having discussed how the k-Ward computation can be improved for a single partitioning, we now propose a recursive and superior mechanism, referred to as *Recursive k-Ward Micro-aggregation*, kWR, for further minimizing the computations for the entire data set. Our strategy is the following: Rather than process all the data using a k-Ward method, we propose that the data be recursively subdivided into smaller subsets. We emphasize that the smaller subsets need not be obtained as the result of invoking a k-Ward algorithm on the original data. This strategy leads to a sequence of so-called divide-k-Ward-coalesce steps, which are invoked recursively to ultimately yield the desired micro-aggregated records. This recursive subdivision can be "arbitrary" (*i.e.*, based on any meaningful criterion), and it does not need to utilize any underlying clustering methodology. Furthermore, we propose that each subset be micro-aggregated independently. Finally, the micro-aggregated records are combined in order to obtain the entire set of records appropriately grouped.

The algorithm that implements the kWR can be formalized as follows. Let the input data set be given by $InSet$ and the output micro-aggregated records be $OutSet$. We first partition the $InSet$ set into J mutually exclusive subsets $InSet_1, InSet_2, \ldots, InSet_J$. We then compute the SST for this set and for each subset. Subsequently, we calculate the value of $\frac{\Sigma_{i=1}^{J} SST(InSet_i)}{SST(InSet)}$ and compare it to a user-defined threshold, θ. If the value of $\frac{\Sigma_{i=1}^{J} SST(InSet_i)}{SST(InSet)}$ is smaller than θ, a traditional k-Ward MAT is invoked to yield the micro-aggregated records. Otherwise, the original data set is further recursively subdivided into the J sub-subsets, and the process

is recursively invoked, thus proceeding towards the leaves of the recursive tree. Observe that a traditional k-Ward MAT is invoked only when the corresponding input set is not only small enough but when it also preserves the minimization of the IL.

Algorithm 4 kWR

Input: InSet: A set of sorted data set records; θ: The user-defined threshold; J: A fixed constant.
Output: OutSet: The micro-aggregated records
Method:

1: **if** (Size(InSet) $\geq 2Jk$) **then**
2: Partition $InSet$ into J mutually exclusive Sets $InSet_1, InSet_2, \ldots, InSet_J$
3: **if** ($\theta > \frac{\Sigma_{i=1}^{J} SST(InSet_i)}{SST(InSet)}$) **then**
4: call k-Ward $(InSet)$
5: **return** $OutSet$
6: **else**
7: **for** $i = 1 \leftarrow J$ **do**
8: call Recursive-k-Ward_MAT ($InSet_i$,$OutSet_i$)
9: **end for**
10: **return** $OutSet = OutSet_1 \bigcup OutSet_2 \cdots \bigcup OutSet_J$
11: **end if**
12: **else**
13: call k-Ward $(InSet)$
14: **return** $OutSet$
15: **end if**
16: **End Algorithm** kWR

The proposed recursive method is consistent with our overall objectives of minimizing the IL, minimizing the DR, and reducing the required time to micro-aggregate all the records in the data set. The main advantage of this modification, compared to the corresponding non-recursive version, is that kWR can micro-aggregate the data set in *significantly* less time without sacrificing either the IL or the DR. Another advantage of such a strategy is that such a modification will provide a natural way to design an efficient algorithm inherently suitable for multi-processor machines, and especially for shared-memory systems where the communication of data between processors does not have to be planned in advance because distinct subsets can be

executed on different processors. Apart from this, such a method can also make efficient use of memory caches because after a subset is small enough to be stored in cache the partitioning can be achieved without accessing the slower main memory.

The most intricate part of designing such a recursive method involves determining a suitable threshold value, θ, because it depends on the structure, distribution, and statistical measures of the data set. Indeed, from our experience, a good initial guess of θ is generally around the value of $\frac{\Sigma_{i=1}^{J} SST(InSet_i)}{SST(InSet)}$.

Using a recursive strategy ensures that at least J subsets are obtained from the entire data set, which leads to a noticeable saving on time. Further, by this recursive method we not only reduce the required time to micro-aggregate the records, but also attempt to preserve the minimization of the IL. This is achieved by invoking the base (terminating) step, where the IL is minimized for each atomic partition.

Theorem 3. *The running time for the kWR Algorithm is bounded by* $[\bigcirc(n \log n), \bigcirc(n^2)]$.

Proof.
The proof of our complexity claim contains two parts:

1. We first prove that the upper bound of the complexity of algorithm kWR is given by $\bigcirc(n^2)$.
   To do this, we observe that the worst case of algorithm kWR will occur when the data elements are maximally close to each other. This occurs when there are no gaps between the various data elements, leading to a satisfaction of the θ condition on Line 3. Such a scenario will lead to the further invocation of the original k-Ward algorithm, which, in the worst case, leads to an obvious complexity of order $\bigcirc(n^2)$ in.

2. We now prove that the lower bound of the time complexity of algorithm kWR is given by $\bigcirc(n \log n)$.
   The best case of algorithm kWR occurs when the data is full of gaps that renders the θ condition (see Line 3) to be always *False*. The size of the *Inset*

data set will then determine the complexity of the algorithm. We see that there are three mutually exclusive and exhaustive cases:

- **Case 1**. This is the case when the size of $InSet$, n, is greater than or equal $2Jk$. Such a case will cause the recursive invocation of algorithm kWR J times after the θ condition fails. This will require $\bigcirc(n)$ time to compute the SST for each subset.
- **Case 2**. This is the case when the size of $InSet$, n, is less than $2Jk$ and greater than $2k$. In this case, the latter cannot be further divided into J subsets. Hence, the original k-Ward algorithm will be invoked, leading to a time complexity of $\bigcirc(n^2)$.
- **Case 3**. When the size of $InSet$ is equal to $2k$, this will further invoke the k-Ward algorithm. Such an invocation will place the first k elements in the first subset, and the last k elements in the second group. Thus, since there are no elements between the two subsets, the algorithm terminates with $\bigcirc(1)$ computations[8]. Now, it is clear that the event that $InSet$ has a size smaller than $2k$ is an event of probability *zero*, because when the size of $InSet$ is less than $2Jk$, it will not be further divided into J subsets. This implies that this event cannot occur.

Merging these arguments, we see that the running time of algorithm kWR is $T(n)$, where:

$$T(n) = \begin{cases} JT(n/J) + \bigcirc(n), & n \geq 2Jk \\ \bigcirc(n^2), & 2k < n < 2Jk \\ \bigcirc(1), & n \leq 2k. \end{cases}$$

By considering the different "paths" of the event tree when the above scenarios can occur, it can be shown that[9]:

$$T(n) = \sum_{i=0}^{\log_J(n-2Jk)} J * \frac{n}{J} + \sum_{i=2k}^{2Jk} \bigcirc(i^2) + \bigcirc(1). \tag{3.4}$$

[8]This is, quite simply, because the set is already sorted.

[9]The details are omitted in the interest of brevity.

Consider the first term. By using the so-called *Mastering* method [18], we can see that $\sum_{i=0}^{\log_j(n-2Jk)} J * \frac{n}{J} = \bigcirc(n \log n)$. Similarly, when considering the second term, we see that $\sum_{i=2k}^{2Jk} \bigcirc(i^2) = \bigcirc(1)$ because k and J are constants, and the values of both of these terms are negligible comparable to n, the size of the data. Thus, the best case running time of algorithm kWR is $\bigcirc(n \log n)$.

The theorem follows. □

3.5 Experimental Results

In this section, we describe our data sets, explain our experimental methods, and present the results we have obtained by using kWD, kWR, and $kWDR$ for micro-aggregation. All the programs were written in the C^{++} language, and the tests were performed on an Intel(R) Pentium (R)M Processor with the clock speed of 1.73 GHz., and with 512 MB of RAM.

The data used in the experiments is given in Section 3.5.1. Sections 3.5.2 and 3.5.3 report the results for the univariate MAT. The corresponding results for the projected multi-variate data sets are given in Section 3.5.4.

3.5.1 Data Sets

We tested the various versions of the k-Ward_MAT using the two reference data sets[10] that have been used as benchmarks in previous studies [45, 53]:

1. **Tarragona Data Set** [45]

   This set involves 834 companies in the Tarragona area. Each company has 13 associated quantitative variables, as follows: Fixed_Assets ($Var1$), Current_Assets ($Var2$), Treasury ($Var3$), Uncommitted_Funds ($Var4$), Paid-up_Capital

[10] In future, we shall merely refer to these data sets as the *Tarragona* and *Census* data sets.

($Var5$), Short-term _Debt ($Var6$), Sales ($Var7$), Labor_Cost ($Var8$), Description ($Var9$), Operating_Profit ($Var10$), Financial_Outcome ($Var11$), Gross_Profit ($Var12$), and Net_Profit ($Var13$). Table 3.1 shows the main descriptive statistics of the Tarragona data set.

Table 3.1: Statistical Summary for Tarragona data set

Tarragona Data Set						
Variable	Maximum	Minimum	Mean	Median	Mode	Std.Dev.
Var1	4994098	0	105338.8261	33930.5	0	278227.8001
Var2	4539074	-26538	210314.0204	90545	0	432394.3559
Var3	497490	-38776	19352.86811	5656	0	44704.9887
Var4	3939325	-515464	120218.6571	40580	769	303023.5211
Var5	1493787	90	39946.43165	11000	10000	105988.4465
Var6	3601182	-2701	159344.012	67045	10393	326846.4954
Var7	15382214	0	546958.2818	244794.5	277939	1155792.694
Var8	1372449	0	74447.95923	36200	0	135407.5982
Var9	378927	-119	10855.10312	3614	0	27193.58137
Var10	1207056	-251798	27622.76499	8475.5	4750	88221.54817
Var11	108058	-357071	-8366.248201	-3548	0	25690.3612
Var12	1267541	-301172	21243.48441	4055	2580	80287.94397
Var13	873843	-301172	14133.80096	2844.5	1955	56155.53656

2. **Census Data Set** [53]

This set was obtained on *July* 27, 2000, using the Data Extraction System of the U.S. Bureau of the Census. It contains 1,080 records that were obtained using the extraction procedure described in [14], and it has 13 quantitative variables[11]: Final_Weight ($Var1$), Adjusted_Income ($Var2$), Insurance_Contribution ($Var3$), Business_Earnings ($Var4$), Federal_Tax ($Var5$). Payroll_Deduction ($Var6$), Interest_Income ($Var7$), Total_Earnings ($Var8$), Total_OIncome ($Var9$), Total_Income ($Var10$), State_Tax ($Var11$), Taxable_Income ($Var12$), and Amount (*i.e.,* total wage and salary) ($Var13$). Table 3.2 shows the main descriptive statistics of the Census data set.

[11]The full forms of the variables can be found in [14].

Table 3.2: Statistical Summary for Census data set

Census Data Set						
Variable	Maximum	Minimum	Mean	Median	Mode	Std.Dev.
Var1	689039	13567	196039.81	180349.00	13567	101251.42
Var2	99894	6539	56222.76	58412.50	6539	24674.84
Var3	7091	16	3173.14	3215.50	16	1401.83
Var4	21260	1	7544.66	7068.00	1	4905.20
Var5	116721	3570	45230.84	43278.00	3570	21323.47
Var6	11480	2	2597.18	2322.00	2	1826.44
Var7	83454	8	39712.95	41155.00	8	21224.16
Var8	105941	1	5162.23	1586.50	5000.00	9449.64
Var9	49425	1	1421.41	353.00	500.00	3750.89
Var10	97604	80	40068.61	39000.00	30000.00	20816.01
Var11	7932	6	2962.65	3002.00	4207.00	1427.23
Var12	97604	80	39523.38	38000.00	30000.00	20601.28
Var13	97604	80	38444.56	36000.00	30000.00	20677.57

3.5.2 Results for Uni-variate MATs

The run time characteristics of the optimized k-Ward micro-aggregation algorithms, (for the uni-variate case), for the data sets are discussed in more detail in this section.

The sorted data approach was chosen to obtain the specific implementation of kWD. Table 3.3 presents the results obtained by executing kWD and the IL 100 times for each variable. From the table we see that kWD leads to a huge reduction in the running time required to micro-aggregate all the records in the data set compared to the original k-Ward_MAT. For example, the required time to micro-aggregate the values of $Var11$ in the Tarragona data set using k-Ward_MAT was 19.86 seconds while it was only 3.95 seconds using kWD. The percentage of time improvement reaches up to 80.11% on the Tarragona data set. Similarly, the required time to micro-aggregate all values of $Var2$ in the Census data set was 42.41 seconds using k-Ward_MAT, but the kWD required only 8.36 seconds. Again, the percentage of time improvement is as high as 80.29% on the Census data set. Interestingly, both the k-Ward_MAT and kWD obtain the same value of the IL for each variable in both the data sets.

Table 3.3: Comparison between the original k-Ward_MAT, kW, and the optimized kWD on the Tarragona and the Census data sets. In this case we used the value of k=3.

Tarragona Data Set						Census Data Set					
Var	kW		kWD			Var	kW		kWD		
	IL	Time	IL	Time	Improv.(%)		IL	Time	IL	Time	Improv.(%)
Var1	7.176	19.89	7.176	5.41	72.80%	Var1	0.131	42.44	0.131	9.48	77.66%
Var2	0.557	19.72	0.557	4.98	74.75%	Var2	0.001	42.41	0.001	8.36	80.29%
Var3	0.775	19.77	0.775	4.99	74.76%	Var3	0.008	42.58	0.008	8.42	80.23%
Var4	1.487	19.61	1.487	4.88	75.11%	Var4	0.005	42.03	0.005	8.84	78.97%
Var5	2.018	49.03	2.018	12.78	73.93%	Var5	0.024	41.92	0.024	8.91	78.75%
Var6	0.586	19.69	0.586	5.03	74.45%	Var6	0.034	42.00	0.034	9.14	78.24%
Var7	1.921	19.81	1.921	5.03	74.61%	Var7	0.002	42.06	0.002	8.36	80.12%
Var8	0.338	19.73	0.338	4.95	74.91%	Var8	0.443	41.69	0.443	10.66	74.43%
Var9	1.287	20.02	1.287	5.47	72.68%	Var9	0.732	44.84	0.732	10.05	77.59%
Var10	1.905	19.69	1.905	4.67	76.28%	Var10	0.003	47.02	0.003	10.34	78.01%
Var11	2.966	19.86	2.966	3.95	80.11%	Var11	0.008	47.11	0.008	10.28	78.18%
Var12	4.748	19.72	4.748	4.72	76.06%	Var12	0.004	47.31	0.004	10.28	78.27%
Var13	4.958	19.84	4.958	4.83	75.66%	Var13	0.005	48.34	0.005	10.59	78.09%

In the case of kWR, the data set was partitioned into 2 subsets ($J = 2$) whenever the base condition was not satisfied. As explained previously, finding the best threshold value, θ, is far from trivial. Indeed, the value of θ plays an important role in determining the required time, and also in determining the value of the IL. Thus, if θ has a certain value, (which does not allow the base condition to be satisfied), it is clear that the recursion will be invoked more often; this will lead to a huge reduction in the running time, but to an increased value in the IL. It is important to mention that the number of recursions invoked differs from one variable to another, but for each specific variable the number of recursive calls seems to be the same regardless of the value of k. This makes sense because the recursion divides the data set into subsets, as per the base condition, before invoking the micro-aggregation method. Besides, we also observe that the base terminating condition is generally satisfied at the same positions for each specific variable regardless of the value of k. Finally, we have observed that changing the value of k affects the value of the IL and the running time, but it does not seem to affect the total number of recursive invocations.

Tables 3.4 and 3.5 show a comparison between the original k-Ward_MAT and kWR on the Tarragona and Census data sets, respectively. In the table, the sequence of recursive calls is represented by an expression of well-matched parentheses. Thus, for example, there are two recursive calls for the variable $Var2$ in the Tarragona data set, and this is represented by the parenthetic string "(()(()))". The first time the algorithm is invoked, it divided the entire data set into two partitions, each recursively invoking the algorithm again. The first partition was continued with no further divisions or recursions, while the second partition was divided into another two sub-partitions and the algorithm was recursively invoked for each sub-partitions, which ultimately terminated the process. It turns out that in this case the value of $\theta = 0.8$ is suitable for all the 13 variables of the Tarragona data set, and the percentage of the time improvement reaches up to 93.65% on $var12$ after 8 recursive calls. But for the Census data set, we observe that we have to set θ differently for each variable, and that values of θ- which are respectively 0.4, 0.28, 0.31, 0.28, 0.31, 0.38, 0.29, 0.70, 0.80, 0.31, 0.27, 0.32, and 0.30 (respective to the corresponding variables $Var1$ to $Var13$)- lead to the best solution. kWR yielded the same value of the IL in most of these variables, except for the cases of $Var4$, $Var5$, $Var6$, $Var11$, and $Var12$ where it gave a value close to the value of IL obtained by the original k-Ward_MAT as shown in Table 3.5. We believe that the reason for this minor deviation from the optimal is due to the nature of the data. An examination of the values of data in the Tarragona data set shows that since they are relatively close to each other, there are not many gaps between the values. But in the case of the Census data set such gaps exist. Thus, there are many divisions on the Census data set, which will marginally increase the IL. However, we still believe that such a small loss is worth the significant computation gain.

3.5.3 Dynamic threshold

Since we are attempting to reduce the computational time and the IL, we have also experimented the schemes using a dynamic threshold, which varies at each step. This

Table 3.4: Comparison between the original k-Ward_MAT, kW, and the optimized kWR using a fixed threshold on the Tarragona data set. In this case we used the value of k=3.

Var	kW		kWR			
	IL	Time	IL	Time	Improv.(%)	Recursion
Var1	7.176	19.89	7.176	5.31	73.30%	(()(()(()())))
Var2	0.557	19.72	0.557	3.75	80.98%	(()(()()))
Var3	0.775	19.77	0.775	5.91	70.11%	(()())
Var4	1.487	19.61	1.487	1.49	92.40%	((((((())())())())(()()))
Var5	2.018	49.03	2.018	22.36	54.40%	(()(()()))
Var6	0.586	19.69	0.586	3.70	81.21%	(()(()()))
Var7	1.921	19.81	1.922	3.75	81.07%	(()(()()))
Var8	0.338	19.73	0.338	3.75	80.99%	(()(()()))
Var9	1.287	20.02	1.287	3.83	80.87%	(()(()()))
Var10	1.905	19.69	1.909	1.23	93.75%	(((((())())())(()(()())))
Var11	2.966	19.86	2.967	1.99	89.98%	((()())(()()))
Var12	4.748	19.72	4.748	1.19	93.97%	(((((((())())())())())(()(()())))
Var13	4.958	19.84	4.958	1.26	93.65%	(((((((())())())())())(()(()())))

Table 3.5: Comparison between the original k-Ward_MAT, kW, and the optimized kWR using a fixed threshold on the Census data set. In this case we used the value of k=3.

Var	kW		kWR			
	IL	Time	IL	Time	Improv.(%)	Recursion
Var1	0.13	42.44	0.13	12.73	70.00%	(()(()()))
Var2	0.00	42.41	0.00	7.78	81.66%	(()(()()))
Var3	0.01	42.58	0.01	6.94	83.70%	(()(()(()(()(()()))))
Var4	0.00	42.03	0.01	7.02	83.30%	(()(()(()(()(()(()))(()(()(()(()())))))))))
Var5	0.02	41.92	0.03	7.13	82.99%	(()(()(()(()(()(()(()(()))(()(()())))))))))
Var6	0.03	42.00	0.08	6.97	83.40%	(()(()(()(()(()()))(()(()())))))
Var7	0.00	42.06	0.00	12.45	70.40%	(()())
Var8	0.44	41.69	0.44	12.36	70.35%	(()())
Var9	0.73	44.84	0.73	9.95	77.81%	(()(()()))
Var10	0.00	47.02	0.01	13.59	71.10%	(()(()(()(()())))))
Var11	0.01	47.11	0.11	5.23	88.90%	((((((())(((()(()()))(()(()(()(()()))(...))))))))))
Var12	0.00	47.31	0.03	13.67	71.11%	(()(()(()(()()))))
Var13	0.00	48.34	0.01	16.44	65.99%	(()(()(()(()()))))

dynamic threshold was computed as:

$$\theta = \frac{SST(InSet_1) + SST(InSet_2)}{SST(InSet)}.$$

At each step, the dynamic threshold was passed as a *parameter* to the invoked recursion. Tables 3.6 and 3.7 show the power of the dynamic threshold scheme that yields the same value of the IL obtained by original k-Ward_MAT on both the Tarragona and Census data sets. Additionally the computation time is reduced significantly and, thus, in one example, the percentage of the time improvement reaches up to 91.04% on Var 11 in the Tarragona data set, while it is 45.44% on Var 3 in the Census data set.

Table 3.6: Comparison between the original k-Ward_MAT, kW, and the optimized kWR using a dynamic threshold on the Tarragona data set. In this case we used the value of k=3.

Var	kW		kWR			
	IL	Time	IL	Time	Improv.(%)	Recursion
Var1	7.18	19.89	7.18	10.30	48.22%	(()())
Var2	0.56	19.72	0.56	5.78	70.69%	(()())
Var3	0.78	19.77	0.78	5.78	70.76%	(()())
Var4	1.49	19.61	1.49	5.66	71.14%	(()())
Var5	2.02	49.03	2.02	24.89	49.24%	(()())
Var6	0.59	19.69	0.59	5.70	71.05%	(()())
Var7	1.92	19.81	1.92	5.78	70.82%	(()())
Var8	0.34	19.73	0.34	5.72	71.01%	(()())
Var9	1.29	20.02	1.29	5.91	70.48%	(()())
Var10	1.91	19.69	1.91	5.76	70.75%	(()())
Var11	2.97	19.86	2.97	1.78	91.04%	((()())(()()))
Var12	4.75	19.72	4.75	3.88	80.32%	((()())())
Var13	4.96	19.84	4.96	3.73	81.20%	((()())())

Since it is prudent to investigate the computational advantage obtained by using a combination of the two modifications, kWD and kWR, we have also devised an augmented modification, $kWDR$, which is the overall optimized version of the original k-Ward_MAT. Thus, $kWDR$ computes the micro-aggregation by recursive calls, and in each case it computes only the distance values advocated by Theorem 1. The

Table 3.7: Comparison between the original k-Ward_MAT, kW, and the optimized kWR using a dynamic threshold on the Census data set. In this case we used the value of k=3.

Var	kW		kWR			
	IL	Time	IL	Time	Improv.(%)	Recursion
Var1	0.131	42.44	0.131	30.78	27.47%	(()())
Var2	0.001	42.41	0.001	23.48	44.64%	(()(()()))
Var3	0.008	42.58	0.008	23.23	45.44%	(()(()(()(()()))))
Var4	0.005	42.03	0.005	23.33	44.49%	(()(()(()(()()))))
Var5	0.024	41.92	0.024	23.31	44.39%	(()(()()))
Var6	0.034	42.00	0.034	23.52	44.00%	(()(()(()())))
Var7	0.002	42.06	0.002	23.56	43.98%	(()())
Var8	0.443	41.69	0.443	23.55	43.51%	(()())
Var9	0.732	44.84	0.732	26.13	41.73%	(()())
Var10	0.003	47.02	0.005	29.63	36.98%	(()(()(()())))
Var11	0.008	47.11	0.008	29.73	36.89%	((()(()()))(()(()(()()))))
Var12	0.004	47.31	0.004	29.94	36.72%	(()(()(()())))
Var13	0.005	48.34	0.005	30.83	36.22%	(()(()(()())))

results of $kWDR$ are shown in Table 3.9. The reduction of time reaches up to 97.36% on the Tarragona data set by using a combination of kWD and kWR with a fixed threshold, while a combination of kWD and kWR with a dynamic threshold yields the 95.23% advantage on the Census data set.

To conclude, Figure 3.3 shows a comprehensive comparison between the original k-Ward_MAT, kWD, kWR, and $kWDR$ respectively using the Tarragona Data set. The power of our contribution is, thus, visually obvious.

3.5.4 Results for Multi-variate MATs

The previous experimental results display the value of the quantity IL, and the required computational time for uni-variate $MATs$. We now demonstrate its applicability for multi-variate data, by considering several optimized versions of the k-Ward method on such data. Any multi-variate MAT is designed to permit the simultaneous micro-aggregation of several variables, so as to yield a single k-partition for the entire

Table 3.8: Comparison between the original k-Ward_MAT, kW, and the optimized $kWDR$ which is a simultaneous combination of both kWD and kWR on the Tarragona data set. In this case we used the value of k=3 and a fixed threshold.

Var	kW		$kWDR$		
	IL	Time	IL	Time	Improv.(%)
Var1	7.176	19.89	7.176	3.08	84.51%
Var2	0.557	19.72	0.557	1.23	93.76%
Var3	0.775	19.77	0.775	2.16	89.07%
Var4	1.487	19.61	1.487	0.59	96.99%
Var5	2.018	49.03	2.018	5.70	88.37%
Var6	0.586	19.69	0.586	1.22	93.80%
Var7	1.921	19.81	1.921	1.22	93.84%
Var8	0.338	19.73	0.338	1.24	93.72%
Var9	1.287	20.02	1.287	1.31	93.46%
Var10	1.905	19.69	1.909	0.56	97.16%
Var11	2.966	19.86	2.967	0.81	95.92%
Var12	4.748	19.72	4.748	0.52	97.36%
Var13	4.958	19.84	4.958	0.56	97.18%

Table 3.9: Comparison between the original k-Ward_MAT, kW, and the optimized $kWDR$ which is a simultaneous combination of both kWD and kWR on the Census data set. In this case we used the value of k=3 and a dynamic threshold.

Var	kW		$kWDR$		
	IL	Time	IL	Time	Improv.(%)
Var1	0.131	42.44	0.131	13.39	68.45%
Var2	0.001	42.41	0.001	2.42	94.29%
Var3	0.008	42.58	0.008	2.03	95.23%
Var4	0.005	42.03	0.005	2.20	94.77%
Var5	0.024	41.92	0.024	2.52	93.99%
Var6	0.034	42.00	0.034	2.27	94.60%
Var7	0.002	42.06	0.002	4.28	89.82%
Var8	0.443	41.69	0.443	4.39	89.47%
Var9	0.732	44.84	0.732	5.67	87.36%
Var10	0.003	47.02	0.005	5.70	87.88%
Var11	0.008	47.11	0.011	4.59	90.26%
Var12	0.004	47.31	0.008	6.55	86.16%
Var13	0.005	48.34	0.009	6.92	85.68%

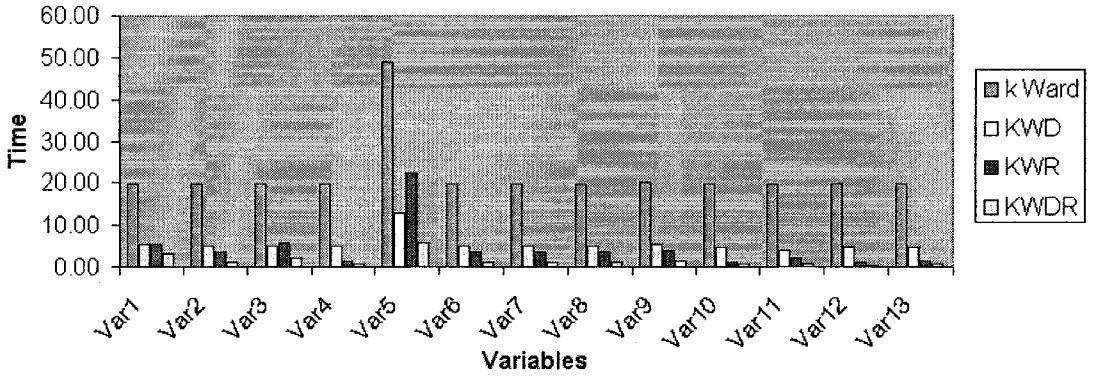


Figure 3.3: Comparison between k-Ward_MAT , kWD, kWR, and $kWDR$ using the Tarragona Data set. In this case we used the value of k=3

data set. As mentioned in Chapter 2, two approaches have been reported for such multi-variate $MATs$. The projected data MAT is one in which the multi-variate data set is projected onto a single axis *prior* to invoking micro-aggregation. Since the projected data leads to a uni-variate data set, it can be sorted, after which any of the optimized versions referred to above can be used for micro-aggregation. As reported in [45], it is important, however that the multi-variate data set be normalized *prior* to the projection and the micro-aggregation phases, to prevent any of the undesired properties discussed there.

To avoid repetition, the results we present are necessarily brief.

Tables 3.10 and 3.11 respectively show a comprehensive comparison between the original k-Ward and the different optimized versions of the projected multi-variate MAT on the Tarragona and Census data sets. The projection is done by using the sum of the z-scores of the variables, ZS, or the first principle component, FPC. The methods by which the data is preprocessed and examined are identical to those reported in [43, 45]. The improvement in the micro-aggregation time by using the

Table 3.10: Results for multi-variate $MATs$ for the Tarragona data set obtained by projecting using the FPC and the ZS indices, In the table we compare the original k-Ward_MAT, kW, and the optimized kWD, kWR using either a fixed threshold, $(kWR-F)$, or a dynamic threshold, $(kWR-D)$, and $kWDR$ which is a simultaneous combination of both kWD and $kWR-D$, by setting $k=3$.

Projection	Method	Tarragona Data Set			
		IL	Time	Improv.(%)	Recursion
	kW	30.13	22.43		
	kWD	30.13	4.89	78.20	
SZ	$kWR-F$	30.13	3.27	85.42	((()(()()))
	$kWR-D$	30.13	10.06	55.15	(()())
	$kWDR$	30.13	6.39	71.51	(()())
	kW	25.35	28.32		
	kWD	25.35	10.86	61.65	
FPC	$kWR-F$	25.35	3.28	88.42	((()(()()))
	$kWR-D$	25.35	5.81	79.48	(()())
	$kWDR$	25.35	2.09	92.62	(()())

Table 3.11: Results for multi-variate $MATs$ for the Census data set obtained by projecting using the FPC and the ZS indices, In the table we compare the original k-Ward_MAT, kW, and the optimized kWD, kWR using either a fixed threshold, $(kWR-F)$, or a dynamic threshold, $(kWR-D)$, and $kWDR$ which is a simultaneous combination of both kWD and $kWR-D$, by setting $k=3$.

Projection	Method	Census Data Set			
		IL	Time	Improv.(%)	Recursion
	kW	28.58	48.82		
	kWD	28.58	0.86	98.24	
SZ	$kWR-F$	28.61	7.33	84.99	((()(()(()(()))(()(()()))(()(()(()(()())))))))
	$kWR-D$	28.58	11.64	76.16	(()(()(()(()(()(()(())))))))
	$kWDR$	28.58	6.56	86.56	(()(()(()(()(()(()(())))))))
	kW	29.54	61.813		
	kWD	29.54	21.63	65.01	
FPC	$kWR-F$	29.54	6.72	89.13	(((((()(()))(()(()(()))))))()())
	$kWR-D$	29.54	7.72	87.51	((((()())()))())
	$kWDR$	29.54	2.61	95.78	((((()())()))())

kWD was as high as 78.20% and 61.65% on the ZS and FPC projections of the Tarragona data set, respectively. Similarly, the time improvement reached up to 98.24% and 65.01% on the ZS and the FPC projection of the Census data set. The kWR method was also applied by using a fixed threshold, $kWR-F$, or a dynamic threshold, $kWR-D$ on the projected data sets as explained in Section 3.4.2. In the case of $kWR-F$, the threshold value was set to 0.8 for the projected Tarragona data set, and to 0.3 for the projected Census data set. These "optimal" setup parameters were obtained by trial-and-error testing. Generally speaking, the time improvement using the $kWR-F$ was more than that obtained by using the $kWR-D$- it reached up to 89.13% by using $kWR-F$ on the FPC projection for the Census data set, and up to 87.61% by using the $kWR-D$ on the FPC projection for the same data. Unfortunately, the value of the IL was sometimes affected by invoking $kWR-F$, (as can be observed in Table 3.11 for the projected ZS of the Census data sets), while the value of the IL remains the same by using the $kWR-D$. Finally, it is important to mention that the combination of the kWD and $kWR-D$ increases the improvement of the required time to 95.78% on the FPC projection of the Census data set without any noticeable change in the value of the IL. It is, thus, clear that the enhanced versions of such proposed schemes are applicable to the processing of multi-variate data sets, by projecting them onto a single axis using the FPC or ZS methods.

3.6 Conclusion

In this chapter, we have considered the problem of securing a statistical database. We have resorted to the well-known micro-aggregation technique, and in particular, considered the k-Ward strategy introduced in [139] and utilized in the state-of-the-art methods [45]. The latter scheme coalesces the sorted data attribute values into groups, and on being queried, reports the means of the corresponding groups. We have shown that such a scheme, as reported in the literature, can be enhanced by minimizing the computations done in evaluating the between-class *distance* matrix and

by recursively partitioning the data set before k-Ward strategy is invoked. When the latter is invoked on the "primitive" sub-groups which terminate the recursion. Our experimental results, done on two benchmark data sets, report a marked improvement. While the IL is comparable to the k-Ward_MAT proposed by Domingo-Ferrer *et.al.* [45], the computations required to achieve this loss is a fraction of the computations required in the latter. This provides a computational advantage that sometimes exceeds 80% if one method is used by itself, and more than 90% if both enhancements are invoked simultaneously.

Finally, although all these methods, were formally proven to be efficient for univariate data sets, we were also able to demonstrate the power of the schemes for multi-variate data. This was achieved by using a projection mechanism, such as the FPC or ZS, and then invoking the MAT on the projected data. Again, the favorableness of the result demonstrated the applicability of the methods.

Chapter 4

A Fixed Structure LA-Based MAT

4.1 Introduction

We consider the MAP which involves partitioning a set of individual records in a micro-data file into a number of mutually exclusive and exhaustive groups. This problem, which seeks for the best partition of the micro-data file, is known to be NP-hard and has been tackled using many heuristic solutions. In this chapter[1] we present the first reported Learning Automata (LA) based solution to this problem. The LA modify a fixed-structure solution to the *Equi-Partitioning Problem* (EPP) to solve the MAP. The newly proposed method competes in a superior manner against the state-of-the-art methods not only from the IL perspective, but when it concerns a criterion involving a combination of the IL and the DR. The scheme has been implemented, tested and evaluated for different real-life and simulated data sets. The results clearly demonstrate the applicability of LA to the MAP, and its ability to yield a solution that obtains the best trade-off between IL and DR when compared to the state-of-the-art.

[1] A preliminary version of some of the results from this chapter appeared in the *Proceedings of PSD'06, the Privacy in Statistical Databases Conference*, in Rome, Italy, in December 2006 [64]. The journal version of these results is currently under review.

The structure of this chapter is as follows: Section 4.2 initially provides a basic introduction to the field of LA, and also explains how LA principles can be used to formulate a solution to the MAP. Section 4.3 presents the Object Migrating Micro-aggregated Automaton both informally and algorithmically. In Section 4.4 we describe how we compute the composite score, which we refer to as the Scoring Index (SI), used for ranking the $MDAV$ and the new LA-based methods with certain specific parameterizations. Experimental results using real-life benchmark and simulated data sets, are discussed in Section 6.3. Finally, Section 6.4 presents the overall conclusions of this work.

4.1.1 Contribution of the Chapter

As mentioned in the literature, this problem in its multi-variate setting is known to be NP-hard [103], and has been tackled using different approaches such as hierarchical clustering [45, 91, 92], genetic algorithms [45, 91, 92, 128], graph theory [74, 84], and fuzzy clustering [57, 133]. All the heuristic $MATs$ seek to minimize the value of the IL. Observe that minimizing the IL, though important, is difficult to enforce, essentially because any SDC strategy is primarily intended to enhance the security of the data. Indeed, the definition of optimality for an SDC is defined in the literature as being equivalent to offer the best trade-off between the IL and DR [27, 93]. In this vein, the recent development of $MATs$ [60] leaves the researcher no excuse to circumvent the problem of trying to reduce the value of the IL as much as possible [49].

As argued by most researchers, maintaining a happy medium between the IL and DR is not trivial. Indeed, even the question of how such a compromise can be attained has not been fully investigated. In this chapter, we shall argue that a good MAT should be capable of minimizing the IL, and yet be able to attain to suitable value for a well-defined composite measure. One such measure, which we refer to as the SI, is a linear combination of the IL and DR. Our aim is to find a good strategy to optimize the latter.

In general, minimizing the IL directly follows the goal of maximizing the similarity between the records in each group. The latter, in turn, can be maximized by using two criteria, namely that of requiring the total SSE to be as low as possible and, simultaneously, preserving the relationships between the individual records in the micro-data file. This is precisely the rationale of our current methodology. In this chapter, we achieve this by attempting to solve the problem using the principles of LA. We have described one such LA called the Object Migrating Micro-aggregated Automaton ($OMMA$), which is capable of grouping similar records, and simultaneously differentiating records which do not belong together. The unique strength of this methodology lies in its applicability to cluster *multi-variate* data sets without resorting to projection in any preferred direction. The chapter, thus, demonstrates the power of LA to minimize the IL, leading to results comparable to those obtained from the best available heuristic methods for micro-aggregation such as the *Maximum Distance to Average Vector* ($MDAV$) [60] and the *Minimum Spanning Tree Partitioning Algorithm* (MST) [84]. The remarkable reduction obtained in the IL can be noticed when micro-aggregating multi-variate data (which sometimes exceeds 11% when compared to the $MDAV$ and MST), and it renders the contribution of this chapter significant. Apart from this, the $OMMA$ also yields the best reported values for the above mentioned index, the SI. We argue that the applicability of LA to the MAT provides a promising strategy to effectively protect sensitive data in the micro-data file not only based on minimizing the value of the IL but also on offering the best trade-off between the IL and DR. The limitations and possible extensions of the $OMMA$ are explained in a subsequent section.

It should be mentioned that by using our strategy, the $OMMA$ can actually be applied to many types of attributes - continuous, ordinal or nominal. The only difference that we foresee will be in designing the aggregation operator (*i.e.*, one which yields the "mean" for the continuous data, or the median for categorical data), which is computed for each group, and which is used to replace the original records. But the main strength of the methodology lies in its applicability to cluster *multi-variate* data sets without resorting to projection in any preferred direction. This is because

the scheme does not *merely* utilize the distance between the individual records, but it also attempts to keep the SSE minimum.

4.2 Learning Automata (LA)

Our solution to the MAP involves LA. LA have been used to model biological learning systems and to learn the optimal action which a random environment offers. The learning is accomplished by actually interacting with an environment and processing its responses to the actions that are chosen, while gradually converging toward an ultimate goal. There are various applications that use LA including parameter optimization, statistical decision making, telephone routing, pattern recognition, game playing, natural language processing, modeling biological learning systems, string taxonomy, detection and tracking, distributing the process of a parallel application, routing in communication networks and object partitioning [71, 75, 95, 98, 99, 105, 109, 111, 112, 113, 147]. Since the literature on LA is extensive, we refer the reader to [98] for a good survey.

The functionality of the LA can be described as a sequence of repetitive feedback cycles. The feedback cycle involves two entities, the *Random Environment* and the LA. During each cycle the automaton chooses an action, which triggers a response from the Environment, and uses the received response - that can be either a reward or a penalty- with the knowledge gained from the previous cycles to determine which is the next action to be chosen. By the process of learning, the automaton adapts itself to the Environment and determines the optimal action, i.e., the action which has the minimum penalty probability, or has a maximum reward probability.

Incorporating LA in any application domain is an evidence of the power of the philosophy. Basically, LA learn from the random environment. The actual technique involved in applying the LA philosophy in the different applications involves modeling the actions, simulating the transforming functions, and representing the system's output in order to have reward or penalty responses. This is where the creativity of

the researcher becomes apparent.

Stochastic LA can be classified into two main classes:

1. Fixed Structure Stochastic Automata ($FSSA$): These $FSSA$ have the property that their transition and output functions do not change with time. These LA seem to possess powerful properties useful for solving different NP-hard problems, as we will show in this chapter.

2. Variable Structure Stochastic Automata ($VSSA$): These $VSSA$ have a dynamically changing structure, because their transition and output matrices are time varying. In practice, however, they are merely defined in terms of action probability updating rules which are either of a continuous or discrete nature [3, 110]. Automata with a variable structure are generally much faster in their convergence.

The basic idea used to solve the MAP is based on a sub-class of LA solutions that has been used to solve the object partitioning problem [71, 111, 112]. As documented in the literature, the object partitioning problem involves partitioning a set of $|\mathbb{P}|$ objects into $|\mathbb{N}|$ groups or classes, where the main aim is to partition the objects into groups that mimic an underlying unknown grouping. In other words, the objects which are accessed together must reside in the same group [112]. In the special case when all the groups are required to contain the same number of objects, the problem is also referred to as the EPP. Many solutions involving LA have been proposed to solve the EPP, but the most efficient algorithm is the *Object Migrating Automaton* (OMA) [112]. The latter was first proposed by Oommen and Ma [112], and some modifications to the original algorithm were added by Gale *et.al.* [71] to create the *Adaptive Clustering Algorithm* (ACA).

For the rest of this section, in all brevity, we discuss the fundamental of LA and present the algorithm due to Oommen and Ma [112] and the modified version by Gale *et.al.*.

4.2.1 Similarities Between the MAP and the EPP

The idea behind using LA as an MAT comes from the elegance of using LA in solving the EPP. LA possess many attractive features that make them applicable to the MAT such as simplicity, ease of implementation, and the ability to study the relation between the individual objects/records without computing any statistical queries. We will show that the similarity between the EPP and the MAP render LA as one of the promising candidate tools to solve the latter. This is because :

- The EPP and MAP are NP-hard problems, primarily due to the exponential growth in the number of partitions of records/objects.
- The EPP dictates that each partition must have the same number of objects, which is a condition analogous to the one associated with the MAP for the fixed-size micro-aggregation criterion, which states that each group must have approximately the same number of records.
- The EPP and MAP seek to partition the records/objects into groups that mimic the underlying unknown groups. In the case of the EPP, the objects which are accessed together more frequently by a random sequence of queries are said to be in the same partition. As opposed to this, in the MAP, the records which are similar to each other, are required to be in the same group so as to maximize the within-group and to minimize the between-group similarities.

4.2.2 Fundamentals

An $FSSA$ is a quintuple ($\underline{\alpha}$, $\underline{\Phi}$, $\underline{\beta}$, F, G) where

- $\underline{\alpha} = \{\alpha_1, \ldots, \alpha_R\}$ is the set of actions that it must choose from.
- $\underline{\Phi} = \{\phi_1, \ldots, \phi_S\}$ is a set of states.

- $\underline{\beta} = \{0, 1\}$ is its set of inputs. The "1" represents a penalty, while the "0" represents a reward.
- F is a map from $\Phi \times \beta$ to Φ. It defines the transition of the internal state of the automaton on receiving an input. F may be stochastic.
- G is a map from Φ to α, and it determines the action taken by the automaton if it is in a given state. With no loss of generality, G is deterministic.

As discussed above, the automaton is offered a set of actions, and it is constrained to choose one of them. When an action is chosen, the Environment gives out a response $\beta(n)$ at a time "n". The automaton is either penalized or rewarded with an unknown probability c_i or $1 - c_i$, respectively. On the basis of the response $\beta(n)$, the state of the automaton $\phi(n)$ is updated and a new action is chosen at (n+1). The penalty probability c_i satisfies:

$$c_i = Pr[\beta(n) = 1 | \alpha(n) = \alpha_i] \qquad (i = 1, 2, \ldots, R).$$

4.2.3 Object Migrating Automaton (OMA)

The Object Migrating Automaton (OMA) is an ergodic automaton that has R actions $\{\alpha_1, \ldots, \alpha_R\}$ representing the possible underlying classes. Each action α_i has its own set of states $\{\phi_{i1}, \phi_{i2} \ldots, \phi_{iM}\}$, where M is the depth of memory, and $1 \leq i \leq R$ represents the number of classes. ϕ_{i1} is called the most internal state and ϕ_{iM} is the most external state.

A set of W physical objects $\{A_1, A_2, \ldots, A_W\}$ is accessed by a random stream of queries, and the objects are to be partitioned into groups so that the frequently jointly-accessed objects are clustered together. The OMA utilizes W abstract objects $\{O_1, O_2, \ldots, O_W\}$ instead of migrating the physical objects. Each abstract object is assigned to a state belonging to an initial random group but in its boundary state. The objects within the automaton move from one action to another, and so, in this

case, all the W abstract objects move around in the automaton. If the abstract objects O_i and O_j are in the action α_h, and the request accesses $< A_i, A_j >$, then the OMA will be rewarded by moving them towards the most internal state ϕ_{h1}. But a penalty arises if the abstract objects O_i and O_j are in different classes, say α_h and α_g, respectively. Assuming O_i is in $\zeta_i \; \varepsilon \; \{\phi_{h1}, \phi_{h2}, \ldots, \phi_{hM}\}$ and O_j is in $\zeta_j \; \varepsilon \; \{\phi_{g1}, \phi_{g2}, \ldots, \phi_{gM}\}$, they will be moved as follows:

- If $\zeta_i \neq \phi_{hM}$ and $\zeta_j \neq \phi_{gM}$, O_i and O_j are moved one state toward ϕ_{hM} and ϕ_{gM}, respectively.
- If exactly one of them is in the boundary state, the object which is not in the boundary state is moved towards *its* boundary state.
- If both of them are in their boundary states, one of them, say O_i is moved to the boundary state of the other object ϕ_{gM}. In addition, the closest object to them is moved to the boundary state ϕ_{hM}, so as to preserve an equal number of objects in each group.

It is important to point out that the random stream of queries contains information about an optimal partition, and the OMA attempts to converge to it. The automaton is said to have converged when all the objects associated with a class are in the deepest (or second deepest) most-internal state.

The OMA can be improved by the following: Assume a pair of objects $< O_i, O_j >$ is accessed, where O_i is in the boundary state, while O_j is in a non-boundary state. In this case, a general check should be made to find another object, in the boundary state of the partition containing O_j. If there is an object, then swapping is done between this object and O_i in order to bring the two accessed objects into the same partition. In turn, instead of waiting for a long time to have these accessed objects in the same partition, the convergence speed can be increased by swapping the objects into the right partitions.

The formal algorithm for the OMA can be found in [71, 112]; it is shown in

Algorithm 5 Enhanced *OMA* [71]

Input: The abstract set of objects, a number of state per action, a sequence of random queries in form (O_i, O_j)

Output: A periodic clustering of the objects into R partitions

Notation: ζ_i is the state of the abstract object O_j. It is an integer in the range1 $\ldots RN$, where, if $(h-1)N+1 \leq \zeta_i \leq hN$, then the object O_i is assigned to α_h.

Method:

Initialize ζ_i for $1 \leq i \leq W$ randomly among the boundary state of classes, each class having W/R objects.
for a sequence of T queries **do**
  Read query (A_i, A_j)
  if ((ζ_i div N)=(ζ_j div N)) **then**
    if (ζ_i mod N $\neq$ 1) **then**
      $\zeta_i = \zeta_i - 1$
    end if
    if (ζ_j mod N $\neq$ 1) **then**
      $\zeta_j = \zeta_j - 1$
    end if
  else
    if (((ζ_i mod N)$\neq$ 0) and ((ζ_j mod N)$\neq$ 0)) **then**
      $\zeta_i = \zeta_i + 1$
      $\zeta_j = \zeta_j + 1$
    else if (ζ_i mod N $\neq$ 0) **then**
      if (O_v: unaccessed object in group of O_i where ζ_v mod $N = 0$) **then**
        temp = ζ_j; $\zeta_j = \zeta_v$; ζ_v = temp;
      end if
      $\zeta_i = \zeta_i + 1$
    else if (ζ_j mod N $\neq$ 0) **then**
      if (O_v: unaccessed object in group of O_j where ζ_v mod $N = 0$) **then**
        temp = ζ_i; $\zeta_i = \zeta_v$; ζ_v = temp;
      end if
      $\zeta_j = \zeta_j + 1$
    else
      $temp = \zeta_i$
      $\zeta_i = \zeta_j$
      t = index of an unaccessed object in group of O_j where O_t is closest to ζ_j
      $\zeta_t = temp$
    end if
  end if
end for
return Partitions based on the state$\{\zeta_i\}$
End Algorithm Enhanced *OMA*

Algorithm 5.

4.2.4 Restrictions of the OMA to the MAP

The reported instances of the OMA are not directly applicable for the MAP. To develop our solution, we highlight the main restrictions, and the necessary enhancements which must be added to the OMA in order for it to be useful in our present application domain.

- In case of the MAP, the user does not have access to the stream of random queries. Rather, the only available data is the set of individual records stored in the micro-data file. It is thus apparent that we have to artificially "generate" a sequence of "queries" (or pairs) which can be used to operate on a machine similar to the OMA. The above restriction has a "two-edged" implication. First of all, in the EPP, the user usually requests the system to obtain a query pair of the form $< O_i, O_j >$. However, in the MAP, it is the responsibility of the user to determine which records are similar or dissimilar, and this determination is a problem to be solved in its own right. Secondly, the placement of the objects in the automaton and the stream of random queries, together, serve to either reward or penalize the automaton. However, in the case of the MAP, the question of obtaining a reward/penalty response is not provided by the user, but it has to be *inferred*. This again has to be solved.

- Unlike the EPP, which has no way of penalizing "non accessed elements", a solution to the MAP must develop a strategy for penalizing such records by considering how similar the records within the same groups are. Clearly, this is superfluous for the EPP because, in that problem, the automaton is absolutely dependent on the user's queries. In the present problem, it is crucial that an automaton can quantify how fitting a record is for any given group.

- The optimal partition for the EPP yields crucial information in the stream of random queries. As opposed to this, in the context of the MAP, the system

has no notion of how to characterize the optimal partition. This renders the problem of adapting the OMA to solve the MAP more difficult.

- In the same vein, the definition of the optimal partition for the EPP is quite different from that of the corresponding solution for the MAP. In the case of the EPP, all objects which are accessed together more frequently should be in the same partition, while in the MAP all similar records should be in the same group.

- The criteria which are used to reward and penalize the automaton in the EPP is quite unlike the one used for the MAP. In the EPP, the automaton is rewarded or penalized based on the (unknown) probability of any two objects being jointly accessed. But in the context of the MAP, the automaton is reward or penalized by conducting a comprehensive study of the relation between the individual records, quantified by the various error criteria, such as the distance between each record and the other records in the micro-data file, and more importantly, the overall SSE .

- Although the EPP and MAP utilize analogous rules for a reward phenomenon, as we shall see, they differ in performing the penalty rules. In case of the EPP, the automaton enforces the rule that the pertinent object migrates, if and only if at least one of the accessed objects is at the boundary state of the different partitions. As opposed to this, in the MAP, the automaton enforces the rule that the records are migrated to another group whenever migrating the object reduces the overall SSE.

- The automaton used to solve the EPP is said to have converged, when all the objects are found in the most (or the last two) internal states of each partition. However, we propose that the convergence in the MAP occur when the measure of the IL is unchanged.

4.3 Object Migrating Micro-aggregated Automaton ($OMMA$)

In this section, we define the $OMMA$ as an 7-tuple as: ($\mathbb{U}$, $\underline{\Phi}$, $\underline{\alpha}$, $\underline{\beta}$, $\mathbb{Q}$, $\mathbb{G}$, $\mathbb{L}$), where

- $\mathbb{U} = \{U_1, U_2, \ldots, U_n\}$ is the micro-data file.
- $\underline{\Phi} = \{\phi_1, \phi_2, \ldots, \phi_{hM}\}$ is the set of states.
- $\underline{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_h\}$ is the set of h actions, each representing a group into which the records of $\mathbb{U}$ must fall.
- $\underline{\beta} = \{0, 1\}$ is its set of inputs. The "1" represents a penalty, while the "0" represents a reward.
- $\mathbb{Q}$ is the transition function, which specifies how the records should move between the various states. This function is quite involved and will be explained in detail presently.
- The function $\mathbb{G}$ partitions the set of states for the groups. For each group, α_j, there is a set of states $\{\phi_{(j-1)M+1,\ldots,\phi_{jM}}\}$, where M[2] is the depth of memory. Thus,

$$G(\phi_i) = \alpha_j \qquad (j-1)M + 1 \leq i \leq jM. \tag{4.1}$$

  This means that the record in the automaton chooses α_1, if it is in any of the first M states, and that it chooses α_2 if it is in any of the states from ϕ_{M+1} to ϕ_{2M}, etc. We assume that $\phi_{(j-1)M+1}$ is the most internal state of group α_j,

[2]Generally speaking, the depth of the memory, M in the LA could play an important role in determining the accuracy of the LA, while the eignevalues of the underlying chain would determine the rate of convergence. In practice, the LA is assumed to have converged when all objects are in the most internal states in each partition. To the best of our knowledge we are not aware of any method used to determine the best value for M except the trial-and-error approach. In our case, since the convergence criterion depends on attaining to a non-increasing value of IL, it is not implicitly tied to the value M. Thus, in the case of the $OMMA$, M seems to have no effect on the computational time and the algorithm's convergence and accuracy .

and that ϕ_{jM} is the boundary state. These are called the state of *"Maximum Certainty"* and *"Minimum Certainty"*, respectively.

- $\mathbb{L}$ is the similarity list specifying the records deemed to be collectively similar. It is stored as a list of tuples of the form $< R_i, R_j >$, where the records included are those for which the similarity index is greater than or equal to a predefined threshold, θ.

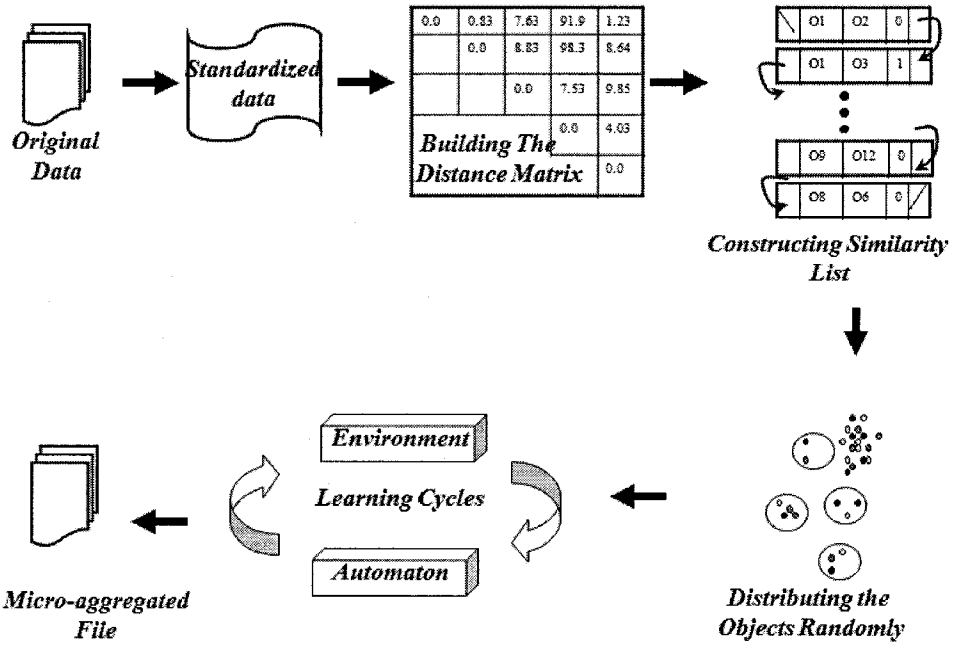


Figure 4.1: The $OMMA$ process used to generate the micro-aggregated file.

Before we explain in greater detail the $OMMA$, it is worth mentioning that this new scheme is applicable when the number of individuals in the micro-data file is an exact multiple of the minimum group size k. When the number of records is not an exact multiple of k, we propose that the MAT be invoked for a marginally smaller file (by removing a few elements in the set), and then including them subsequently in the appropriate partitions - after the MAT process.

The overall model for this application domain is given in Figure 4.1 which shows that the $OMMA$ process consists of three phases involved in generating the micro-aggregated data file as follows:

1. Loading and standardizing the micro-data file.
   The variables encountered in the micro-data file have to be standardized *prior* to the micro-aggregation process. The standardization is performed by replacing the individual values, x_i, for a specific variable, V_j, by $(x_i - \bar{X}_j)/\sigma_j$, where $\bar{X}_j$ and σ_j are, respectively, the average and the standard deviation of the values taken by the variable V_j. The importance of this phase is evident when we consider multi-variate micro-data, because having standardized variables *prior* to the micro-aggregation process removes many undesirable properties, and yields equal weights for all the variables encountered in the micro-data file.

2. Constructing the similarity list that simulates the similarities between micro-records.
   The similarity list, $\mathbb{L}$, simulates the probability of any two records being together in the same group or not. This simulation is based on the real distance value between every pair of records. Therefore, building the similarity list requires creating the symmetric $N \times N$ distance matrix. Each entry in the matrix is computed using the Euclidean metric between the two records in question as follows, $D(R_i, R_j) = \sqrt{\sum_{d=1}^{p}(R_{id} - R_{jd})^2}$, where $D(R_i, R_j)$ is the computed distance between the two records R_i and R_j, and the entry is placed in the i^{th} row and the j^{th} column. The rationale behind building the similarity matrix is to simulate the probability of these two records being together in the same group, and this, in turn, is used in constructing the similarity list. It is worth mentioning that constructing the similarity list is done only once, and this can be considered as a preprocessing operation, and hence the $OMMA$ method is

suitable for small and moderate-sized data sets[3]. Thereafter, it requires comparing the elements of the distance matrix with a certain user-defined threshold, θ. The question of determining θ is open. As it stands now, we have used the following single heuristic: θ is quite simply assigned to be related to the value of the difference between the largest and smallest values in the distance matrix. More explicitly,

$$\theta = \frac{1}{2}[Max_{i,j}\ D(i,j) - Min_{i,j}\ D(i,j)],$$

where $D(i,j)$ is the computed distance value between R_i and R_j. In general, this threshold value is used to make the scheme arrive at an efficient grouping. Finally, if the distance value between R_i and R_j is less than or equal to θ, we will add a tuple $< R_i, R_j >$ to the similarity list. By computing the quantities $D(R_i, R_j)$ between each pair of records in the micro-data file, and by systematically utilizing the similarity list structure, the automaton must adaptively learn how to group the records effectively.

3. Performing the learning cycle, which is where the MAP is solved.
   The learning phase is the core of the clustering, and the $OMMA$ model is initialized by placing all the records at the boundary state of their initially randomly-chosen groups. This indicates that the $OMMA$ is initially uncertain of the placement of the records, because the different states within a given group quantify the measure of certainty that the scheme has for a given record belonging to that group. As the learning cycle proceeds, similar records will be rewarded for their being together in the same group, and they will be penalized by either moving toward their boundary state, or to another group. After the randomly distribution of the micro-records, a certain index will be assigned to each record to specify the state of group it belong to, and the value of the SSE for each group will be calculated to quantify the homogenousity factor of that

[3]The associated computational time can be reduced by using the blocking technique [127]. Since the $OMMA$ runs in a very reasonable time in all the small and medium-sized cases, we have not resorted to the blocking technique in our implementation. The concept of using a hierarchy of $OMMA$ for large data sets has to be investigated.

group.

The $OMMA$ moves into its main learning loop by setting the old value of the IL to be equal to ∞ (a large positive number), and by then processing the constructed similarity list, $\mathbb{L}$, one tuple at a time. It is worth mentioning that the length of the similarity list might be decreased with time by deleting the tuple after being sure that, although these two records are close to each other, their being in the same group will lead to *increase* the total SSE. The latter consideration is relevant since the SSE functionally determines the value of the IL[4].

The list $\mathbb{L}$ is now traversed repeatedly, and similar records in the tuples are processed. If they are both assigned to the same group, the automaton is rewarded. However, if they are assigned to distinct groups, the automaton is penalized. After $\mathbb{L}$ has been processed, we compare the newly computed value of the IL, IL_{new}, with the old value, IL_{old}, produced in the previous cycle. If $IL_{new} < IL_{old}$, the learning phase continues by entering the next learning cycle. But, if $IL_{new} = IL_{old}$, the learning phase terminates and the micro-aggregated file is generated. It should be mentioned that IL_{new} cannot be greater than IL_{old}, because the records will move, if and only if, the SSE remains the same or is reduced. Clearly, the value of IL is non-increasing.

We now describe the actual transitions represented by $\mathbb{Q}$ for each of these operations.

(a) **Transitions for Rewards**

On being rewarded, since the records R_i and R_j are in the same group, say α_u, both of them are moved toward the most internal state of that group, one step at a time. See Figure 4.2 and Algorithm 7.

(b) **Transitions for Penalties**

This case is encountered when two similar records, R_i and R_j, are allocated in distinct groups. say, α_u and α_v respectively (*i.e.*, R_i is in state ζ_i, where

[4]The clustering in the $OMMA$ is achieved based on two *conflicting* criteria - minimizing the distance between the records, and maintaining the total SSE to be as small as possible.

$\zeta_i \epsilon \{\Phi_{(u-1)M+1}, \ldots, \phi_{uM}\}$, and R_j is in state ζ_j, where $\zeta_j \epsilon \{\Phi_{(v-1)M+1}, \ldots, \phi_{vM}\}$). In this case, they are moved as follows:

- Case 1: If both ζ_i and ζ_j are not the boundary states ϕ_{uM} and ϕ_{vM}, respectively, R_i and R_j are moved one state toward the corresponding boundary state. See Figure 4.3 and Algorithm 8.

- Case 2: This occurs when at least one of R_i or R_j is at the boundary state of *Minimum Certainty*, say $\zeta_j = \phi_{vM}$. Studying the effect of migrating any record R_x in α_u to R_j, under the condition that $R_x \neq R_i$, requires investigating the effect of the potential moves on the summation of the SSE_{α_u} and SSE_{α_v}, which we shall refer to as $SSE_{\alpha_{uv}}$. Here we need to consider the $k-1$ different swapping possibilities. If the new value of $SSE_{\alpha_{uv}}$ is less than or equal to the previous value of $SSE_{\alpha_{uv}}$, we have to physically swap the chosen record R_x, which leads to the minimum $SSE_{\alpha_{uv}}$ value with R_j, and proceed to update the values of the SSE for both groups. Subsequently, we assign both R_j and R_x to the boundary state of α_u and α_v, respectively. Otherwise, the tuple $< R_i, R_j >$ will be deleted from the similarity list. This is clarified in Figure 4.4 and Algorithm 9.

- Case 3: If both R_i and R_j are at the boundary states of their different groups, say $\zeta_i = \phi_{uM}$ and $\zeta_j = \phi_{vM}$, we have to study all the different possibilities of swapping any record R_x, under the condition that $R_x \neq R_i$ and $R_x \neq R_j$, in α_u or α_v with R_j or R_i, respectively, on the $SSE_{\alpha_{uv}}$. We then choose the option that leads to reducing the value of $SSE_{\alpha_{uv}}$. In such a case, we physically swap the records, update the values of the SSE for both groups, and assign the swapped records at the boundary state. If we failed to find a candidate to be swapped so as to reduce the value of $SSE_{\alpha_{uv}}$, we should not delete this tuple because we are not in a position to confirm the proposition that these

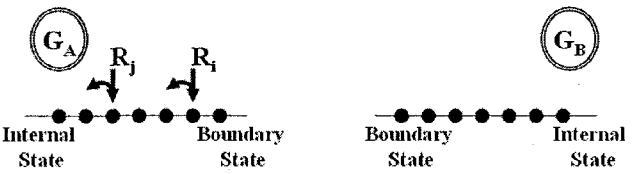


Figure 4.2: The automaton has been rewarded, since both R_i and R_j are similar and located in the same group.

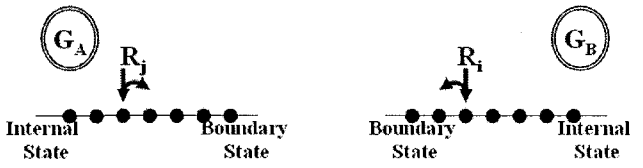


Figure 4.3: The automaton has been penalized, since R_i and R_j are similar but located in distinct groups. Neither of them is at the boundary state.

two records, being in the same group, will decrease or increase the total SSE. This is because both records are in the state of minimum certainty, and, thus, it is likely that during the processing of $\mathbb{L}$ another of the tuples in the similarity list will force one of these records to migrate to another group, rendering the reprocessing of this tuple useful. This scenario is described in Figure 4.5 and Algorithm 10.

For the sake of completeness, (and to not disturb the continuity) the discussed scheme is algorithmically described in Algorithm 6 with the respective modules being given as separate procedures.

Algorithm 6 Learning Phase in the *OMMA*

Input: The standarized micro-data file $\mathbb{U}$, M: the number of states per action, k: the number of records per group, the Similarity List $\mathbb{L}$.

Output: Calculate the value of the IL after this learning cycle.

Note: R_i is in state ζ_i, which belongs to α_u, and R_j is in state ζ_j, which belongs to α_v

Method:

1: Randomly initialize ζ_i for $1 \leq i \leq N$ (N: represents the index of records in $\mathbb{U}$) among the boundary state of classes, each class having k records.
2: Calculate the value of the SSE for each group in the $OMMA$.
3: **repeat**
4: Read tuple (R_i, R_j) from $\mathbb{L}$.
5: **if** $((\zeta_i \text{ div M}) = (\zeta_j \text{ div M}))$ **then**
6: Reward_Record(ζ_i)
7: Reward_Record(ζ_j)
8: **else**
9: **if** $(((\zeta_i \text{ mod M}) \neq 0)$ and $((\zeta_j \text{ mod M}) \neq 0))$ **then**
10: Penalize_Two_Unbound_Records(ζ_i,ζ_j)
11: **else if** $(\zeta_i \text{ mod M} \neq 0)$ **then**
12: Penalize_Bound_Unbound_Records(R_i,R_j,*Activity_Attribute*,G)
13: **else if** $(\zeta_j \text{ mod M} \neq 0)$ **then**
14: Penalize_Bound_Unbound_Records(R_j,R_i,*Activity_Attribute*,G)
15: **else**
16: Penalize_Two_Bound_Records(R_i,R_j,G)
17: **end if**
18: **end if**
19: **until** (all tuples in the Similarity List have been processed)
20: Replace the the individual records by the mean value corresponding to their group
21: **return** The IL value at this learning cycle
22: **End Algorithm Learning Phase in the** *OMMA*

Algorithm 7 Procedure Reward_Record

Input: ζ_i, which represents the index of the state where the record R_i is located in the LA.

Output: The updated value of ζ_i.

Method:

1: **if** (ζ_i mod M $\neq 1$) **then**
2: $\zeta_i = \zeta_i - 1$
3: **end if**
4: **return** ζ_i
5: **End Procedure Reward_Record**

Algorithm 8 Procedure Penalize_Two_Unbound_Records

Input: ζ_i, which represents the index of the state where the record R_i is located in the LA, and ζ_j, which represents the index of the state where the record R_j is located in the LA.

Output: The updated values of ζ_i and ζ_j.

Method:

1: $\zeta_i = \zeta_i + 1$
2: $\zeta_j = \zeta_j + 1$
3: **return** ζ_i and ζ_j
4: **End Procedure Penalize_Two_Unbound_Records**

Algorithm 9 Procedure Penalize_Bound_Unbound_Records

Input: R_i, which represents the record located at ζ_i in the LA, and which is not at the boundary state of α_u. R_j, which represents the record located at ζ_j in the LA, and which *is* at the boundary state of α_v. The other inputs are the Similarity List, $\mathbb{L}$, and the values of the SSE_{α_u} and SSE_{α_v}.

Output: The updated values of ζ_x, ζ_j, the Similarity List $\mathbb{L}$ and the value of the SSE_{α_u} and SSE_{α_v}.

Method:

1: **for all** R_x in the group, α_u, which contains R_i and $R_x \neq R_i$ **do**
2: Compute the new values of SSE_{α_u} and SSE_{α_v} after swapping(R_x, R_j)
3: **end for**
4: Search for the minimum value of $SSE_{\alpha_{uv}}$ which is equal to summation of the new value of SSE_{α_u} and SSE_{α_v}
5: **if** (the new value of $SSE_{\alpha_{uv}} \leq$ the previous value of $SSE_{\alpha_{uv}}$) **then**
6: Swap (R_x, R_j)
7: Update the SSE_{α_v} and SSE_{α_v}.
8: $\zeta_x = vM$, the boundary state of α_v.
9: $\zeta_j = uM$, the boundary state of α_u.
10: **else**
11: delete this tuple from the Similarity List, $\mathbb{L}$.
12: **end if**
13: **return** ζ_x, ζ_j, Similarity List $\mathbb{L}$, and the updated values of the SSE_{α_u} and SSE_{α_v}.
14: **End Procedure Penalize_Bound_Unbound_Records**

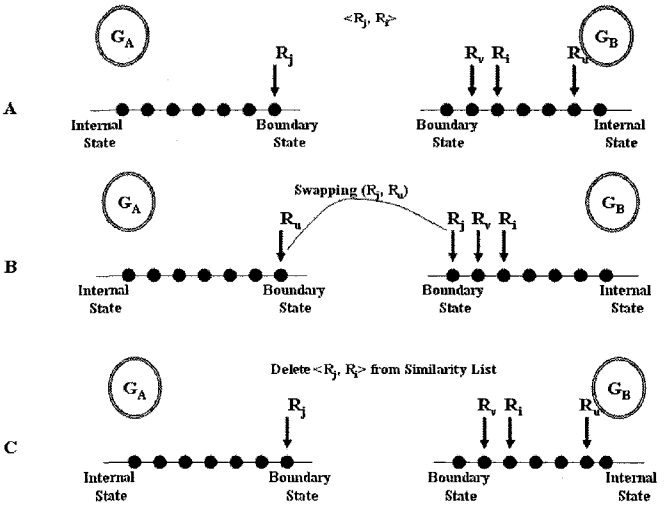


Figure 4.4: (A) The automaton has been penalized, since R_i and R_j are similar but located in distinct groups. R_j is at the boundary state, while R_i is not at the boundary state. After searching for the most suitable record that leads to the minimum value of the SSE, we have two choices for the scenario when $k = 3$. In the first case (see (B) above), we perform a physical swap between (R_u, R_j). In the second case, (in (C)), no record-swap leads to a smaller value of the SSE, and thus, the tuple $< R_i, R_j >$ is deleted from the similarity list.

Algorithm 10 Procedure Penalize_Two_Bound_Records

Input: R_i, which represents the record located at ζ_i in the LA, and which is at the boundary state of α_u. R_j, which represents the records located at ζ_j in the LA, and which is also at the boundary state of α_v. The other input are the values of the SSE_{α_u} and SSE_{α_v}.

Output: The updated values of ζ_i, ζ_j, ζ_x and the value of the SSE_{α_u} and SSE_{α_v}.

Method:

1: **for all** R_x in the group, α_u, which contains R_i and $R_x \neq R_i$ **do**
2: Compute the new values of SSE_{α_u} and SSE_{α_v} after swapping(R_x, R_j)
3: **end for**
4: Search for the minimum value of $SSE_{\alpha_{uv}}$ which is equal to summation of the new value of SSE_{α_u} and SSE_{α_v} and assign it to SSE_1.
5: **for all** R_x in the group, α_v, which contains R_j and $R_x \neq R_j$ **do**
6: Compute the new values of SSE_{α_v} and SSE_{α_u} after swapping(R_x, R_i)
7: **end for**
8: Search for the minimum value of $SSE_{\alpha_{uv}}$ which is equal to summation of the new value of SSE_{α_u} and SSE_{α_v} and assign it to SSE_2
9: **if** $(SSE_1 \leq SSE_2)$ **then**
10: **if** $(SSE_1 \leq$ the previous value of $SSE_{\alpha_{uv}})$ **then**
11: Swap (R_x, R_j)
12: Update the values of SSE_{α_u} and SSE_{α_u}.
13: $\zeta_x = vM$ is assigned to the boundary state of α_v
14: $\zeta_j = uM$ is assigned to the boundary state of α_u
15: **end if**
16: **else**
17: **if** $(SSE_2 \leq$ previous value of $SSE_{\alpha_{uv}})$ **then**
18: Swap (R_x, R_i)
19: Update the SSE_{α_u} and SSE_{α_v}.
20: $\zeta_x = uM$ is assigned to the boundary state of α_u
21: $\zeta_i = vM$ is assigned to the boundary state of α_v
22: **end if**
23: **end if**
24: **return** ζ_i, ζ_j, ζ_x and the updated value of the SSE_{α_u} and SSE_{α_v}.
25: **End Procedure Penalize_Two_Bound_Records**

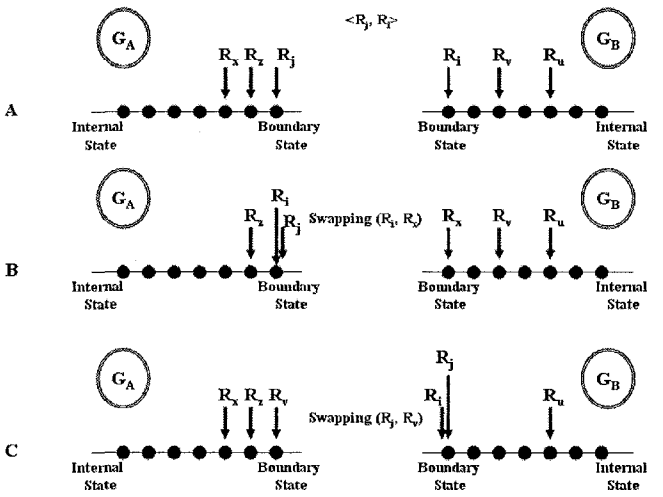


Figure 4.5: (A) The automaton has been penalized, since R_i and R_j are similar but located in distinct groups. However, both of them are in a boundary state. After searching for the most suitable record that leads to the minimum value of the SSE, we encounter four swapping options for the scenario when $k = 3$ - which, in turn, depend on the best candidate record. For example, a physical swap between (R_i, R_x) is done when R_x is the best candidate (as in case (B)). Similarly, a physical swap between (R_j, R_v) is done, when R_v is the best candidate, (as in (C)).

4.3.1 Formal Properties of The $OMMA$

We now state and formally prove a property of the $OMMA$.

Theorem 4. *The learning process in the $OMMA$ leads to a monotonically decreasing IL.*

Proof.
We shall prove the theorem by contradiction. To do this we assume that the value of the IL at time n, $IL(n)$, is less than the value of the IL at time $n+1$, $IL(n+1)$. If the condition $IL(n+1) > IL(n)$ has to be satisfied, it implies that there is at least one movement of two objects (data elements) that causes this increment. This movement could either be a consequence of a reward or a penalty, as seen in Case 1, Case 2, or Case 3 of the LA transition.
Consider the case of a reward. In this case, the pair of objects move one state towards the most internal state. This means that each group continues to have the same objects as in the previous step, which further implies that the value of the SSE of each group does not change. Clearly, such a movement will only result in an identical IL.
Consider now the response referred to as Case 1 in the scenario of a penalty. In this case, the respective objects move a single state toward the most external state. Again, each group still has the same objects as in the previous step, which again implies that the value of the SSE of each group does not change. Thus, clearly, an IL- increasing movement could not have occurred.
Consider, finally, the effect of responding as per Case 2 or Case 3 of a penalty. These are the cases when either of the objects are migrated to another partition. However, by virtue of the pre-checking mechanism (Line 5 of Algorithm 4 and Lines 10 and 17 of Algorithm 5) such an IL- increasing could not have occurred. Note that if the summation of the SSE for the groups after the migration is less than or equal to the summation of the SSE for groups prior to it, the objects will migrate; otherwise no movements is permitted.
Since all the four different types of movements lead to values of the SSE which are

either equal to or less than the total SSE of the previous step, it is clear that the value of the IL at the end of each step is less than or equal to the corresponding value in the previous step. The results follow. □

A straight forward Corollary of this is that the $OMMA$ converges to a local minimum of the IL. We conjecture (from our experiment results) that it also attains the global minimum with the ϵ-optimal property, but this is, as yet, unproven.

4.4 Comparing the $MDAV$and $OMMA$ Methods

Quantifying the quality of $MATs$ is based on two criteria, namely the IL and the DR, both of which have been explained in [53, 146]. In this chapter, we also consider how we can compare $MATs$ using a *composite* measure involving *both* the IL and the DR.

Information Loss: To evaluate the loss in the data utility caused by an MAT on a continuous micro-data file, we seek a measure which assesses how different the masked file is from the original one. Earlier, the quantity IL was measured in a generic way, by essentially estimating how much the data was "harmed" by using the MAT [22, 45, 74, 84, 92]. Such a measurement does not demonstrate how the statistics have been structurally modified by virtue of invoking the micro-aggregation process, and how much the modification is.

We believe that any criterion to evaluate the quality of the published data set should not be only based on the proximity between the masked and the original data files. Rather, it should additionally examine whether the published file preserves the natural statistical *characteristics* of the original file, namely, those which can be evaluated by computing some positional moments such as the means, variances, covariances and correlations. Examinations of this sort will definitely help in building a clear image about the potential of the $MATs$, and in choosing the technique which has the least impact on the desired statistics.

As mentioned in [21, 48, 94, 102, 119], several measures to quantify the IL have been proposed. Five of these measures will be used in this chapter to construct a comparative benchmark. This will be done by assuming that the original and masked micro-data sets are specified in terms of the n ordered individuals such as $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathcal{X}' = \{X'_1, X'_2, \ldots, X'_n\}$, respectively. Observe that each X_i is an instantiation of the random vector (of dimension d) $\underline{X}$, whose mean is $\bar{X}$, and each X'_i is an instantiation of the random variable $\underline{X'}$, whose mean is $\bar{X}'$. Thus, each data vector in the original and masked data sets can be represented as $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T$ and $X'_i = [x'_{i1}, x'_{i2}, \ldots, x'_{id}]^T$, respectively, where both x_{ij} and x'_{ij} are the values associated with the j^{th} variable. Symbolically:

- $\mathcal{X}$ and $\mathcal{X}'$: Are the original and masked data sets.
- $\bar{X}$ and $\bar{X}'$: Are the mean vectors of $\underline{X}$ and $\underline{X'}$, respectively, where those are computed as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{X}' = \frac{1}{n}\sum_{i=1}^{n} X'_i$.
- $\mathcal{V}$ and $\mathcal{V}'$: Are the covariance matrices of $\underline{X}$ and $\underline{X'}$, respectively. Thus, $\mathcal{V} = \mathrm{E}[(\underline{X} - \bar{X})(\underline{X} - \bar{X})^T]$ and $\mathcal{V}' = \mathrm{E}[(\underline{X'} - \bar{X}')(\underline{X'} - \bar{X}')^T]$.
- S and S' : Are the vectors representing the variance of the components of $\underline{X}$ and $\underline{X'}$ respectively, and are given as $S = \mathrm{Diag}[\mathcal{V}]$, and $S' = \mathrm{Diag}[\mathcal{V}']$.
- $\mathcal{R}$ and $\mathcal{R}'$: Are the correlation matrices of $\underline{X}$ and $\underline{X'}$ respectively. If Γ is the diagonal matrix with the standard deviation of the variables along the main diagonal (and with zero's elsewhere), then, $\mathcal{R} = \Gamma^{-1}\ \mathcal{V}\ \Gamma^{-1}$, and $\mathcal{R}' = \Gamma'^{-1}\ \mathcal{V}'\ \Gamma'^{-1}$.

The difference or the "discrepancy" between two matrices can be measured in at least three ways: Mean square error, Mean absolute error, and Mean variation. In this chapter, we use the mean variation[5] when it concerns data structures, data means, data variance, and data covariance, while the mean absolute error is used to measure the data correlation difference[6], as shown in Table 4.1. This leads to five

[5]The following rule is applied to all the mean variation formulae. If $x_{ij} = 0$ and $x'_{ij} \neq 0$, then we divide the difference by $|x'_{ij}|$. If $x_{ij} = x'_{ij} = 0$, the term is not added to the sum.

[6]The mean variation is commonly used because it is a unit-free measure, since the variation is

distinct metrics $M1$, $M2$, $M3$, $M4$, and $M5$, whose significance and explicit forms are tabulated in Table 4.1.

Finally, the overall IL, referred to (G_{IL}), is defined as follows:

$$G_{IL} = 100 * \frac{M1 + M2 + M3 + M4 + M5}{5}, \tag{4.2}$$

where the explicit form for $M1 - M5$ are given in Table 4.1.

Table 4.1: Information loss measures.

Metric	Measurement	Matrix discrepancy	Mathematical expression
$M1$	Mean variation of data	$\mathcal{X} - \mathcal{X}'$	$\frac{\sum_{j=1}^{d}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{nd}$
$M2$	Mean variation of data means	$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^{d}\frac{\lvert \bar{x}_j-\bar{x}'_j\rvert}{\lvert \bar{x}_j\rvert}}{d}$
$M3$	Mean variation of data variances	$S - S'$	$\frac{\sum_{j=1}^{d}\frac{\lvert v_{jj}-v'_{jj}\rvert}{\lvert v_{jj}\rvert}}{d}$
$M4$	Mean variation of data covariates	$\mathcal{V} - \mathcal{V}'$	$\frac{\sum_{j=1}^{d}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{d(d+1)}{2}}$
$M5$	Mean absolute error of data correlations	$\mathcal{R} - \mathcal{R}'$	$\frac{\sum_{j=1}^{d}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{d(d-1)}{2}}$

Disclosure Risk: The effect of applying different $MATs$ is not, really, limited to the IL. We contend that the loss of *confidentiality* as a result of disseminating the published micro-aggregated file must also be analyzed. This is true because the DR does not only depend on the data, but also on the Intruder who knows something about the population. Therefore, any DR assessment must evaluate the risk of providing additional information which can assist in linking a masked record with the corresponding record in the original data set. In addition, we need to

divided by the original variable. However, it is not suitable if the original variable is likely to be zero or close to zero, which is the case for data correlations. For near-zero correlations, even a small change would lead to a significant mean variation, and this would be wrongly interpreted as a large IL. On the other hand, since the correlations are already unit-free, the mean absolute error was used as a good option for this quantity.

evaluate the risk that the values of original records can be accurately estimated from the published masked records [14, 94, 142].

The first kind of risk is evaluated through a strategy known as the Record Linkage Disclosure technique (RLD). Each record in the masked file is linked to its nearest[7] record in the original file. The linkage computation proceeds by evaluating the distances between each record in the masked file and all the records in the original file. Thereafter, the *"nearest"* and *"second nearest"* records for each record in the masked file are considered. A *Match* is said to have occurred when the index of the published record is equal to the index of either the nearest or the second nearest record in the original file [23, 44, 48, 53]. The number of *Matches* yields an estimate of the number of published records whose original identity can be re-identified by the Intruder. DR is defined as the proportion of *Matches* among the total number of records in the original or published file.

To apply this method, researchers generally assume that an Intruder has an external file containing a subset of the key variables that are common to the published file. The Intruder attempts to match the published file and the external file using the subset of common shared variables so as to discover additional information about the original record. The literature reports several disclosure scenarios depending on the number and the identity of the key variables in the original file – which are also assumed to be known to the Intruder.

The second kind of risk is evaluated by means of a so-called confidential Interval Disclosure (ID) technique. In this case, we assume that the Intruder is not attempting to pair masked records with original records. Rather, the Intruder is only interested in determining an approximate value for the original record. The ID operates as follows: Each variable is independently ranked, and a rank interval is defined in the neighborhood of the value that the variable takes on for a specific record, say r. The rank of this value should be less than $P\%$ of the total number of records, and the

[7]This distance is typically measured using the Euclidean distance. It is worth mentioning that both files are usually standardized [23, 53].

value in the center of the interval should correspond to the value of the variable in record r [23, 44, 48]. For example, let us assume that a variable is sorted, and that r is a value taken by that variable in a certain record. Then, the interval centered around value r consists of values rather than ranks, and so the lower bound of the interval equals the value with $Rank_{Lb} = Rank(r) - P\%$, and the upper value of the interval equals the value $Rank_{Ub} = Rank(r) + P\%$. A *Match* is said to have occurred when the values of all variables in the original record fall in the corresponding calculated intervals[8].

A confidence index for the Intruder for each interval size is evaluated by using this measure, and a smaller interval implies a lower value for the confidence. Normally, the average confidence is computed by using a certain fixed range of percentages (say, between 1% and 10%) of the records. Obviously, a large value of P implies that less information is revealed, and a large value for DR.

The overall global DR, given as G_{DR}, is defined as the average of the DR as computed by both the above strategies, and has the form:

$$G_{DR} = \frac{RLD + ID}{2}. \tag{4.3}$$

Overall Scoring Index: As we know, the indices of the IL and DR reflect completely different aspects of an MAT. Most researchers have concentrated on either of these criteria. While this is an accepted practice, we feel that a more fair index would be one which considers both of them simultaneously. In this vein, we define the Scoring Index, SI, to be a linear combination of the indices obtained for the IL and DR as:

$$SI = x\ G_{IL} + (1 - x)\ G_{DR}. \tag{4.4}$$

[8]At the extreme points, the bounds of the interval are computed as follows: For r_{min}, the value of $Rank_{Lb}$ is set at r_{min}, and $Rank_{Ub}$ is assigned the value $Rank(r_{min}) + P\%$. Analogously, for r_{max}, the value of $Rank_{Lb}$ is assigned at the value with $Rank(r_{max}) - P\%$, and $Rank_{Ub}$ is set at r_{max}.

Through out our studies, we have set the value of x to be 0.5, implying that we give equal weight to the IL and DR measures.

4.5 Experimental Results

4.5.1 Data Sets

The $OMMA$ has been rigorously tested and the results obtained seem to be very promising. We have tested it using the two real-life benchmark reference data sets used in previous studies and various simulated data sets[9] using *Matlab's* built-in-functions:

1. The *Tarragona* data set which contains 834 records with 13 variables [45].
2. The *Census* data set which contains 1080 records with 13 variables [53].
3. A uniform distribution (min=0; max=1000).
4. A normal distribution (μ=0; σ=0.05).
5. A uniform distribution (min=0; max=40,000) which contains $10,000$ records and 16 dimensions.
6. A normal distribution (μ=500; σ=150) which contains $6,000$ records and 16 dimensions.

In all the data sets, discrete variables were approximated as being *continuous.*

The data sets in items 3 and 4 above involved vectors with dimensions ranging from 12 up to 22, and sets of cardinality from $1,200$ up to $5,400$.

[9]The experiments were also done to investigate the scalability of the $OMMA$ with respect to the size of the data, its dimensionality, and the number of records per group.

It is worth mentioning that the generated simulated data sets have been generated to mimic real-life scenarios [102]. To achive this, the resulting simulated data sets, described in items 5 and 6 above, had two properties that were important to our experiments:

1. Key variables are necessary to estimate the disclosure risk using RLD. Therefore, the selection criterion of the key variables was based on choosing the minimum number of repetitions of values in each variable. More specifically, 3 key variables were chosen for the uniform data sets, while 5 key variables were chosen for the normal data set.

2. The number of records in each data set is based on the number of key variables. In general, the size of the simulated data should be low, as one would not expect repeated values for continuous variables. However, there were repetition in the data set. Our selection of $4,800$ records in the uniform data, and $1,560$ in the normal data set, because this was the cardinality of the set which corresponded to the largest integer which is a multiple of 3, 4, 5 and 6. Thus, the MAT could be invoked with a minimum group size of $k = 3$, 4, 5 or 6.

4.5.2 Results

For a given value of the minimum group size k, we compared the percentage value of the $IL = SSE/SST$ resulting from the $OMMA$ and the $MDAV$ strategies. It is important to mention that the $MDAV$ was implemented based on the centroid concept and not a diameter concept[10]. The memory value, M, was set to be 5 for all the experiments, since it yielded the best solution and a reasonable computational time. All the programs were written in the C^{++} language, and the tests were performed on an Intel(R) Pentium (R)M Processor with the clock speed of 1.73 GHz., and with 512 MB of RAM. The $OMMA$ was applied to the uni-variate data sets and the

[10]We did not program the $MDAV$ scheme. We are extremely thankful to Dr. Francesc Sebe for giving us his source code.

multi-variate data sets.

Table 4.2: Comparison of the percentage of the IL between the $MDAV$ and the $OMMA$ on the Tarragona and Census data sets (uni-variate Methods). In this case the value of k was set to be $k = 3$. The column "Converge" indicates the index of the cycle in the learning phase when the value of the total SSE does not change.

Tarragona Data Set				**Census Data Set**			
Variable	$MDAV$	$OMMA$		Variable	$MDAV$	$OMMA$	
	IL	IL	Converge		IL	IL	Converge
Var1	7.15200	7.15199	2	Var1	0.13155	0.13155	4
Var2	0.63586	0.63586	4	Var2	0.00138	0.00138	4
Var3	0.51702	0.51702	5	Var3	0.00828	0.00828	5
Var4	1.48854	1.48854	4	Var4	0.00489	0.00489	4
Var5	1.69394	1.69394	5	Var5	0.02449	0.02449	4
Var6	0.47503	0.47503	4	Var6	0.03262	0.03262	4
Var7	1.96623	1.96623	5	Var7	0.00171	0.00172	4
Var8	0.42182	0.42182	4	Var8	0.43418	0.43418	5
Var9	1.28625	1.28627	4	Var9	0.72176	0.72176	5
Var10	1.74929	1.74929	4	Var10	0.00611	0.00611	5
Var11	2.58368	2.58367	3	Var11	0.01353	0.01353	4
Var12	4.14703	4.14703	5	Var12	0.00689	0.00689	5
Var13	5.00563	5.00563	4	Var13	0.00808	0.00808	8

Table 4.2 shows a comparison between the $MDAV$ and the $OMMA$. In this table, both strategies have been applied on uni-variate data sets, for each of the "Tarragona" and "Census" data sets containing 13 continuous variables, and the value of k was set to 3 in all the experiments. The results clearly show that the $OMMA$ obtains values of the IL exactly same as those obtained by the $MDAV$ scheme. The reason for this appears to be the uniqueness of the optimal solution in the uni-variate case when n is multiple of k. It is worth mentioning, that the $OMMA$ reaches the minimum value of the IL with an average of 4 successive learning cycles in the Tarragona data set, while 5 cycles are required, on the average, to reach the minimum value of the IL in the Census data set. The computation time required to micro-aggregate each variable independently was as low as 1.93 seconds (on the average) for the Tarragona data set, while for the Census data set it was, on the average, about 2.93 seconds. We also can unequivocally conclude that for uni-variate micro-aggregation the time required

to micro-aggregate all the individual records to reach the minimum IL increases with respect to the data size. Besides, the number of learning cycles, required to lead to the minimum value of the IL, is proportional to the computation time.

Table 4.3: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $OMMA$ on the real-life data sets (Tarragona and Census), and simulated data sets (Uniform and Normal distributions) for multi-variate methods.

Data Set	k value	*MDAV*		*OMMA*			Improv. (%)
		IL	Time	IL	Time	Converge	
Tarragona	2	9.2750	0.28	9.2435	5.44	7	0.34%
	3	16.9661	0.20	15.1290	11.69	5	10.83%
	6	26.4047	0.11	24.2609	36.08	5	8.12%
Census	3	5.6535	0.33	5.2290	20.16	5	7.51%
	4	7.4414	0.25	6.7623	30.00	4	9.13%
	5	8.8840	0.27	8.0900	48.84	4	8.94%
	6	10.1941	0.24	9.1429	63.97	4	10.31%
Uniform distribution	3	22.4608	4.18	21.3214	33.50	7	5.34%
	4	29.1714	4.45	26.2564	60.21	7	11.10%
	5	33.9636	4.62	29.8415	70.11	7	13.81%
	6	37.1577	4.75	32.8529	125.59	7	13.10%
Normal distribution	3	26.9348	0.45	25.1386	1.15	6	7.15%
	4	33.9256	0.78	30.7705	1.80	6	10.25%
	5	39.564	0.59	35.1803	2.03	6	12.46%
	6	43.4592	0.60	38.5145	2.23	6	12.84%

The power of the $OMMA$ is clearly evident when it is used on multi-variate data sets, since the $OMMA$ has the ability to measure the similarity between the individual records based on two different criteria. The first criterion, C_1, involves studying the relation between the records quantified in terms of the distance between each record and the other records. As opposed to this, the second criterion, C_2, attempts to maintain the SSE as low as possible. The $OMMA$ has the ability to prioritize these two criteria, and to reflect the corresponding evaluation in the process of awarding rewards and penalties. In the case of a reward, the potential of having any two records in the same group is increased by moving both of them, one state towards the most internal states. Thus, in this setting, C_1 is given a higher priority than C_2. Similarly, in the scenario of a penalty (see Case 1), processing the tuple

and moving both records one state towards the boundary state having the minimum certainty (while preserving the minimum value of SSE), again lends C_1 a higher priority than C_2. But, in Case 2 and Case 3 when a penalty response is processed, the $OMMA$ forces the corresponding records to migrate to groups that preserve the number of records in each group, and to move towards the condition in which the absolute minimum value of the SSE is obtained. This has the effect of giving C_2 a higher priority than C_1.

Table 4.3 shows the results of using the $OMMA$ on multi-variate data sets, where we micro-aggregate *all* the individual records with all the variables simultaneously. In this case, we have tested the $OMMA$ with different values of k in order to determine the effect of increasing the number of records per group. It seems to be apparent that increasing the number of records per group tends to increase the value of the IL besides increasing the required computational time.

Table 4.3 shows that the values of the IL measured for the $OMMA$ were less than the corresponding values measured for the $MDAV$. The percentage of improvement in the IL was as high as 10.83% in the Tarragona data set when $k = 3$, requiring only 11.68 seconds. Similarly, in the Census data set the improvement was as high as 10.31% when $k = 6$, requiring 63.97 seconds. But in case of simulated data sets, the improvement in the IL reached up to 13.81% when $k = 5$ in the uniform data set, and it was as high as 12.84% when $k = 6$ in the normal data set. In term of comparison, we believe that minimizing the loss in the data utility is more important than minimizing the computational time, especially because micro-aggregation is usually performed off-line where additional time resources are less stringent. However, the question of how the decrease in IL is related to the increase in computation time is still open.

To be consistent with the state-of-the-art, we have also compared our IL values with the corresponding loss values yielded by the MST in [84]. In the Tarragona data set, the percentage improvement in the IL was up to 3.02% when $k = 3$. Similarly, in the Census data set, the percentage improvement in the IL reached up to 3.52% when $k = 3$, 6.34% when $k = 4$, and 8.17% when $k = 5$. This clearly demonstrates that

the $OMMA$ competes in a superior manner against all the state-of-the-art strategies. The authors of [84] had suggested that a MST partitioning-based method should be considered as a potential candidate for any practical application. In this context, we recommend using the $OMMA$ in micro-aggregating a multi-variate data set.

Table 4.4: Scoring the $MDAV$ and $OMMA$ with respect to the G_{IL} and G_{DR} by computing the index SI, for $k = 3$, 4, and 5 by using Census data set and the simulated Uniform and Normal distributed data sets.

Data	*Creterion*	$k = 3$		$k = 4$		$k = 5$	
		$MDAV$	$OMMA$	$MDAV$	$OMMA$	$MDAV$	$OMMA$
Census	G_{IL}	28.61000	22.81600	33.25400	28.40800	38.67200	38.62400
	RLD	60.71330	63.14640	49.70570	53.53510	42.09810	48.58860
	ID	1.98148	2.25926	0.73148	1.02778	0.26852	0.61111
	G_{DR}	31.34739	32.70283	25.21859	27.28144	21.18331	24.59986
	SI	33.80500	30.90800	36.12700	33.70400	38.83600	38.81200
Uniform	G_{IL}	115.43193	41.28212	116.11979	45.16758	120.91989	56.38516
	RLD	0.10880	1.09259	0.06944	0.86343	0.06019	0.55093
	ID	0.00417	4.33611	0.00000	1.25625	0.00000	0.50069
	G_{DR}	0.05648	2.71435	0.03472	1.05984	0.03009	0.52581
	SI	57.74420	21.99824	58.07725	23.11371	60.47499	28.45549
Normal	G_{IL}	108.98124	76.07109	89.03977	40.14807	110.86944	51.84437
	RLD	0.64936	10.74100	0.65705	7.34936	0.2814	5.40449
	ID	0.00855	1.57692	0.00000	0.39103	0.00000	0.09829
	G_{DR}	0.32895	6.15896	0.32853	3.87019	0.14071	2.75139
	SI	54.65510	41.11503	44.68415	22.00913	55.50507	27.29788

The other set of experiments were carried to test the SI of $MDAV$ and $OMMA$. Therefore, they have been scored as per the SI index on the simulated data sets and the Census[11] data set, which contains seven key variables: $Var1$, $Var2$, $Var3$, $Var5$, $Var10$, $Var11$, and $Var12$.

Table 5.2 displays the SI for the $MDAV$ and $OMMA$ methods for various values of k, which was set to be either 3, 4 or 5 as per the accepted requirements stated in [46, 115, 140]. As per Equation (4.2), the G_{IL} was computed by averaging the values of M_1, M_2, M_3, M_4, and M_5. In general, the value of the G_{IL} is directly proportional

[11]Scoring them against the Tarragona data set is meaningless because the latter contains no so-called key variables.

to the number of records per group represented by the value, k. Therefore, in the Census data set, the best value of G_{IL} for the $OMMA$ was obtained when $k = 3$, and was equal to 22.816%, while the best value of G_{IL} for the $MDAV$ was 28.610% when $k = 3$. In the context of the simulated data sets, and in general, the value of G_{IL} for the $OMMA$ is almost half the value that was obtained by using the $MDAV$ (*i.e.,* the value of G_{IL} for the $OMMA$ in the uniform data set and when $k = 3$ was 41.28%, while it was equal to 115.43% for the $MDAV$). Clearly, the $OMMA$ method preserves the data utility more efficiently than the state-of-the-art, $MDAV$.

We also compared the G_{DR} (which estimates the risk of data being disclosed) using the RLD and ID techniques. The RLD was computed using distance-based metric computations, where the average value was computed over several scenarios. Each scenario computed the average of the RLD over all possible $\binom{S}{C} = \frac{S!}{C!(S-C)!}$ combinations, where S is the number of key variables in the micro-data file, and C is the number of chosen variables that are known to the Intruder. The results (Table 5.2) generally shows that estimating the risk of disclosing the secure information using the RLD method is inversely proportional to the number of records per group represented by k. Thus, in the Census data set, the percentage value of the RLD for the $MDAV$ was as low as 60.71% when $k = 3$, 49.70% when $k = 4$, and 42.10% when $k = 5$, while it was 63.15% when $k = 3$, 53.54% when $k = 4$, and 48.59% when $k = 5$ for the $OMMA$. Similarly, simulated data sets report the same fact which is that the $MDAV$ method is more secure than the $OMMA$ method in estimating the RLD, for example, the percentage value of the RLD by using the $MDAV$ on the uniform data set was around 0.1%, while it was around 1.0% by using the $OMMA$. On the other hand, the percentage value of the ID was calculated as the average value for various settings of P, $(1, 2, \ldots, 10\%)$ on Census data set, when the value of k was assigned to be 3, 4, and 5. We observe that the minimum percentage value of the ID was equal to 0.26852%, which was achieved by invoking the $MDAV$ when $k = 5$. As opposed to this, the minimum percentage value of the ID for the $OMMA$ was almost comparable – to 0.6111% when $k = 5$. In the simulated data sets, the percentage value of the ID was equal to *zero* for the $OMMA$ and $MDAV$. Therefor, the range

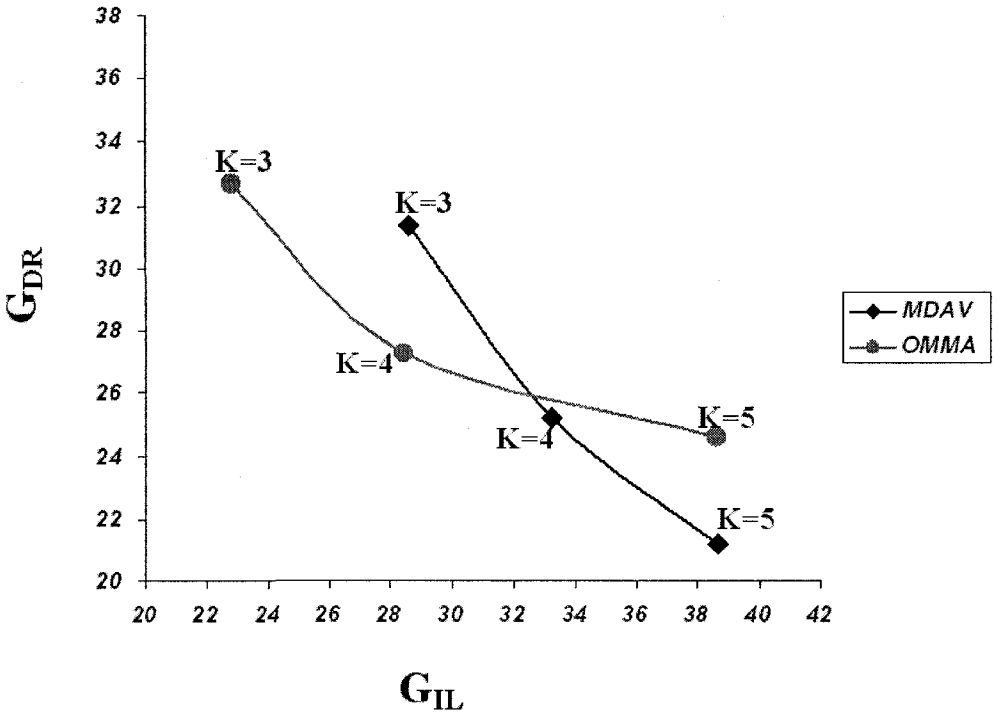


Figure 4.6: The effect of invoking the $MDAV$ and $OMMA$ on the G_{IL} and G_{DR} indices when $k = 3$, 4 and 5 for Census data set.

of the p values was extended to reach up to 30% of the original size of the data sets. The percentage value of the $MDAV$ remained *zero*, while the minimum percentage value for the $OMMA$ reached up to 0.5% and 0.098% when $k = 5$ in the uniform and normal data sets, respectively.

Finally, the SI value was computed for each MAT based on Equation (4.4). Observe that a lower score value implies a superior performance. From the table we see that the $OMMA$ technique has, almost consistently, a better performance index than the $MDAV$ not only from the G_{IL} perspective, but also as from the perspective of a combination of the G_{IL} and G_{DR} for different values of the k. Thus, for example, in Census data set the $OMMA$ method scores the minimum value that was equal to 30.91 when $k = 3$, and to 33.704 when $k = 4$. Besides the $OMMA$ scores almost half

the SI value which was obtained by the $MDAV$ by using the simulated uniform and normal data sets.

An interesting exercise involves studying how the $OMMA$ and the $MDAV$ compared when "all" the factors coming to play. In other words, it would be good if we could understand the conditions when one algorithm is superior to the other and vice verse. While this can be done in many ways, we have plotted the G_{IL} versus the G_{DR} for the Census data set, in Figure 5.2, for both schemes. This figure presents sets of paired values of G_{IL} and G_{DR} for the respective algorithm for different values of k ranging from 3 to 5. The user should observe the influence of a masked method as the value of k changes. From the curve we see that the $OMMA$ optimizes these conflicting criteria in a superior way than the stat-of-the-art $MADV$ method, because as the value of k increases, the increase of G_{IL} does not effect the value of G_{DR} in the $OMMA$ method as it does in the case of the for $MDAV$.

Finally, we discuss the performance of the $OMMA$ and $MDAV$ with respect to three other issues.

1. **The scalability of the *OMMA* with respect to cardinality.**
   We have tested both the schemes on the uniform and the normal distributions with data set sizes cardinalities 1200, 2400, 3000, 4500, and 5400, with 16 variables and $k = 3$, as shown in Table 4.5. The impact of increasing the size of the data set leads to minimizing the IL value, increasing the computational time and, in the case of the $OMMA$, increasing the number of learning cycles required to reach to the minimum value of the IL. The percentage of the improvement in the IL that the $OMMA$ obtains, ranges from 6.69% to 3.03% when the data size equals $1,200$ and $4,500$, respectively, for the uniform distribution, and for the normal distribution, this ranges from 5.46% when the data size equals $1,200$ to 3.03% when the data size is $4,500$. As the reader can observe, Figures 4.7 and 4.8 display similar trends for minimizing the value of the IL as a function of the data set size for uniform and normal distributions, respectively.

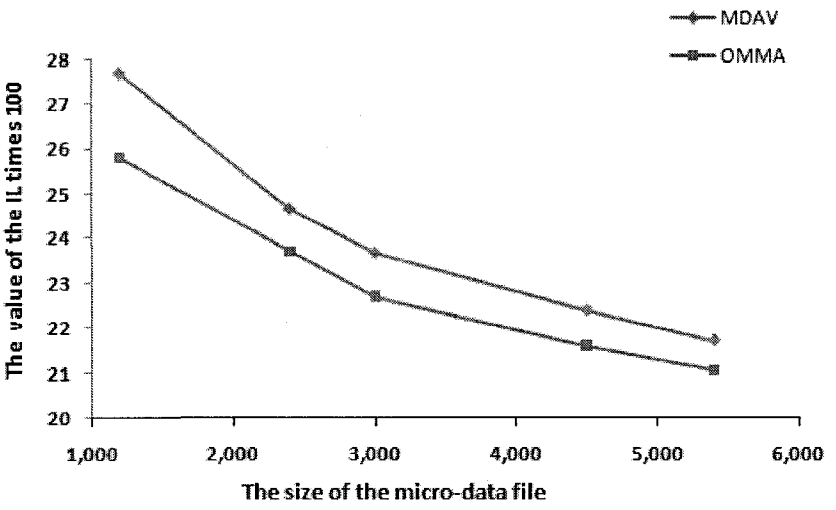


Figure 4.7: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the cardinality of the data set using uniformly distributed data when $k = 3$.

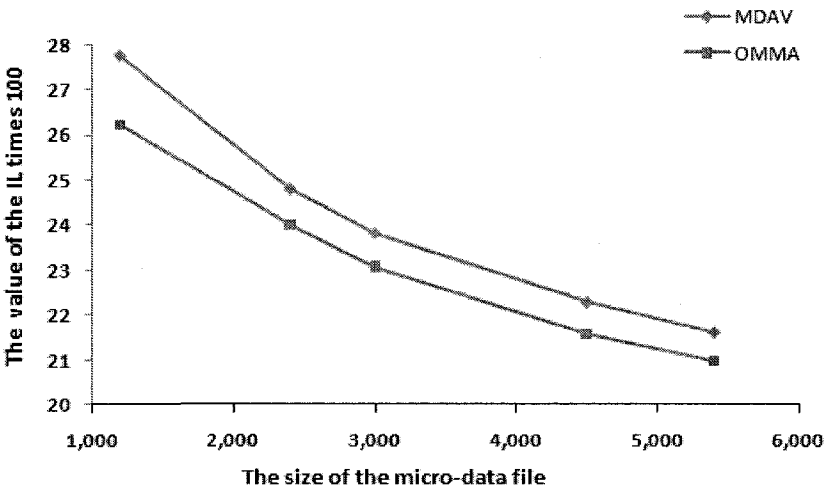


Figure 4.8: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the cardinality of the data set using normally distributed data when $k = 3$.

Table 4.5: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $OMMA$ on simulated uniform and normal data sets for multi-variate methods. This table investigates the scalability with respect to data size.

Data Set	No. of Records	No. of Groups	Dim.	k value	$MDAV$		$OMMA$			Improv. (%)
					IL	Time	IL	Time	Converge	
Uniform Distribution	1,200	400	16	3	27.67671	0.42	25.82418	0.66	6	6.69
	2,400	800			24.64686	1.28	23.71665	5.13	6	3.77
	3,000	1000			23.66435	2.00	22.69465	9.88	8	4.10
	4,500	1500			22.38678	4.45	21.60147	21.64	7	3.51
	5,400	1800			21.71397	6.42	21.05564	33.05	7	3.03
Normal Distribution	1,200	400	16	3	27.74187	0.36	26.22697	0.65	7	5.46
	2,400	800			24.80187	1.28	24.00719	3.76	6	3.20
	3,000	1000			23.82211	1.99	23.06686	6.45	7	3.17
	4,500	1500			22.29773	4.45	21.57716	16.23	6	3.23
	5,400	1800			21.62495	6.46	20.96986	29.97	9	3.03

2. **The scalability of the *OMMA* with respect to dimensionality.**
   We have also tested both strategies on the uniform and the normal distributions for different numbers of variables, including 12, 14, 16, 18, 20, 22 and 24, when the data size was set to $3,000$ and the value of k was set to 3, as shown in Table 4.6. Again, we observe that, the value of the IL is proportional to the number of the variables used in the micro-aggregation process. Increasing the dimensionality implies increasing the loss in the utility of the information. This makes sense, because, informally speaking, increasing the dimensionality leads to minimizing the similarity between the individual records and, at the same time, reducing the correlation between the different multi-variate records. The interesting point here is that the computational time required to micro-aggregate all the individual records seems to be inversely proportional to the dimensionality in the $OMMA$ scheme, but proportional for the $MDAV$ scheme. This can be justified as follow: In the $OMMA$ case, increasing the dimensionality seems to minimize the within-group similarity, and simultaneously reduce the correlation between them. This, thus, leads to a smaller size for the similarity list structures, which in turn, obviously, requires less time to process all the tuples. As opposed to this, in the case of the $MDAV$, increasing the dimensionality

requires more time to achieve the computation of the distances between the multi-variate records. The highest percentage of the improvement in the IL is 4.04% when the number of variables is 16 for the uniform distribution, while for the normal distribution, the highest percentage advantage is 4.52% when the number of variable is 12. It should be mentioned that, the $OMMA$ works perfectly when the number of variables used to micro-aggregate the data set is moderate and not too high. Thus, for the uniform distribution with the dimensionality of 22, and the normal distribution with a dimensionality of 24, the $OMMA$ is faster than the $MDAV$, while the value of the IL increases. Figures 4.9 and 4.10 display similar trends for minimizing the value of the IL as a function of the dimension of the data set for uniform and normal distributions, respectively. It is worth mentioning that the value of the IL is proportional to the dimension of the variables used in the micro-aggregation process.

Table 4.6: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $OMMA$ on simulated uniform and normal data sets for multivariate methods. This table investigates the scalability with respect to the data dimensionality.

Data Set	No. of Records	No. of Groups	Dim.	k value	*MDAV*		*OMMA*			Improv. (%)
					IL	Time	IL	Time	Converge	
Uniform Distribution	3,000	1000	12	3	17.55012	1.61	16.92778	20.75	7	3.55
			14		20.80110	1.81	20.05538	12.38	6	3.59
			16		23.64962	2.00	22.69465	9.88	8	4.04
			18		26.07906	2.23	25.23600	6.25	7	3.23
			20		28.59153	2.45	27.45749	4.17	7	3.97
			22		30.51047	2.67	29.41102	1.97	6	3.60
Normal Distribution	3,000	1000	12	3	17.65475	1.59	16.85623	29.66	6	4.52
			14		20.93491	1.81	20.30601	18.66	9	3.00
			16		23.82211	1.99	23.06686	6.47	7	3.17
			18		26.29432	2.24	25.36438	5.03	7	3.54
			20		28.62054	2.42	27.56436	3.67	7	3.69
			22		30.57049	2.64	29.41102	2.97	6	3.79

3. **The scalability of the *OMMA* with respect to the number of records per group.**

   The scalability of the $OMMA$ and the $MDAV$ regarding the group size has

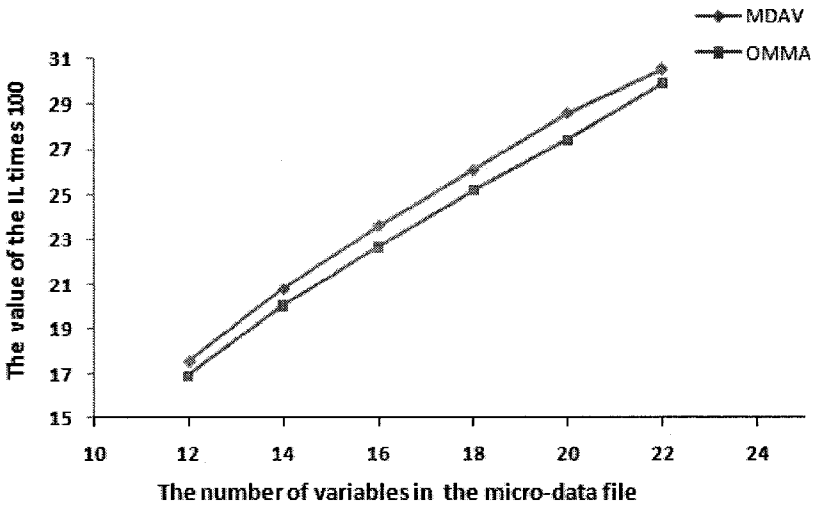


Figure 4.9: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the dimension of the data set using uniformly distributed data when $k = 3$ and $n = 3,000$.

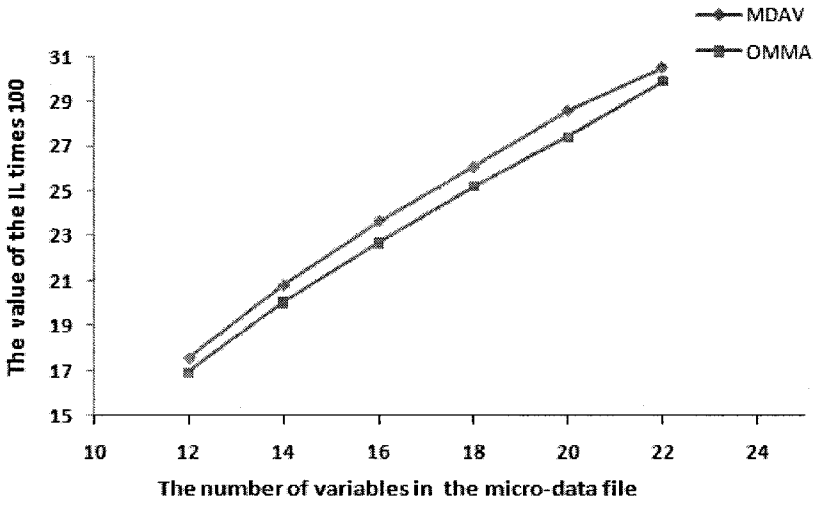


Figure 4.10: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the dimension of the data set using normally distributed data when $k = 3$ and $n = 3,000$.

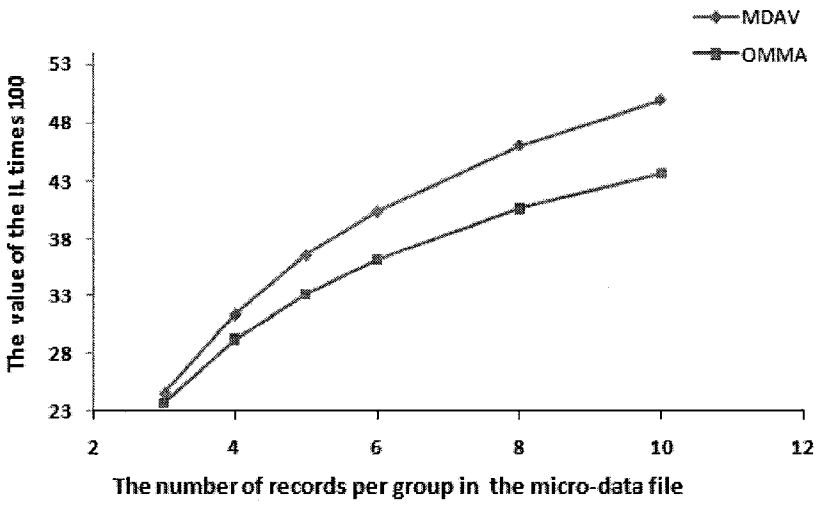


Figure 4.11: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the number of records per group using uniformly distributed data when $n = 2,400$ and $d = 16$.

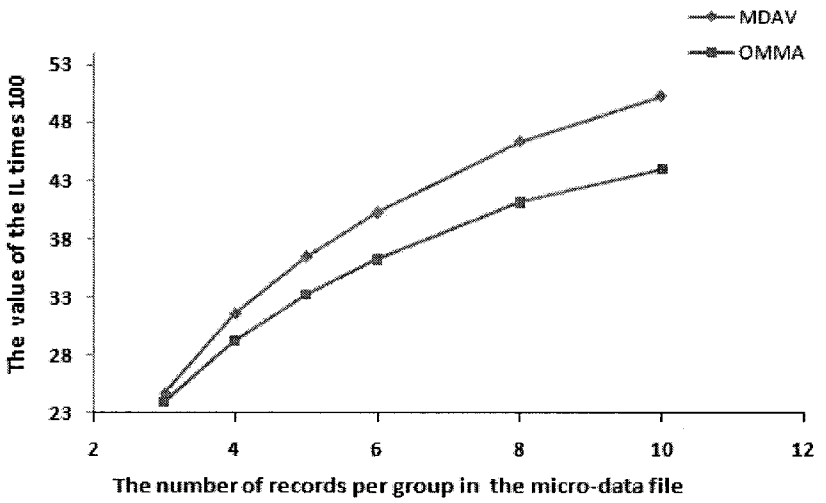


Figure 4.12: The improvement of the $OMMA$ in reducing the percentage value of the IL as a function of the number of records per group using normally distributed data when $n = 2,400$ and $d = 16$.

Table 4.7: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $OMMA$ on simulated uniform and normal data sets for multivariate methods. This table investigates the scalability with respect to the number of records per group.

Data Set	No. of Records	No. of Groups	Dim.	k value	$MDAV$ IL	$MDAV$ Time	$OMMA$ IL	$OMMA$ Time	$OMMA$ Converge	Improv. (%)
Uniform Distribution	2,400	800	16	3	24.64686	1.27	23.71665	5.13	6	3.77
		600		4	31.39147	1.28	29.17778	7.28	7	7.05
		480		5	36.55690	1.30	33.12187	10.42	7	9.40
		400		6	40.37307	1.30	36.19842	15.89	9	10.34
		300		8	46.01553	1.30	40.65105	19.64	7	11.66
		240		10	49.91325	1.30	43.61896	28.50	7	12.61
		200		12	53.22947	1.30	46.16533	37.58	7	13.27
Normal Distribution	2,400	800	16	3	24.80187	1.28	24.00719	3.71	6	3.20
		600		4	31.68612	1.31	29.27261	8.19	10	7.62
		480		5	36.54596	1.33	33.24953	8.23	8	9.02
		400		6	40.23421	1.31	36.30028	13.55	8	9.78
		300		8	46.32321	1.30	41.14944	21.86	9	11.17
		240		10	50.22509	1.33	43.99177	27.86	7	12.41
		200		12	52.78059	1.32	46.64860	37.38	7	11.62

also been studied for both the uniform and the normal data sets, where the group sizes were 3, 4, 5, 6, 8, 10, and 12, and for the data set cardinality of $2,400$, with the dimensionality of 16, as shown in Table 4.7. Here, the value of the IL increases with the number of records per group. This is a fair observation, because having many records in the group tends to minimize the within-group similarity, and to simultaneously maximize the similarity between the groups. This also tends to increase the value of the IL. In the $OMMA$ the computational time required to micro-aggregate the records increases with the number of records per group. This too can be justified because when the $OMMA$ forces the records to migrate to another group, choosing the record is not an easy job, as it depends on a comprehensive study of the effect of migrating every other possible record- which equals $(k-1)$ possibilities for Case 2, and $2(k-1)$ possibilities for Case 3. As opposed to this, the computational time is fixed in the $MDAV$ strategies (because it depends only on the size of the data set and the dimensionality). Thus, obviously, increasing the group size

tends to increase the efficiency of the $OMMA$ by leading to a minimum value of the IL, as opposed to the $MDAV$. In the uniform data set the percentage of improvement for the IL is as high as 13.27%, which is obtained when the group size equals 12. As opposed to this, this improvement reaches 12.41% for the normal data set, when the group size equals 10. Figures 4.11 and 4.12 show that both schemes possess similar trends for the IL as a function of the group size, and that the value of the IL is proportional to the number of records per group.

4.6 Conclusions

In this chapter we have presented, to our knowledge, the first reported LA-based solution to the MAP, which is known to be NP-hard. We have shown that our newly devised scheme, the $OMMA$ can successfully be used to micro-aggregate a multi-variate micro-data file. The $OMMA$ competes in a superior manner to the state-of the art $MDAV$ and MST methods in minimizing the loss in the information for such data. The percentage of improvement reaches up to 13% when compared to the $MDAV$ scheme on real-life and simulated data sets. The proposed strategy also scales well with respect to the size of the data set, the dimensionality, and the group size. By defining a score index, SI, as a composite measure involving the IL and DR, we see that proposed strategy also obtains a minimum score value when compared to the $MDAV$ method. This indicates that the $OMMA$ technique is probably the best MAT not only from the IL perspective, but also from the viewpoint of a measure which is as a combination of the IL and DR. Therefore, the $OMMA$ can thus be highly recommended for advantageous micro-aggregation.

Chapter 5

Using Neural Methods to Solve the MAP

5.1 Introduction

This chapter[1] presents a possibly pioneering endeavour to tackle the MAP in secure statistical databases, by resorting to the principles of associative Neural Networks (NN). The prior art has improved the available solutions to the MAT, by incorporating proximity information, and this is done by recursively reducing the size of the data set by excluding points which are farthest from the centroid, and those which are closest to these farthest points. Thus, although the method is extremely effective, arguably it uses only the proximity information, while ignoring the mutual Interaction between the records. In this chapter, we argue that inter-record relationships can be quantified in terms of two entities, namely their "Association" and "Interaction". This means that the records which are not necessarily close to each other may

[1] A preliminary version of some of the results from this chapter appeared in the *Proceedings of ICICS'07, the 9th International Conference on Information and Communications Security*, in Zhengzhou, China, in December 2007 [107]. The journal version of these results is currently under review.

still be "grouped" because their mutual Interaction could be significant. Based on the theoretically sound principles of NNs, we believe that the proximity information can be quantified using the mutual Association, and their mutual Interaction can be quantified by invoking transitive-closure like operations on the latter. By repeatedly invoking the inter-record Associations and Interactions, the records are grouped into sizes of cardinality "k", where k is the security parameter in the algorithm. Our experimental results, which are done on artificial data and on the benchmark real-life data sets, demonstrate that the newly proposed method is superior to the state-of-the-art not only from the IL perspective, but also when it concerns a criterion involving a combination of the IL and the DR.

The structure of this chapter is as follows: We first present a full description of the Associative Cluster Neural Network algorithm, which is used in unsupervised classification, in Section 5.2. In Section 5.3 the Interactive-Associative Micro-Aggregation Technique is presented informally and algorithmically. Then, in Section 5.4, we present the results of experiments we have carried out for synthetic and real data sets. The chapter finishes in Section 5.5 with some conclusions.

5.1.1 Contribution of the Chapter

As mentioned in Chapter 2 and 4, this problem in its multi-variate setting is known to be NP-hard [103], and has been tackled using different approaches such as hierarchical clustering [45, 91, 92], genetic algorithms [45, 91, 92, 128], graph theory [74, 84], fuzzy clustering [57, 133] and machine learning [64]. As we know from our previous discussions, all the reported heuristic $MATs$, seek to minimize the value of the IL. Although minimizing the loss in the data utility is an important issue, we again argue that maintaining a happy medium between the IL and DR is also crucial [27, 49, 60, 93]. In this chapter, we shall argue that a good MAT should be capable of minimizing the IL, and yet be able to attain to suitable value for a well-defined composite measure, SI (as mentioned in Chapter 4). Our aim is to find a good strategy to optimize the latter. Observe that the DR in an MAT is solely determined

by the minimum group size parameter k, namely, by ensuring k-anonymity, where each data subject is distinguishable from at least $k-1$ other data subjects [39, 60].

In general, minimizing the IL directly follows maximizing the similarity between records in each group. The state-of-art $MATs$ depend on utilizing the "Euclidean" distance, which serves as the criterion playing a central role in estimating the similarity between the records. However, this distance function does not completely capture the appropriate notion of similarity for any data set. Our position is that the notion of similarity should be measured by using a metric that also unravels the relationship between the inter-records. We believe that this can be quantified in terms of two quantities, namely the mutual "Association" between the individual records and their mutual "Interaction". We propose to measure these quantities using Association Similarity Rules[2] ($ASRs$). In this context, we mention that the concepts of Association and Interaction are derived from the Associative Cluster Neural Network ($ACNN$), which estimates the similarity between neurons by building a dynamic model evaluated through the interaction between the neurons inside each group, and the interaction among the groups themselves.

The main contribution of this chapter is to integrate the foundational concepts of $ASRs$ with $MATs$ so as to devise a new strategy for estimating the similarity. This new method demonstrates that the IL can be reduced taking two measurements into consideration. First of all, we consider the mutual Association between the records. Secondly, and more importantly, we also consider the mutual Interaction between the records by using a transitive-closure like operation when $k \geq 3$. This, in turn, is achieved by invoking our newly proposed Interactive-Associative Micro-Aggregation Technique ($IAMAT$). The effect of these considerations can be seen to minimize the IL by up to 13% when compared to the state-of-the-art. Apart from this, the $IAMAT$ also yields the best reported values for the above mentioned index, the SI. We argue that the applicability of the new strategy in estimating the similarity provides a promising strategy to effectively protect sensitive data in the micro-data

[2]Association Similarity Rules are well-known data mining techniques used to discover the relationships between patterns in different application domains [6, 7, 9, 68, 87, 116].

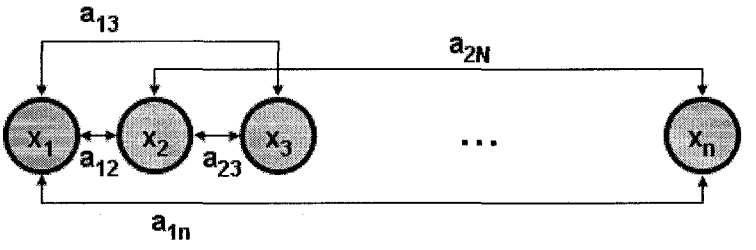


Figure 5.1: The structure of ACNN

file not only based on minimizing the value of the IL but also on offering the best trade-off between the IL and DR.

5.2 Associative Clustering Neural Network ($ACNN$)

The Associative Cluster Neural Network ($ACNN$) was proposed [148] as a recurrent NN model that dynamically evaluates the Association of any pair of patterns through the Interaction between them and the group of patterns. The $ACNN$ possesses many attractive features, such as its simple structure, its respective learning mechanism, and its efficient clustering strategy, which uses the Association as a new similarity measure. Its superiority in clustering and analyzing gene expression data has also been demonstrated [149]. The rationale behind this superiority probably lies in the inherent advantages of $ASRs$, that possess the potential to ensure that the similarities between patterns within the same cluster increase, whereas the similarities between different clusters decrease.

The $ACNN$ initializes the Association between any two neurons by evaluating the relationship between them and by setting the learning ratio, α, to the most suitable value. The learning ratio should guarantee that the initial association is large when the distance (*i.e.*, proximity in the feature space) between the patterns is small. The $ACNN$ studies the Interaction level of each pair of patterns based on the

Association made by the other patterns, and defines the similarity threshold which ensures a robust performance. The association value between any two patterns has to be updated based on the result of the Interaction level, and this is, in turn, scaled by using the well-known sigmoid function. This procedure has to be iteratively executed until there is no noticeable change in the successive Associations. Subsequently, the $ACNN$ constructs the cluster characteristic matrix to describe the cluster property at the end of the learning phase, after which it determines the number of clusters and labels the patterns with the index of the cluster that they belong to. We describe the structure and operation of the $ACNN$, more formally, below.

The structure of the $ACNN$ is depicted in Figure 5.1, which clusters the patterns based on evaluating the similarities between them. In the $ACNN$, each pattern is represented by a neuron, while the similarity between them is measured by their mutual Association denoted as a_{ij}, and which satisfies the following conditions:

- $a_{ij} = a_{ji}$.
- $\begin{cases} a_{ij} > 0 & : \ X_i \text{ and } X_j \text{ are associated} \\ a_{ij} \leq 0 & : \ X_i \text{ and } X_j \text{ are unrelated.} \end{cases}$

The $ACNN$ initializes the association between any two neurons, say X_i and X_j, as follows:

$$a_{ij}(0) = r(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{\alpha}}, \tag{5.1}$$

where X_i and X_j are two patterns, $r()$ is a function that evaluates the relationship between any two patterns, and α reflects the decay coefficient that uses the Mean Square Distance in the exponential function, which, in turn, is utilized to determine the initial association a_{ij} between them. The value of α is assigned so as to guarantee that the initial association is large when the distance between X_i and X_j is small and vice versa. Typically, it is given by:

$$\alpha = \frac{\sqrt{n}}{\frac{1}{n}(\sum_{i=1}^{n} ||X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)||^2)}. \tag{5.2}$$

Learning the successive Association between the patterns is done through a number of learning cycles (or epochs). In the cycle with index 't', the training of the NN simulates the Interactions of each pair $\langle X_i, X_j \rangle$, and this is based on the Associations made by the other patterns on each of them. In other words, for a pattern X_p, the $ACNN$ attempts to associate X_i and X_j with values a_{pi} and a_{pj}, respectively. If the values of the Association are greater than 0, X_p has the potential to be in the same cluster containing the pair $\langle X_i, X_j \rangle$, where the latter is given as:

$$y_{ij}(t) = \begin{cases} \sum_{p=1}^{n} a_{ip}(t-1) \bigotimes a_{pj}(t-1) & p \neq i,j \text{ and } i \neq j \\ 0 & i = j, \end{cases} \tag{5.3}$$

and in which

$$u \bigotimes v = \begin{cases} u \times v & u > 0 \text{ and } v > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

The association value between any two patterns has to be updated based on the result of the Interaction level, as follows:

$$a_{ij}(t) = \begin{cases} \varphi(y_{ij}(t)) & i \neq j \\ 1 & i = j, \end{cases} \tag{5.5}$$

where φ is the sigmoid function given as:

$$\varphi(y_{ij}(t)) = \frac{e^{y_{ij}(t)-\theta_{ij}(t)} - 1}{e^{y_{ij}(t)-\theta_{ij}(t)} + 1}, \tag{5.6}$$

and where θ is a similarity threshold that ensures that the characteristics of the unsupervised learning are preserved. The interesting issue here is not one of having a fixed threshold. Rather, each threshold value is derived by making use of the means of the mutual Interactions over all the data points to adjust the sigmoid function, and to thus attain a robust performance. To achieve this, θ_{ij} is defined as:

$$\theta_{ij}(t) = \frac{1}{2(n-1)} \sum_{p=1}^{n} (y_{pi}(t) + y_{pj}(t)). \tag{5.7}$$

All the above equations ensure that the similarities within the same cluster increase, whereas the similarities among different clusters decrease as the dynamic system evolves.

The learning process is terminated if the change in the successive Associations is less than a user-defined threshold. In other words, the learning terminates when $E(t) \approx 0$, where,

$$E(t) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}(t) - a_{ij}(t-1)|. \tag{5.8}$$

The cluster characteristic matrix, $\mathcal{M} = [m_{ij}]$, that describes the clustering property at the end of the $ACNN$ learning phase, is defined in terms of its components as:

$$m_{ij} = \begin{cases} 1 & a_{ij} > 0 \\ 0 & a_{ij} \leq 0. \end{cases} \tag{5.9}$$

It can be seen that each pattern X_i has its cluster characteristic vector $M_i = \{m_{i1}, m_{i2}, \ldots, m_{in}\}$ that records its cluster property by identifying the associated patterns in the same group. The number of clusters is determined as $C = rank(\mathcal{M}) - |\Psi|$, where $rank(\mathcal{M})$ is the rank of the matrix $\mathcal{M}$, and $|\Psi|$ is the cardinality of rejected patterns (Ψ), (*i.e.,* those which are isolated from all the other patterns), and is defined[3] as:

$$\Psi = \left\{ M_i \mid \sum_{p=1}^{n} m_{pi} = 1;\ i = 1, \ldots, n \right\}. \tag{5.10}$$

Each pattern will be labeled with the index of the cluster it belongs to. This is done by first excluding the rows of $\mathcal{M}$ which are repetitive, yielding the matrix $\bar{\mathcal{M}}$ of size $C \times n$. The rows of $\bar{\mathcal{M}}$ satisfy:

$$\bar{\mathcal{M}}_L = \{M_i \mid \sum_{p=1}^{n} m_{pi} > 1;\ M_i \neq M_j \quad i, j = 1, \ldots, C;\ i \neq j\}. \tag{5.11}$$

[3]This is equivalent to: $\Psi = \{M_i \mid m_{ii} = 1;\ m_{pi} = 0 \ \forall p \neq i\}$.

Any pattern X_i in the data set will be labeled as

$$\Theta(X_i) \begin{cases} j & \text{if } \exists M_j \in \bar{\mathcal{M}}_L, \ s.t. \ \ m_{ji} = 1 \\ 0 & otherwise. \end{cases} \tag{5.12}$$

where Θ is a labeling function in which 0 represents no label, implying that X_i is a rejected pattern without an associated label.

The $ACNN$ learning mechanism is algorithmically described below in Algorithm 11.

5.3 Interactive-Associative Micro-Aggregation Technique ($IAMAT$)

The state-of-the-art $MATs$ use a proximity function, and in particular, the Euclidean distance, to measure the similarity between records. To the best of our knowledge, the combination of the Association *and* the Interaction between individual records has not been taken into consideration while micro-aggregating the data file. We now discuss how these two criteria are applicable to micro-aggregate the data file so as to further minimize the IL.

5.3.1 Inadequacy of Using the $ACNN$ Directly

Although the basic $ACNN$ dynamically evaluates the Association and the Interaction between the patterns, it is not directly applicable to the MAP in its virgin form. We shall now explain how the principles motivating the $ACNN$ can be extended to solve the MAP.

Algorithm 11 Associative Cluster Neural Network (ACNN)

Input: A set of patterns, where each pattern represents a neuron.
Output: The set of labeled patterns with the cluster number.
Method:

1: Compute the decay coefficient $\alpha = \frac{\sqrt{n}}{\frac{1}{n}(\sum_{i=1}^{n} ||X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)||^2)}$. α determines the decay in the exponential function in the association function given below.
2: Initialize the association value between every pair of neurons $a_{ij}(0) = e^{-\frac{||X_i - X_j||^2}{\alpha}}$.
3: $t = 0$.
4: **Repeat**
5: $t = t + 1$.
6: Stimulate all the interactions on each pair, say $\langle X_i, X_j \rangle$, as follows:

$$y_{ij}(t) = \begin{cases} \sum_{p=1}^{n} a_{ip}(t-1) \bigotimes a_{pj}(t-1) & p \neq i,j \text{ and } i \neq j \\ 0 & i = j. \end{cases}$$

where $u \bigotimes v = \begin{cases} u \times v & u > 0 \text{ and } v > 0 \\ 0 & \text{otherwise.} \end{cases}$

7: Compute the dynamic similarity threshold between every pair of neurons: $\theta_{ij}(t) = \frac{1}{2(n-1)} \sum_{p=1}^{n} (y_{pi}(t) + y_{pj}(t))$.
8: Compute the sigmoid function between every pair of neurons as $\varphi(y_{ij}(t)) = \frac{e^{y_{ij}(t) - \theta_{ij}(t)} - 1}{e^{y_{ij}(t) - \theta_{ij}(t)} + 1}$.
9: Update the association value between every pair of neurons based on the Associations made by other neurons linked with them as follow: $a_{ij}(t) = \begin{cases} \varphi(y_{ij}(t)) & i \neq j \\ 1 & i = j. \end{cases}$
10: Compute $E(t) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}(t) - a_{ij}(t-1)|$ as the change in successive Associations.
11: **Until**$(E(t) \approx 0)$
12: Compute the cluster characteristic matrix, which describes the cluster property at the end of the *ACNN* learning phase. It is built as follows: $m_{ij} = \begin{cases} 1 & a_{ij} > 0 \\ 0 & a_{ij} \leq 0. \end{cases}$
13: Determine the rejected patterns, $\Psi = \{M_i | \sum_{p=1}^{n} m_{pi} = 1 \quad i = 1, \ldots, n\}$. The rejected patterns do not associate with any other pattern.
14: Calculate the number of clusters $C = rank(\mathcal{M}) - |\Psi|$.
15: Select C different cluster characteristic vectors $\bar{\mathcal{M}}_L = \{M_i \mid \sum_{p=1}^{n} m_{pi} > 1;\ M_i \neq M_j \quad i,j = 1, \ldots, C;\ i \neq j\}$.
16: Label each pattern with the indices of the cluster it belongs to as: $\Theta(X_i) \begin{cases} j & \text{if } \exists\ M_j \in \bar{\mathcal{M}}_L,\ s.t.\ m_{ji} = 1 \\ 0 & otherwise. \end{cases}$
17: **Return** the labeled patterns with the index of the clusters they belong to.
18: **End Algorithm** *Associative Cluster Neural Network (ACNN)*

5.3.1.1 Feature Values

In a neural setting, the weights of the neurons are updated based on the relative relationship between them. These weights are usually updated by gradient-like considerations, so that a change in the weights leads to a better classifications or performance. Consequently, in all such networks (including the $ACNN$) the weights could be both positive or negative, quite simply because increasing the values of certain features may have a negative impact on the optimization problem.

As opposed to this, it is meaningless to have weights that are negative in the MAP. This is because the fundamental reason for tackling the problem is to determine how the records are associated with each other, and clearly, the concept of the records being negatively associated is irrelevant. Thus, if we are to use the principles of the $ACNN$ to solve the MAP, it is imperative that the weights are never allowed to become negative. Rather, we would prefer that they stay within the interval $[0, 1]$. This is achieved by, first of all, computing the initial Association using the distances, and secondly, by updating the Interactions only as long as they are non-negative. This is achieved by resorting to a single-shot training phase, as will be explained presently.

5.3.1.2 Ineffectiveness of Sigmoidal Mappings

When Minsky suggested the weakness of the Perceptron, he showed that it was incapable of resolving the basic XOR problem. However, the field of NNs received a huge "boost" when it was discovered that if these primitive neural units were cascaded and interconnected, the discriminant could be arbitrarily complex. To effectively model the switching and clipping effects in such complex domains, researchers introduced functions such as the sigmoidal function whose purpose was to transform the input space using highly non-linear mappings.

It is our position that such switching and clipping effects are not pertinent to the

study of the MAP. The reason for this is quite straightforward: The Associations and the Interactions between the records are, in one sense, related to their relative proximity, and we have no reason to believe that these quantities fall off or change *abruptly*. Rather, our experience is that these quantities vary smoothly with the relative proximity of the records.

5.3.1.3 Transitive-Closure-like Properties

Obtaining the set of shortest paths on a graph can be achieved by using a transitive-closure algorithm that traverses all the edges of the graph. In this case, the shortest paths are obtained by using the operation of "Addition" on the weights of the edges along any given path, and invoking the "Minimum" operation over all the possible paths. However, the underlying algorithm has been proven to be much more powerful if it is mapped using the properties of a semi-ring $(S, \bigoplus, \bigotimes)$, where (i) S is the set of elements (weights) associated with an edge, (ii) $\bigoplus$ represents an abstract "Addition" operation over the elements of S, and (iii) $\bigotimes$ represents an abstract "Multiplication" operation over the elements of S. In particular, if S is the set of reals, and $\bigoplus$ and $\bigotimes$ represent the arithmetic addition and product operations respectively, the transitive closure algorithm would lead to a matrix multiplication scheme, which is central in determining the multi-step Markov matrix for a Markov chain.

The basic $ACNN$ computes the Interaction between the neurons using the product involving a_{ip} and a_{pj}. The total two-step Interaction is thus, effectively, the contribution of the transitive-closure operation of the path from X_i to X_j via all the possible intermediate nodes, X_p. In our case, the issue of interest is *not* the total Interaction between the relevant nodes, but rather the node X^* which contributes to the maximal Interaction between X_i and X_j. Thus, unlike the $ACNN$, in our solution, we do not compute the sum of all the Interactions between the nodes. Rather, we report the one which is maximally interacting with the nodes already in the same cluster, say X_i and X_j. This is a fundamental difference between our present scheme and the $ACNN$, because it renders the computations both easier and faster, and is

yet able to coalesce the nodes based on the inferred interactions.

5.3.1.4 One-shot Training

The final difference between our present scheme and the $ACNN$ is the fact that we have resorted to a one-shot "training" mechanism. This is *atypical* for most NNs. Indeed, most NNs repeatedly run the NN over the data set till their respective weights converge. Some families of NNs (for example, the Adachi's network [1]) have been reported, which actually yield the final weights using a single pass over the data set.

In our case, we argue that repeatedly running the updating algorithm over the data set is superfluous. Rather, by initially computing the Associations, we are able (by the strategy mentioned in the earlier sub-section) to arrive at the best Interactions. The $ACNN$ requires that the set of associations are then re-computed. But, since these associations are computed based on the relative proximities of the records, and since the interactions are computed based on the latter, it is meaningless, in the case of the MAP, to re-compute the associations. Indeed, if we resorted to doing this, it would lead to weights that are negative which again, as argued above, is unacceptable. It would also lead to the "rejection" of many records - which is inappropriate for the MAP. Thus, in the $IAMAT$, the corresponding matrices are computed in a one-shot manner. Subsequent computations are required only after the learned groups of size k are removed in each learning cycle.

Based on the above principles, we now present the design of our newly-proposed scheme, the $IAMAT$.

5.3.2 Design of the $IAMAT$

We propose the $IAMAT$ to micro-aggregate the records in the data set by using a new methodology to evaluate the similarity between them. This similarity is intuitively

expressed by their inter-record relationships, and is estimated by measuring the "Association" and "Interaction" as *modeled* in the $ACNN$. The resulting measurements are similar to the ones that cluster the records based on the distance between them. Consequently, instead of merely assigning relatively "close" records to be in the same group, we choose to "estimate" the Association *and* the Interaction between them, and if the combination of these indexes is relatively high, we assign them to be in the same group. Otherwise, we determine that they should be in two different groups. We believe that using this pair of measurements will help to achieve a more robust performance than other existing measures, which is a claim that we have verified. From a top level, we can describe it as below.

The $IAMAT$ is a consequence of incorporating the above considerations into the elegant $MDAV$ strategy. Consider the $IAMAT$ for any specific value of k. The $IAMAT$ uses the centroid of the data set to relatively determine the least associated record, say X_r. Subsequently, we achieve a quick search to obtain the record that is most associated to X_r, say X_s. After this, we propose to choose $k-2$ records based on the mutual *Interaction* between each record inside the group and the remaining unassigned records. Consequently, the next step consists of creating a cluster that comprises of the associated pair $\langle X_r, X_s \rangle$ and the most interactive $k-2$ records. At the end of this stage, the cluster is micro-aggregated and removed from the original data set. The above steps are iteratively repeated until no more that $k-1$ records remain in the original data set. The $IAMAT$ terminates by assigning the remaining unassigned records to the last group. The scheme is algorithmically described below in Algorithm 12, after which each step is explained in greater detail.

Unlike the $MDAV$, instead of measuring the distance between the records, the $IAMAT$ utilizes the association as per the $ACNN$. The $ACNN$ classifies the records as being associated if the value of the Association index, a_{ij}, is positive. Otherwise the neurons will be classified as being unrelated, leading to its "rejection". Clearly, rejecting records will not comply with the spirit and goal of the MAP whose aim is to minimize the IL. We believe that an Association between any pair of records exists regardless of its value, and this could be very small when it is close to zero, or

Algorithm 12 Interactive-Associative Micro-Aggregation Technique ($IAMAT$)

Input: The original micro-data file, $\mathcal{D}$, that contians n unassigned records, and the parameter,k.

Output: The micro-aggregated micro-data file, $\mathcal{D}'$.

Method:

1: Compute the centroid of $\mathcal{D}$ as $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2: Compute the scaling factor α as $\alpha = \frac{\sqrt{n}}{\frac{1}{n}(\sum_{i=1}^{n} ||X_i - \mu||^2)}$.
3: Compute the association values between μ and each record, X_i, as $a_{\mu i} = e^{-\frac{||X_i - \mu||^2}{\alpha}}$.
4: Initialize the number of groups to zero.
5: **while** there are more than $(k-1)$ *Unassigned* records in $\mathcal{D}$ **do**
6: Increment the number of groups by unity.
7: Initialize the number of records inside the group to zero.
8: Select the least associated *Unassigned* record, X_r, to the centroid μ as follows $X_r = Min\ \ a_{\mu i}$ and mark it as *Assigned* record.
9: Compute the association values between X_r and each *Unassigned* record, X_i, in $\mathcal{D}$.
10: Select the most associated *Unassigned* record, X_s, to X_r as follows $X_s = Max\ \ a_{ri}$ and mark it as an *Assigned* record.
11: Compute the association values between X_s and each *Unassigned* record, X_i, in $\mathcal{D}$.
12: Add X_r and X_s to the group and increment the number of records inside the group by two units.
13: **while** the number of records inside the group is less than k **do**
14: **for all** *Unassigned* records, X_p, in $\mathcal{D}$ **do**
15: Initialize the Interaction of X_p, η_p, to 1.
16: **for all** *Assigned* records inside the group, X_i **do**
17: Update the value of Interaction as follows $\eta_p = \eta_p * a_{ip}$.
18: **end for**
19: **end for**
20: Let X^* be the record which has the highest value for η_p and mark it as an *Assigned* record.
21: Add X^* into this group and increment the number of records inside the group by unity.
22: Compute the association values between the most interactive record, X^* and each *Unassigned* record, X_i, in $\mathcal{D}$.
23: **end while**
24: Remove the present cluster from the set $\mathcal{D}$.
25: **end while**
26: Assign the remaining *Unassigned* records to the last group.
27: **return** the micro-aggregated data file, $\mathcal{D}'$.
28: **End Algorithm** ($IAMAT$)

very large when is close to unity. Therefore, the $IAMAT$ quantifies the value of the Association between two records, say X_i and X_j, to belong to the interval $[0, 1]$, and this is computed as follows:

$$a_{ij} = a_{ji} = r(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{\alpha}}. \tag{5.13}$$

where $r()$ is the identical function used in the definition of the $ACNN$, which evaluates the relationship between any two records, and which also involves α. As mentioned, the value of α is assigned so as to guarantee that the initial Association is large when the distance between X_i and X_j is small and vice versa. Typically, it is given by:

$$\alpha = \frac{\sqrt{n}}{\frac{1}{n}(\sum_{i=1}^{n} ||X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)||^2)}. \tag{5.14}$$

The rationale for incorporating the Association with the Interaction between the records inside a group, is that it leads to more homogeneous groups. The concept of the *Interaction* turns out to be crucial in forming the cluster, because we believe that merely being close to the farthest records is not a reason that is sufficiently important for any record to be grouped with the most distant one. Rather, we propose that the Interaction with respect to all the records inside the group has to be taken into consideration while clustering the records. As mentioned earlier, the latter is computed by invoking transitive-closure like operations. Finding the most interactive record with the associated pair is achieved by searching for the maximum product of the Association between the unassigned records, say X_p, and each record in the associated pair, $\langle X_i, X_j \rangle$, as follows:

$$\eta_{ij} = \begin{cases} a_{ip}(t-1) \times a_{pj}(t-1) & p \neq i, j \text{ and } i \neq j \\ 0 & i = j. \end{cases} \tag{5.15}$$

The above equation is valid when $k = 3$. By increasing the value of k, the transitive-closure is applied by adding one unassigned record at a time. The decision of grouping the unassigned record with other records in the group depends on the

Interaction of that record with respect to other records inside the group. Logically, the most interactive unassigned record has been chosen as follows:

$$Index\ Maximum_{1 \leq p \leq n}\ \eta_p, \tag{5.16}$$

where

$$\eta_p = \prod_{i=1}^{n_j} a_{ip} \tag{5.17}$$

where X_i represents the record inside the group, G_j, of size n_j and X_p represents the unassigned record.

5.4 Experimental Results

5.4.1 Data Sets

The *IAMAT* has been extensively tested[4] and the results obtained seem to be very good, where the "goodness" of a scheme refers to the combination of its being efficiently computed, and its offering a good trade-off between the *IL* and *DR*. We have tested it using the two real-life benchmark reference data sets used in previous studies, and in various simulated data sets obtained using the *Matlab's* build-in functions. These sets are:

1. The *Tarragona* data set which contains 834 records with 13 variables [45].
2. The *Census* data set which contains 1080 records with 13 variables [53].
3. A uniform distribution (min=0; max=1000).

[4]The reader is encouraged to re-visit Section 4.4 of Chapter 4 to see how the *MDAV* and the new method are compared using the composite Scoring Index, namely *SI*.

4. A normal distribution (μ=0; σ=0.05).

5. A uniform distribution (min=0; max=40,000) which contains 10,000 records and 16 dimensions.

6. A normal distribution (μ=500; σ=150) which contains 6,000 records and 16 dimensions.

While that data sets in items 1, 2, 5, and 6 are identical to those in Chapter 4, the data sets in items 3 and 4 involve vectors with dimensions ranging from 10 up to 80, and sets of cardinality from 10,000 up to 100,000. The latter data sets were used to investigate the scalability of the $IAMAT$ with respect to the size of the data, its dimensionality, and the number of records per group.

5.4.2 Results

For a given value of the security parameter k, which represents the minimum number of records per group, we compared the percentage value of the $IL = (SSE/SST)$ (as defined in Chapter 2) resulting from the $IAMAT$ and the $MDAV$ strategies. It is important to mention that the $MDAV$ was implemented based on the centroid concept and not a diameter concept[5]. All the programs were written in the C^{++} language, and the tests were performed on an Intel(R) Pentium (R)M Processor with the clock speed of 1.73 GHz., and with 512 MB of RAM.

Table 5.1 shows the improvement of the solution obtained by using the $IAMAT$ as opposed to the $MDAV$ on the multi-variate real data sets, where all the 13 variables were used simultaneously during the micro-aggregation process. The reduction in the value of the IL attained up to 8% on the Tarragona data set, and 5.12% on the Census data set when the group size was equal to 3. But in the case of the simulated data sets, the improvement in IL reached up to 14.47% when $k = 5$ in the uniform data set,

[5]We did not program the $MDAV$ scheme. We are extremely thankful to Dr.Francesc Sebé for giving us his source code.

Table 5.1: Comparison of the percentage of the IL and the computational time between the $MDAV$ and the $IAMAT$ on the real-life data sets (Tarragona and Census), and simulated data sets (Uniform and Normal distributions) for multi-variate methods.

Data Set	k value	$MDAV$		$IAMAT$		Improv. (%)
		IL	Time	IL	Time	
	3	16.9593	0.17	15.6023	0.31	8.00
Tarragona	4	19.7482	0.12	19.2872	0.22	2.33
	5	22.8850	0.12	22.7164	0.23	0.74
	3	5.6535	0.22	5.3639	0.41	5.12
	4	7.4414	0.19	7.2170	0.44	3.02
Census	5	8.8840	0.17	8.8428	0.42	0.46
	6	10.1941	0.17	9.9871	0.42	2.03
	3	22.4608	4.18	19.9730	8.52	11.08
Uniform	4	29.1714	4.45	25.2008	7.85	13.61
Distribution	5	33.9636	4.62	29.0478	7.48	14.47
	6	37.1577	4.75	32.1441	7.18	13.49
	3	26.9348	0.45	24.1245	0.92	10.43
Normal	4	33.9256	0.78	30.0154	0.81	11.53
Distribution	5	39.564	0.59	34.5886	0.87	12.58
	6	43.4592	0.60	38.0086	0.75	12.54

and it was as high as 12.58% when $k = 5$ in the normal data set. It is, thus, evident that the impact of the group size on the solution is minimized by increasing the number of records per group in the real data stes. To be fair, the computational time required to execute the $IAMAT$ is almost double the computational time required for the $MDAV$, although, in every case, the time was marginal – less than 0.5 second. In term of comparison, we believe that minimizing the loss in the data utility is more important than minimizing the extremely small computational time, especially because the micro-aggregation is usually performed off-line where the additional time requirement is less crucial. However, the question of how the decrease of IL is related to the increase in the computational time is still open.

Table 5.2: Scoring the $MDAV$ and $IAMAT$ with respect to the G_{IL} and G_{DR} by computing the index SI, for $k = 3$, 4, and 5 by using the Census data set and the simulated Uniform and Normal distributed data sets.

Data	*Creterion*	$k = 3$		$k = 4$		$k = 5$	
		$MDAV$	$IAMAT$	$MDAV$	$IAMAT$	$MDAV$	$IAMAT$
Census	G_{IL}	28.6100	22.1500	33.2540	30.3120	38.6720	34.9580
	RLD	60.7133	61.8889	49.7057	51.9070	42.0981	43.1771
	ID	1.9815	1.8704	0.7315	0.8611	0.2685	0.5556
	G_{DR}	31.3474	31.8796	25.2186	26.3840	21.1833	21.8663
	SI	33.8050	30.5750	36.1270	34.6560	38.8360	36.9790
Uniform	G_{IL}	115.4319	37.8225	116.1198	45.1763	120.9199	53.3052
	RLD	0.1088	1.1875	0.0694	0.8449	0.0602	0.5556
	ID	0.0042	4.7194	0.0000	1.2437	0.0000	0.4924
	G_{DR}	0.0565	2.9535	0.0347	1.0443	0.0301	0.5240
	SI	57.7442	20.3880	58.0773	23.1103	60.4750	26.9146
Normal	G_{IL}	108.9812	61.8348	89.0398	80.1520	110.8694	105.2394
	RLD	0.6494	11.8173	0.6571	6.7410	0.2814	5.3885
	ID	0.0085	0.8248	0.0000	0.1987	0.0000	0.0406
	G_{DR}	0.3290	6.3210	0.3285	3.4699	0.1407	2.7145
	SI	54.6551	34.0779	44.6841	41.8109	55.5051	53.9770

The other experiments were carried to test the SI of the $MDAV$ and $IAMAT$. Therefore, they have been scored as per the SI index on the simulated data and the Census[6] data sets, which contains seven key variables: $Var1$, $Var2$, $Var3$, $Var5$,

[6]Scoring them against the Tarragona data set is meaningless because the latter contains no so-called key variables.

$Var10$, $Var11$, and $Var12$.

Table 5.2 displays the SI for the $MDAV$ and $IAMAT$ methods for various values of k, which was set to be either 3, 4 or 5 as per the accepted requirements stated in [46, 115, 140]. As per Eq. (4.2), the G_{IL} was computed by averaging the values of M_1, M_2, M_3, M_4, and M_5. In general, the value of the G_{IL} is directly proportional to the number of records per group represented by the value, k. Therefore, in the Census data set, the best value of G_{IL} for the $IAMAT$ was obtained when $k = 3$, and was equal to 22.15%, while the best value of G_{IL} for the $MDAV$ was 28.61% when $k = 3$. In the context of the simulated data sets, and in general, the value of G_{IL} for the $IAMAT$ is less than half the value that was obtained by using the $MDAV$ (*i.e.,* the value of G_{IL} for the $IAMAT$ in the uniform data set and when $k = 3$ was 37.82%, while it was equal to 115.43% for the $MDAV$). Clearly, the $IAMAT$ method preserves the data utility more efficiently than the state-of-the-art, $MDAV$.

We also compared the G_{DR} (which estimates the risk of the data being disclosed) using the RLD and ID techniques. The RLD was computed using distance-based metric computations, where the average value was computed over several scenarios (as mentioned in Chapter 4). Each scenario computed the average of the RLD over all possible $\binom{S}{C} = \frac{S!}{C!(S-C)!}$ combinations, where S is the number of key variables in the micro-data file, and C is the number of chosen variables that are known to the Intruder. The results (Table 5.2) generally shows that estimating the risk of disclosing the secure information using the RLD method falls inversely "proportional" with the number of records per group represented by k. Thus, in the Census data set, the percentage value of the RLD for the $MDAV$ was as low as 60.71% when $k = 3$, 49.70% when $k = 4$, and 42.10% when $k = 5$, while it was 61.89% when $k = 3$, 51.90% when $k = 4$, and 43.18% when $k = 5$ for the $IAMAT$. Similarly, for the simulated data sets, we report the same fact which is that the $MDAV$ method is more secure than the $IAMAT$ method in estimating the RLD. For example, the percentage value of the RLD by using the $MDAV$ on the uniform data set was around 0.1%, while it was around 1.0% by using the $IAMAT$. On the other hand, the percentage value of the ID was calculated as the average value for various settings

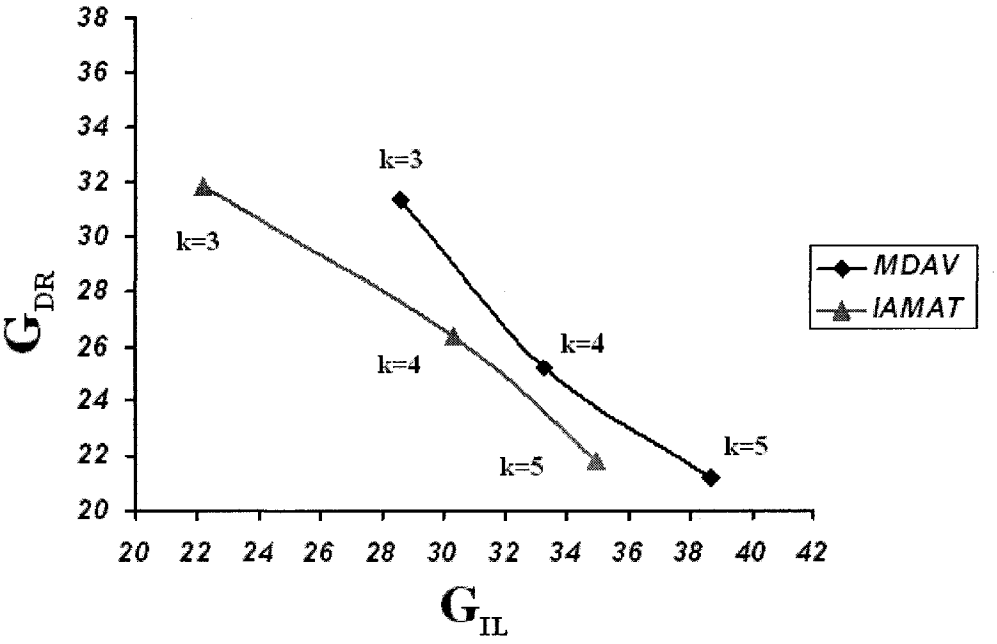


Figure 5.2: The effect of invoking the $MDAV$ and $IAMAT$ on the G_{IL} and G_{DR} indices when $k = 3$, 4 and 5 for Census data set.

of P, $(1, 2, \ldots, 10\%)$ on the Census data set, when the value of k was assigned to be 3, 4, or 5. We observe that the minimum percentage value of the ID was equal to 0.27%, which was achieved by invoking the $MDAV$ when $k = 5$. As opposed to this, the minimum percentage value of the ID for the $IAMAT$ was almost comparable – to 0.56% when $k = 5$. In the simulated data sets, the percentage value of the ID was equal to *zero* for the $IAMAT$ and $MDAV$. Therefore, the range of values for P was extended to reach up to 30% of the original size of the data sets. The percentage value of the $MDAV$ remained *zero*, while the minimum percentage value for the $IAMAT$ reached up to 0.49% and 0.04% when $k = 5$ in the uniform and normal data sets, respectively.

Finally, the SI value was computed for each MAT based on Eq. (4.4). Observe that a lower score value implies a superior performance. From the table we see that the $IAMAT$ technique has, almost consistently, a better performance index than the

$MDAV$ not only from the G_{IL} perspective, but also as from the perspective of a combination of the G_{IL} and G_{DR} for different values of k. Thus, for example, in Census data set the $IAMAT$ method scores the minimum value that was equal to 30.58 when $k = 3$, to 34.66 when $k = 4$, and 36.98% when $k = 5$. Besides the $IAMAT$ scores almost half the SI value which was obtained by the $MDAV$ by using the simulated uniform and normal data sets.

An interesting exercise involves studying how the $IAMAT$ and the $MDAV$ compared when "all" the factors coming to play. In other words, it would be good if we could understand the conditions when one algorithm is superior to the other and vice versa. While this can be done in many ways, we have plotted the G_{IL} versus the G_{DR} for the Census data set, in Figure 5.2, for both schemes. This figure presents sets of paired values of G_{IL} and G_{DR} for the respective algorithm for different values of k ranging from 3 to 5. The user should observe the influence of a masked method as the value of k changes. From the curve we see that the $IAMAT$ optimizes these conflicting criteria in a superior way than the stat-of-the-art $MADV$ method, because as the value of k increases, the increase of G_{IL} does not effect the value of G_{DR} in the $IAMAT$ method as it does in the case of the $MDAV$.

We also undertook a comprehensive evaluation of the performance of the $IAMAT$ scheme so as to investigate the scalability of the technique with respect to the cardinality of the data set, its dimensionality and the number of records per group.

- **The scalability of the $IAMAT$ with respect to cardinality.**
  We tested both the $IAMAT$ and the $MDAV$ schemes using data based on the uniform and the normal distributions with cardinalities ranging from $10,000$ records up to $100,000$ with 10 variables as shown in Tables 5.3. The percentage of the improvement achieved by invoking the $IAMAT$ in the IL, (when the value of k was set to 3) ranges from 10.02% to 11.25% for the normal distribution, and ranges from 10.47% to 11.28% for the uniform distribution. The same experiments were repeated for the value $k = 4$. The improvement increased and reached as high as 13.73% for the normal distribution and 13.67% for the

uniform distribution. It is fair to state that the $IAMAT$ requires almost triple the computational time needed to execute the $MDAV$ scheme. Figures 5.3 and 5.4 display similar trends for minimizing the value of the IL as a function of the data set size for uniform and normal distributions, respectively. In general, increasing the size of the data set tends to minimize the IL value. It should be mentioned that the $IAMAT$ was superior to the $MDAV$ in every single case.

Table 5.3: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $IAMAT$ on simulated data involving the uniform and normal distributions. The results demonstrate the scalability of the $IAMAT$ with respect to the size of the data set (when $k = 3$ and $k = 4$).

Data	k	Normal Distribution					Uniform Distribution				
		$MDAV$		$IAMAT$		improv.	$MDAV$		$IAMAT$		improv.
Size	value	IL	Time	IL	Time	(%)	IL	Time	IL	Time	(%)
10k	3	10.5383	13.53	9.4821	32.96	10.02	10.5437	13.75	9.4400	33.87	10.47
	4	13.9376	12.05	12.2106	29.37	12.39	13.9202	12.24	12.2069	29.96	12.31
20k	3	9.0065	54.33	8.0597	135.14	10.51	9.0458	58.57	8.0903	139.59	10.56
	4	11.9354	48.90	10.3653	120.37	13.15	11.9617	63.15	10.4275	145.09	12.83
30k	3	8.2177	125.75	7.3392	304.38	10.69	8.2435	127.04	7.3442	311.29	10.91
	4	10.8849	111.84	9.4488	277.20	13.19	10.9228	143.17	9.4976	328.02	13.05
40k	3	7.7352	222.23	6.8839	547.57	11.01	7.7576	226.43	6.9063	554.01	10.97
	4	10.2238	200.42	8.8732	494.94	13.21	10.2733	222.74	8.9063	526.37	13.31
50k	3	7.3644	349.66	6.5548	855.77	10.99	7.3724	335.32	6.5611	867.68	11.00
	4	9.7234	315.04	8.4382	773.92	13.22	9.7694	389.45	8.4574	903.39	13.43
60k	3	7.0538	505.63	6.2888	1,247.30	10.85	7.0896	516.40	6.2956	1275.53	11.20
	4	9.3513	457.50	8.1119	1,115.69	13.25	9.3707	644.32	8.1338	1386.01	13.20
70k	3	6.8195	693.42	6.0700	1,683.27	10.99	6.8504	723.90	6.0822	1790.66	11.21
	4	9.0358	623.89	7.8243	1,520.94	13.41	9.0776	629.73	7.8523	1500.12	13.50
80k	3	6.6273	909.36	5.8972	2,194.42	11.02	6.6537	924.03	5.9029	2277.30	11.28
	4	8.7898	809.397	7.6063	1,981.06	13.46	8.8183	1002.77	7.6247	1959.25	13.54
90k	3	6.4592	1,152.77	5.7387	2,776.69	11.15	6.4797	1181.84	5.7529	2855.63	11.22
	4	8.5584	1,044.77	7.4010	2,505.86	13.52	8.5885	1252.61	7.4140	2485.42	13.68
100k	3	6.3125	1,421.75	5.6021	3,418.86	11.25	6.3265	1615.74	5.6218	3209.00	11.14
	4	8.3799	1,512.68	7.2293	3,413.98	13.73	8.3933	1656.90	7.2456	3546.72	13.67

- **The scalability of the $IAMAT$ with respect to dimensionality.**
  We also tested the $IAMAT$ and the $MDAV$ on the uniform and the normal distributions for various dimensions of the variables ranging from 10 to 80,

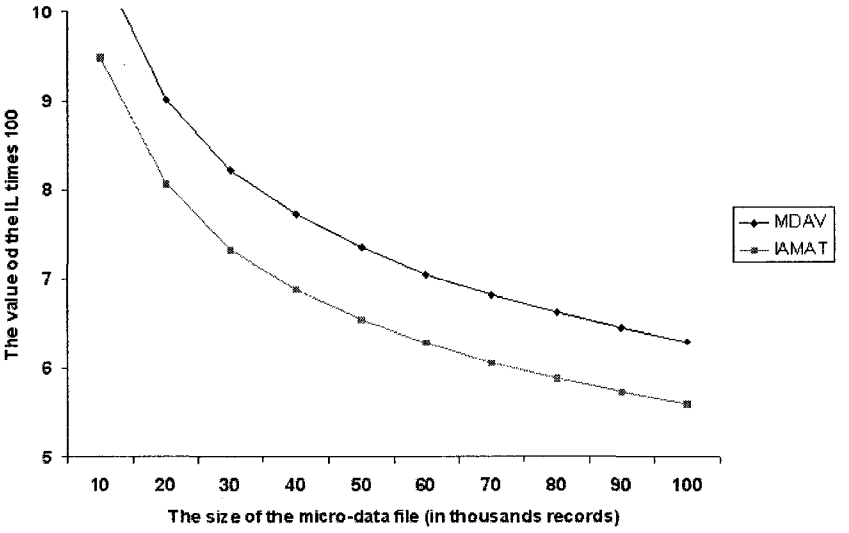


Figure 5.3: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the cardinality of the data set using normally distributed data when $k = 3$.

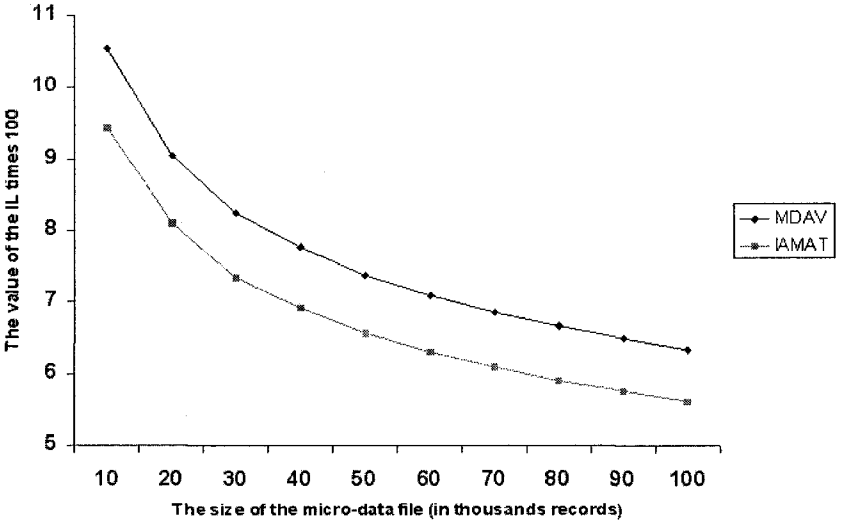


Figure 5.4: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the cardinality of the data set using uniformly distributed data when $k = 3$.

when the data size was set to 10,000 records, and the value of k was set to 3. Table 5.4 shows that the highest percentage of the improvement in the IL was about 10% for both the uniform and normal distributions. The computational time required to micro-aggregate all the individual records was almost directly proportional to the dimensionality in both schemes, and the required computational time in the $IAMAT$ was almost double the required time for the $MDAV$. Figures 5.5 and 5.6 show that both the $IAMAT$ and $MDAV$ schemes have similar trends for the IL *i.e.,* they are almost proportional to the dimension of the variables used in the micro-aggregation process for both distributions. As expected, increasing the dimensionality implies increasing the loss in the data utility. This is intuitively appealing because, informally speaking, increasing the dimensionality tends to minimize the similarity between the individual records and, at the same time, reduce the Association and the Interaction between the different multi-variate records.

Table 5.4: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $IAMAT$ on simulated data involving uniform and normal distributions. The results demonstrate the scalability of the $IAMAT$ with respect to the dimensionality of the micro-data file.

Number of variables	Normal Distribution					Uniform Distribution				
	MDAV		*IAMAT*		improv.	*MDAV*		*IAMAT*		improv.
	IL	Time	IL	Time	(%)	IL	Time	IL	Time	(%)
10	10.5383	13.53	9.4821	30.14	10.02	10.5437	13.75	9.4400	33.87	10.47
20	24.3766	24.30	22.3577	41.26	8.28	24.4692	25.03	22.2454	46.40	9.09
30	32.6044	35.50	29.9938	69.79	8.01	32.5113	36.67	29.9711	60.00	7.81
40	37.5679	47.00	34.8732	64.98	7.17	37.4627	47.18	34.8052	72.68	7.09
50	40.8705	40.87	38.2173	67.87	6.49	40.8427	58.78	38.2470	85.98	6.36
60	43.3104	70.01	40.7761	89.88	5.85	43.3912	69.48	40.8478	98.81	5.86
70	45.3212	82.78	42.7483	104.6	5.68	45.3778	81.45	42.8263	112.48	5.62
80	46.8119	93.78	44.3459	113.17	5.27	46.8533	97.10	44.4367	132.42	5.16

- **The scalability of the $IAMAT$ with respect to the number of records per group.**

  The scalability of the $IAMAT$ and the $MDAV$ with regard to the group size

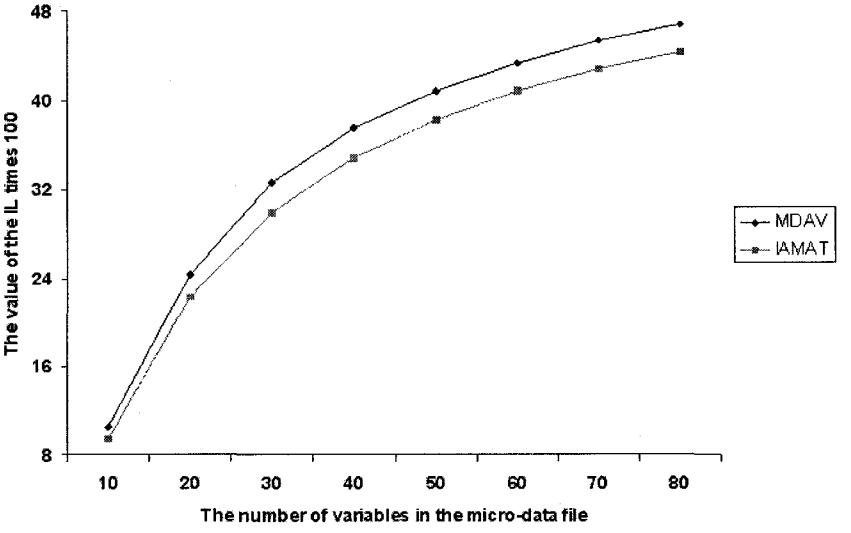


Figure 5.5: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the dimension of the data set using normally distributed data when $k = 3$ and $n = 10,000$.

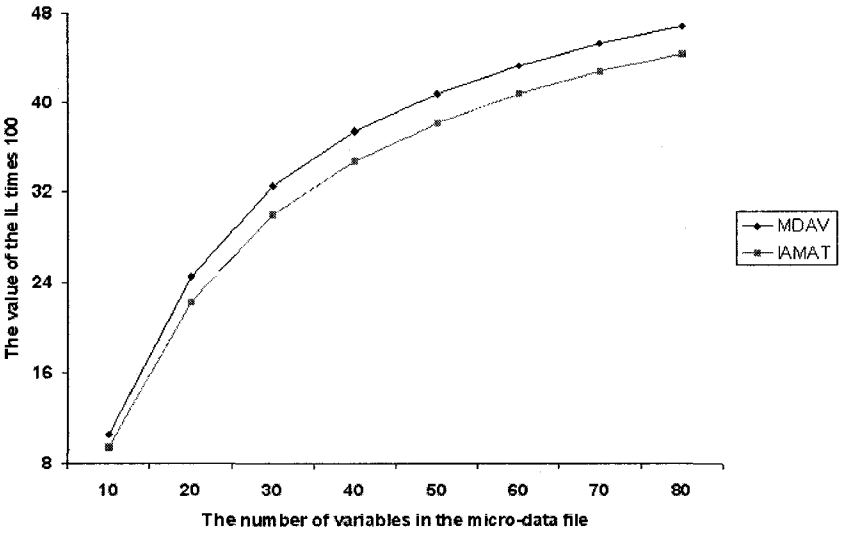


Figure 5.6: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the dimension of the data set using uniformly distributed data when $k = 3$ and $n = 10,000$.

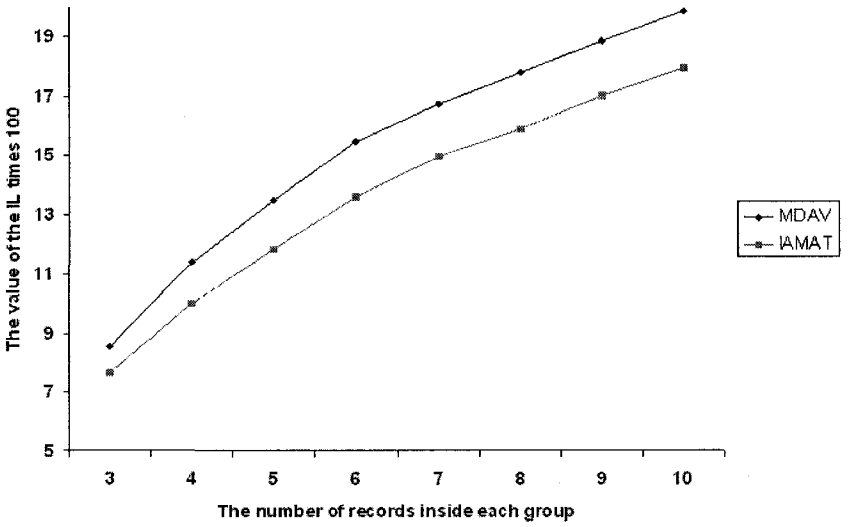


Figure 5.7: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the number of records per group using normally distributed data when $n = 10,000$ and $d = 10$.

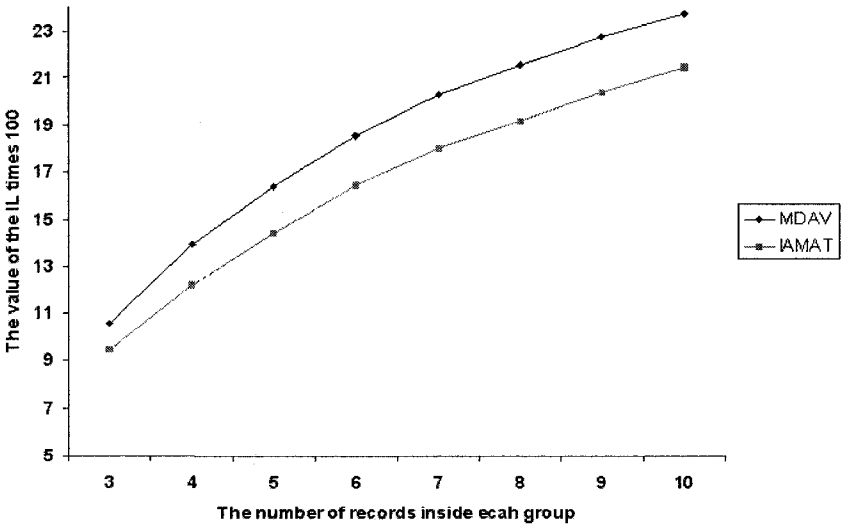


Figure 5.8: The improvement of the $IAMAT$ in reducing the percentage value of the IL as a function of the number of records per group using uniformly distributed data when $n = 10,000$ and $d = 10$.

has also been studied for both the uniform and the normal distributions, where the group size ranged from 3 to 10, and for the cardinality of the data set being 10,000 records, with a dimensionality of 10 variables. Table 5.5 shows the percentage of improvement in reducing the value of the IL. This reduction reached 12.74% for the normal distribution when the group size was 5, and reached 12.31% for the uniform distribution when the group size was 4. Figures 5.7 and 5.8 show that both schemes possess similar trends for the IL as a function of the group size, and that the value of the IL is proportional to the number of records per group. This is, again, a fair observation, because having many records in the group tends to minimize the within-group similarity, and to simultaneously maximize the similarity between the groups. This also tends to increase the value of the IL.

Table 5.5: Comparison of the percentage of the IL and the computation time between the $MDAV$ and the $IAMAT$ on simulated data involving uniform and normal distributions. The results demonstrate the scalability of the $IAMAT$ with respect to the number of records per group.

k	Normal Distribution					Uniform Distribution				
	$MDAV$		$IAMAT$		improv.	$MDAV$		$IAMAT$		improv.
value	IL	Time	IL	Time	(%)	IL	Time	IL	Time	(%)
3	10.5383	13.53	9.4821	32.96	10.02	10.5437	13.75	9.4400	33.87	10.47
4	13.9376	11.93	12.2106	28.79	12.39	13.9202	12.24	12.2069	29.96	12.31
5	16.5221	11.17	14.4167	26.67	12.74	16.4020	11.45	14.4091	29.32	12.15
6	18.7797	10.68	16.4650	26.86	12.33	18.5484	10.96	16.4152	28.78	11.50
7	20.3782	10.32	18.0344	25.48	11.50	20.2887	10.56	18.0431	28.28	11.07
8	21.6869	10.12	19.1562	26.67	11.67	21.5464	10.31	19.1668	26.32	11.04
9	22.8931	9.87	20.4255	24.92	10.78	22.7587	10.07	20.3952	26.02	10.39
10	23.8439	9.07	21.4969	26.06	9.84	23.6922	9.92	21.4520	25.92	9.46

5.5 Conclusions

In this chapter, we have considered the problem of achieving micro-aggregation in secure statistical databases. The novelty of our method involves enhancing the primitive

MAT that merely incorporates proximity information. The state-of-the-art MAT recursively reduces the size of the data set by excluding points which were farthest from the centroid and those which were closest to these farthest points. Thus, although the state-of-the-art method was extremely effective, we have argued that it uses only the proximity information, and ignores the mutual Interaction between the records. In this chapter, we have proved that inter-record relationships can be quantified in terms of two entities, namely their "Association" and "Interaction" that can be measured by invoking transitive-closure like operations, and by mapping the problem into a neural setting using the $ACNN$. By repeatedly invoking the inter-record Associations and Interactions, we have shown that the records can be grouped into sizes of cardinality "k". Our experimental results, which were done on artificial data and on the benchmark data sets for real life data, demonstrate that the newly proposed method is superior to the state-of-the-art by as much as 13%. Thus, we believe that our strategy leads to a very promising tool for solving the MAP.

By defining a score, SI, as a composite measure involving the IL and DR, we see that proposed strategy also obtains a minimum score value when compared to the $MDAV$ method. This indicates that the $IAMAT$ technique is probably the best MAT not only from the IL perspective, but also from the viewpoint of a measure which is as a combination of the IL and DR.

Chapter 6

Utilizing Dependence-Based Information to Enhance MATs

6.1 Introduction

Although Artificial Intelligent (AI) techniques have been used in various applications, their use in maintaining security in $SDBs$ has not been reported. This chapter[1] presents results, that to the best of our knowledge is pioneering, by which concepts from causal networks are used to secure SDBs. We consider the MAP in secure SDBs which involves partitioning a set of individual records in a micro-data file into a number of mutually exclusive and exhaustive groups. This problem, which seeks for the best partition of the micro-data file, is known to be NP-hard, and has been tackled using many heuristic solutions. In this chapter, we would like to demonstrate that in the process of developing $MATs$, it is expedient to incorporate AI-based

[1]A preliminary version of some of the results from this chapter appear in the *Proceedings of ACISP'08, the Thirteenth Australasian Conference on Information Security and Privacy*, in Wollongong, Australia, in July 2008 [108], and in the Proceeding of AI'08, *Twenty-First Australasian Joint Conference on Artificial Intelligence*, in Auckland, New Zealand in December 2008 [106]. The journal version of these results is currently under review.

causal information about the dependence between the random variables in the micro-data file. This can be achieved by pre-processing the micro-data *before* invoking any MAT, in order to extract the useful dependence information from the joint probability distribution of the variables in the micro-data file, and then accomplishing the micro-aggregation on the "maximally independent" variables. Our results, on real-life and artificial-life data sets, show that including such information will enhance the process of determining how many variables are to be used, and which of them should be used in the micro-aggregation process.

The structure of this chapter is as follows: In Section 6.2 describes how we can enhance micro-aggregation by incorporating dependence, and this is done both informally and algorithmically. Then, in Section 6.3, we present the results of experiments we have carried out for synthetic and real data sets. The chapter finishes in Section 6.4 with some conclusions.

6.1.1 Contribution of the Chapter

Central to the study of $SSDs$ are a family of algorithms classified in the literature as being $MATs$. Apart from being fast and efficient, they are also intuitively appealing because they are akin to the family of clustering methods. This chapter considers how such methods can be enhanced, both with regard to "accuracy" and efficiency, by learning, and thereafter incorporating the information that relates to the dependence between the random variables being analyzed. In all brevity, we are not aware of any other reported method which specifically incorporates such dependence-type information to optimize an MAT, or for that matter, to optimize a method which controls the IL and the DR in $SSDs$.

Understanding the presence and structure of dependency between a set of random variables is a fundamental problem in the design and analysis of many types of systems including filtering, pattern recognition etc. As far as we know its application in SDC has been minimal. Utilizing this information is the goal of this chapter. Typically, in

modern day systems, the data protector has been able to choose the technique and set its parameters without a thorough understanding of the characteristics of the micro-data file, and the stochastic dependence of the variables. Although gleaning this information could be particularly difficult and even time-consuming, our hypothesis is that this information is central to the micro-data file, especially when working in a high dimensional space.

In general, the result of the multi-variate *MATs* depends on the number of variables used in the micro-aggregation process. In other words, deciding on the number of variables to be taken into account, and on the *identity* of the variables to be micro-aggregated, is far from trivial. Domingo-Ferrer and Torra have reported in [54] that multi-variate micro-aggregation on unprojected data taking two or three variables at a time (rather than incorporating the information in all the variables) offers the best trade-off between IL and DR. The unanswered question is that of inferring which variables should be used in this process. We believe that a solution to this puzzle lies in the inter-variable "dependence" information.

The authors of [120] have emphasized that the decision about which variables are to be chosen has to be gleaned from *a priori* "knowledge about the characteristics of each variable from the experts". While this is a feasible approach, we argue that it is subjective, and that a formal objective method is desirable. Indeed, what will happen if the researcher encounters a new project for which there is no prior knowledge? Or how we will proceed if an expert for a specific data domain is not available? Our aim is to minimize the necessity to depend on a human expert, but rather to have the ability to study and estimate the characteristics of each variable objectively. Thus, we seek a systematic process by which we can choose the desired variables automatically and, thereafter, micro-aggregate the file.

This chapter involves *MATs*, but rather from a perspective different than the ones that have been considered in the literature. We propose a scheme by which we can avoid using the information in *all* the dimensions (for example, in computing the distance between two records etc.). Furthermore, neither will we resort to projecting

the micro-data file onto a single axis, nor will we attempt to micro-aggregate it using any specific sorting method [31, 32, 33, 34, 89, 91, 92, 115]. The main contribution of this chapter is to extract useful information from the joint probability distribution of the variables in the file to be micro-aggregated. Then, rather than use *all* the variables in the micro-data file, we propose to only process the "maximally independent" variables in the subsequent multi-variate micro-aggregation. Indeed, we propose to use such a method as a pre-processing step before *any* MAT is invoked, and to test the effect of using such a dependency analysis on the micro-aggregation process so as to reduce the computational time, and IL.

6.2 Enhancing Micro-Aggregation with Dependence ($EMAD$)

It is well-known that the result of the multi-variate $MATs$ depends on the number and the *identity* of the variables used in the micro-aggregation process. Since multi-variate micro-aggregation using two or three variables at a time offers the best trade-off between the IL and the DR [54], the question we intend to resolve involves understanding why we have to maintain and use vast dimension-dependent resources in the clustering phase in order to compute the distance between the micro-records. We shall also study how we can minimize the computation time needed to evaluate the distance between a single micro-data record and the mean of the group it belongs to. This computation involves evaluating

$$D(X,Y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \quad , \tag{6.1}$$

where X and Y are two multi-variate data vectors with their components being $\{x_i\}$ and $\{y_i\}$ respectively, and d represents the dimension of the space.

We consider the problem of determining the dependencies between the different variables within a micro-data file, and then combining the latter with the MAT in such a way as to reduce the overall required computational time, and/or reduce the corresponding IL.

The primary goal of any MAT is to reduce the loss in the data utility by choosing the most suitable sub-set of variables with size equal to three [54] prior to invoking the multi-variate micro-aggregate. Theoretically, to know the best sub-set of variables that has to be used in order to obtain the minimum value of the IL, we have to consider all different possibilities of combinations, namely the $\binom{S}{C} = \frac{S!}{C!(S-C)!}$ combinations, where S is the number of variables in the original micro-data file, and C is the number of chosen variables which are used in projecting and micro-aggregating the data file.

We propose that the key idea in choosing a sub-set of the variables by avoiding the combinatorial solution, should be based on the dependence model of the micro-data file. If the variables are highly-correlated, then using any one of them will somehow reflect the stochastic nature of the others. If we, thus, incorporate this logic into our consideration, we believe that we can reduce the number of variables which will be used to measure either the distance between the micro-unit and the mean of the group it belongs to, or the distance between the micro-units themselves. Thus, in turn, this will reduce the dimensionality of the space to $d' < d$. The new distance that will thus be computed will be:

$$D'(X,Y) = \sqrt{\sum_{i=1}^{d'}(x_i - y_i)^2} \qquad \text{where } d' < d. \tag{6.2}$$

The reader should observe that our goal is quite distinct from the reported methods of projecting the multi-dimensional space onto a single axis using a particular variable, the sum z-scores scheme, or a principal component analysis. The reduction in the dimensionality is not done randomly. Rather it is to be done based on a formal criterion. Our aim is to micro-aggregate the multi-dimensional vector by maximally

using the information in the "almost-independent" variables, and we plan to do this by finding the best dependence tree. We believe that we can achieve this by evaluating the dependence between the variables in the micro-data file by using either the method due to Chow and Liu [17] or the method due to Valiveti and Oommen [135, 136].

We formalize these concepts below. The joint probability distribution of the random vector $\mathbf{V} = [V_1, V_2, \ldots, V_d]^T$ in terms of conditional probabilities is given as

$$P(\mathbf{V}) = P(V_1)P(V_2|V_1)P(V_3|V_1, V_2)\ldots P(V_d|V_1, V_2, \ldots, V_{d-1}), \tag{6.3}$$

where each V_i is a random variable.

It is obvious, from the above expression, that each variable is conditioned on an increasing number of other variables. Therefore, estimating the k^{th} term of this equation requires maintaining the estimates of all the k^{th} order marginals. Clearly, it is impractical to gather the estimates for the joint density function $P(\mathbf{V})$ for all the different values which V could assume. We, therefore, simplify the dependency model by restricting ourselves to the lower-order marginals, using the approximation which ignores the conditioning on multiple variables, and retaining only dependencies on at most a single variable at a time. This leads us to the following [135]:

$$P_a(\mathbf{V}) = \prod_{i=1}^{d} Pr(V_i|V_{j(i)}), \tag{6.4}$$

where $P_a(\mathbf{V})$ is the approximated form of $P(\mathbf{V})$, and V_i is conditioned on $V_{j(i)}$ for $0 \leq j(i) < i$.

The dependence of the variables can be represented as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ where $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ is a finite set of vertices, which represents the set of random variables in the micro-data file with d dimensions, $\mathbf{E}$ is a finite set of edges $\{\langle V_i, V_j\rangle\}$, where $\langle V_i, V_j\rangle$ represents an edge between the vertices V_i and V_j. Finally, $\mathbf{W} = \{w_{i,j}\}$ is a finite set of weights, where $w_{i,j}$ is the weight assigned to the edge

$\langle V_i, V_j \rangle$ in the graph. The values of these weights can be calculated based on a number of measures, as will be explained presently.

In **G**, the edge between any two nodes represents the fact that these variables are statistically dependent [17]. In such a case, the weight, $w_{i,j}$, can be assigned to the edge as being equal to the Expected Mutual Information Measure ($EMIM$) metric between them. Generally speaking, the $EMIM$ metric between two variables, given by $I^*(V_i, Vj)$ for discrete distributions, has the form:

$$I^*(V_i, V_j) = \sum_{v_i, v_j} Pr(v_i, v_j) \log \frac{Pr(v_i, v_j)}{Pr(v_i) Pr(v_j)} \ , \tag{6.5}$$

where the summation above is done over all values of v_i and v_j which V_i and V_j can assume.

Observe that any edge, say $\langle V_i, V_j \rangle$ with the edge weight $I^*(V_i, V_j)$ represents the fact that V_i is stochastically dependent on V_j, or that V_j is stochastically dependent on V_i. Although, in the worst case, any variable pair could be dependent, the model expressed by Eq.(6.4) imposes a tree-like dependence. It is easy to see that this graph includes a large number of trees (actually, an $\bigcirc(d^{(d-2)})$ of such spanning trees). Each of these trees represents a unique approximated form for the density function $P(\mathbf{V})$. Chow and Liu proved that searching for the best "dependence tree" is exactly equivalent to searching for the Maximum Spanning Tree[2](MST) of the graph [17]. Further, since the probabilities that are required for computing the edge weights are not known *a priori*, Valiveti and Oommen showed that this could be achieved by estimating them in a maximum likelihood manner [135, 136]. They showed that the Maximum Likelihood (ML) estimate for the best dependence tree, can be obtained

[2]Two generic greedy algorithms can be used to solve the Minimum Spanning Tree problem, namely, the so-called Kruskal and the so-called Prim algorithms. Both of them run in time $\bigcirc(E \lg V)$ by using ordinary binary heaps [18, 118, 125]. Since we are attempting to compute the Maximum Spanning tree, it is obvious that we have to order the edges in a decreasing order (as in Kruskal) or to extract the maximum edges weight (as in Prim). We have used the Kruskal algorithm in our experiments.

by computing the MST of the graph, where the edge weights are computed using the $EMIM$ of the estimated probabilities, as shown in Figure 6.1.

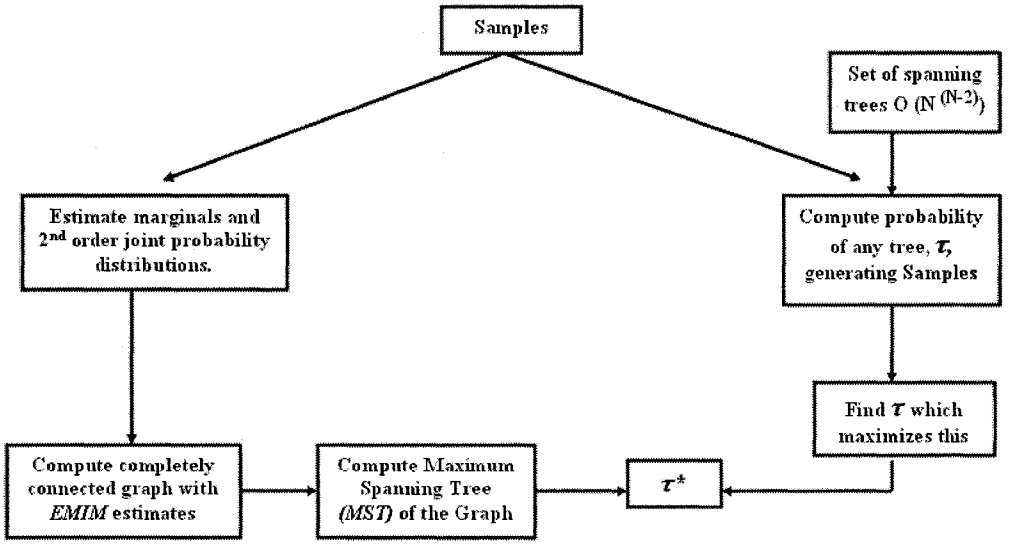


Figure 6.1: Equivalent procedures for finding the Maximum Likelihood Estimate of the tree-based dependence from the samples.

It is worth mentioning that this solution is truly both elegant and efficient. A rigorous ML solution to obtaining the best tree would involve computing *it* from the set of all possible spanning trees, which (at the enumeration level itself) is a combinatorially explosive problem. To solve the ML problem in a formal manner, one has to first obtain the set of all the graph's spanning trees, and then determine the tree which maximizes the likelihood function evaluated in terms of the dependence described by the tree itself. Observe that the solution obtained by solving for the MST is many orders of magnitude less complex. It involves estimating *the probabilities* (and not the structure) of the Binomial (multinomial) distributions using a ML estimate, and then merely computing the MST. The fact that these two processes lead to the same estimate (as shown in Figure 6.1) is far from trivial to prove, but is

indeed, true.

It should be mentioned here that the weights of the edges in the graph, **G**, can be computed using either the $EMIM$ metric or the χ^2 metric proposed by Valiveti and Oommen [135]. The latter, $I_\chi(V_i, V_j)$, is an alternative measure that quantifies the dependence information between pairs of random variables, and is computed by:

$$I_\chi(V_i, V_j) = \sum_{v_i, v_j} \frac{(Pr(v_i, v_j) - P(v_i)P(v_j))^2}{P(v_i)P(v_j)} \ . \tag{6.6}$$

I_χ has the following desirable characteristics relevant to capturing dependence information:

$$\begin{cases} I_\chi(V_i, V_j) = 0 & \text{iff } P(v_i, v_j) = P(v_i)P(v_j) \\ I_\chi(V_i, V_j) > 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

It turns out that for binary and normally distributed random variables, the I_χ metric is exactly equivalent to the I^* metric in finding the dependence tree [135, 136]. But, when the underlying dependence is not actually based on a tree structure, both of them estimate the best dependence tree corresponding to their representative measures. Valiveti and Oommen showed the interesting feature that although their estimation for the best dependence tree does not always match, the total weights are almost always identical.

By way of example, consider a micro-data file which incorporates 6 variables (as in Figure 6.2) and thousands of records. Let us assume that we intend to micro-aggregate this file using any MAT, for example, the $MDAV$ method. In such a case, the prior art will process all the six variables to quantify the relevant distances during the clustering stage. We could choose a sub-set of size three to be used in the micro-aggregation process. In general, we will have to go through the 20 different combinations of size three in order to attain the minimum value of the IL. However, if we are able to discover any existing inter-variable dependencies, this could render

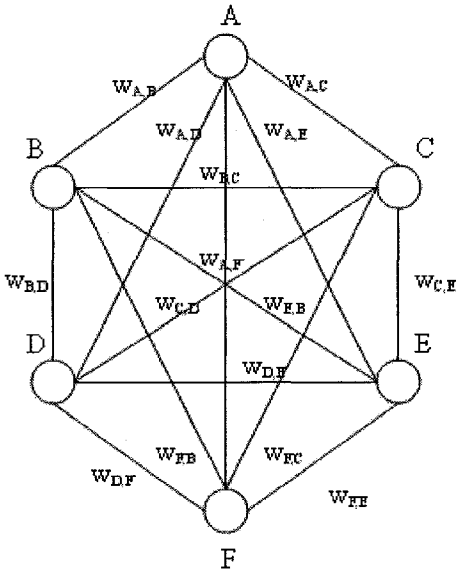


Figure 6.2: The fully-connected undirected graph represents the dependence between six random variables.

the problem simpler. Let us assume that we compute the $EMIM$-based edge weights for all pairs of nodes, and create the fully-connected undirected graph G, as in Figure 6.2. By using the strategy alluded to above, we obtain a tree as in Figure 6.3.a, which shows the case when the MST leads to the ML condition that the variables B, C, and D depend on the variable A, and that variables E and F depend on variable D.

Since these dependent variables are maximally-correlated to the variable that they depend on, we propose to use the vertices that have the maximum number of In/Out edges in the graph to micro-aggregate the micro-file. We believe that the nodes which possess this property are the best candidates to reflect the characteristics of the entire multi-variate data set because they connect to the maximum number of nodes that statistically depend on it, as argued in Conjecture 1.

Conjecture 1. *Micro-aggregating the micro-data file can be best achieved if the nodes which possess the maximum number of In/Out edges in the tree obtained as the MST of the underlying undirected connected graph* $\boldsymbol{G}$*, are used as the input to solve the*

MAT.

Rational for Conjecture

The existence of an edge between two nodes in the connected undirected graph signifies that these two nodes are statistically correlated to each other, and that a variation of one of these variable is reflected by a corresponding change in the other. Thus, the variables which are connected to each other via edges in the skeletal tree represent nodes which are connected to each other based on the best tree-based dependence, and in turn, reflect the maximal shared characteristics within the variables of the micro-data file. Thus, any node which has a larger number of In/Out edges is one which connects to a larger number of nodes, and is thus capable of individually representing more "other" variables. This implies that the best candidates to be used to represent the other variables in the micro-aggregation are those which have the maximum number of In/Out edges.

In order to invoke this property, we first rank the nodes of the graph based on the number of In/Out edges in a descending order and choose the first d' variables, where d' is usually determined by the data protector, and is usually equal to 3 or 4. Thus, for example, based on the above discussion, for the data represented by the variables of Figure 6.3, the micro-aggregation process will be invoked by using two variables instead of using the entire set of six variables in the micro-data file. Figure 6.3(b) shows that the selected sub-set of the variables is $\{A, D\}$, since both of them connect to 3 variables while the other variables in the micro-data file connect to only a single variable. The process outlined above has been formalized in Algorithm 13 which presents an automated way to select a sub-set of the variables to be used in the multi-variate micro-aggregation process.

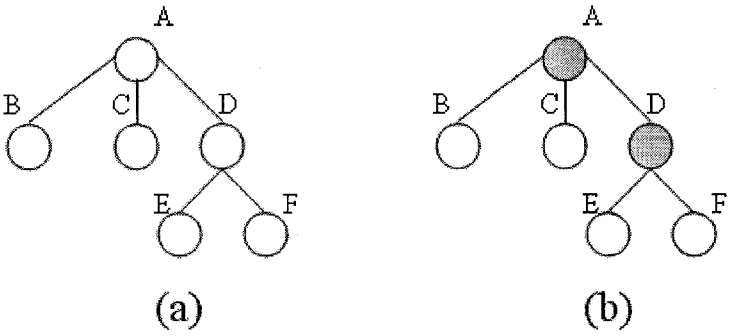


Figure 6.3: An example of a dependence tree used to micro-aggregate the data file containing 6 variables.

Algorithm 13 *EMAD*

Input: $\mathcal{U}$: the micro-data file, and C: the number of variables that will be used in the micro-aggregation process.

Output: d': the sub-set of the variables that will be used in the multi-variate MAT.

Method:

1: Estimate the first and second order marginals of the random variables from the various micro-records.

2: Create a fully-connected undirected graph, where the

Nodes: Represent the random variables in the micro-data file.

Edges: Represent the possible statistical dependence between these variables.

Weights of the edges are computed either by using:

$$EMIM \Rightarrow I^*(V_i, V_j) = \sum_{v_i,v_j} Pr(v_i, v_j) \log \frac{Pr(v_i,v_j)}{Pr(v_i)Pr(v_j)}, \text{ OR}$$

$$\chi^2 \Rightarrow I_\chi(V_i, V_j) = \sum_{v_i,v_j} \frac{(Pr(v_i,v_j)-P(v_i)P(v_j))^2}{P(v_i)P(v_j)}.$$

3: Invoke Kruskal's algorithm to compute the Maximum Spanning Tree of the graph.

4: Rank the nodes of the graph based on the number of In/Out edges in a decreasing order, and reckon the first d' variables to be the sub-set to be used in the MAT.

5: **Return** the sub-set of variables which will be used in the micro-aggregation process before invoking the MAT

6: **End Algorithm** *EMAD*

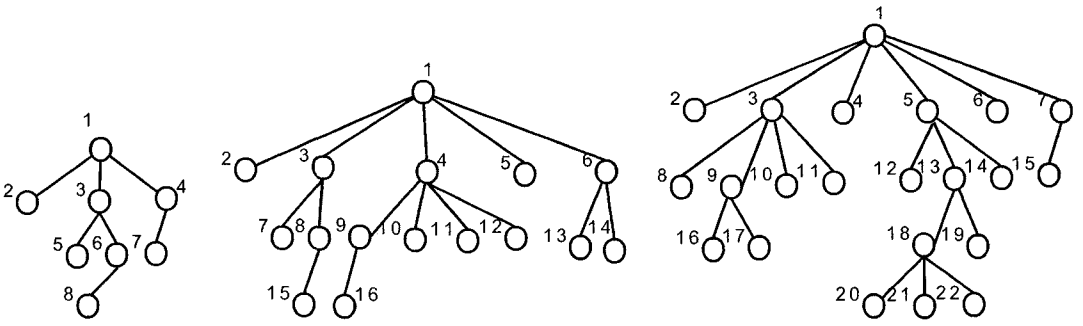
6.3 Experimental Results

6.3.1 Data Sets

In order to verify the validity of our methodology in projecting the multi-variate data set into a subset of random variables to be used in the micro-aggregation process, two benchmark real-life data sets and three simulated data sets were used in the testing phase. Table 6.1 summarizes the characteristics of each data set by defining its type, dimensionality and cardinality.

Table 6.1: The characteristics of various data sets.

Name of the data set	Type	Dimensionality	Cardinality
Tarragona	Real	13	834
Census	Real	13	1080
Sim_1	Simulated	8	5000
Sim_2	Simulated	16	10,000
Sim_3	Simulated	22	20,000



(a) *Sim_1* (b) *Sim_2* (c) *Sim_3*

Figure 6.4: The true structures for the simulated data sets

The Tarragona and Census benchmarks are reference data sets used in previous studies for their special statistical properties [45, 53]. On the other hand, the simulated data were generated or tested for various dimensions of random vectors, as

follows: First of all, the number of random variables was determined. Thereafter, the "true" structure of the defined dependence tree which imposed the dependence relationships between the variables was selected subjectively, as shown in Figure 6.4. Then the second-order marginal distributions were randomly generated. The procedure by which these were generated was as follows: If we define the entire space of each variable to be between 1 and 1000, this space is sub-divided into a number of subspaces with equal width, say 100. That means that we limit ourselves to be dealing with 10 events where each event represents a sub-interval of width equal to 100 from the entire domain as follows: $\{I_1 = [1, 100], I_2 = [101, 200], \ldots, I_{10} = [901, 1000]\}$, thus, effectively simulating a multinomial distribution. In the latter, each outcome is a random number belonging to exactly one of the 10 sub-intervals, I_j, with probability, P_j, where $j = 1, 2, \ldots, 10$. If n_j represents the number of occurrences of values belonging to I_j and n represents the number of independent records, we have

$$\sum_{i=1}^{10} n_i = n, \sum_{i=1}^{10} P_i = 1, \tag{6.8}$$

where the probability mass function of the multinomial distribution is

$$f(n_1, n_2, \ldots, n_{10}) = \frac{n!}{n_1!\ n_2!\ \ldots\ n_{10}!} \prod_{i=1}^{10} P_i^{n_i}. \tag{6.9}$$

Observe that prior to assigning the second order marginal distributions for the rest of the tree, we had to also randomly generate 10 different probabilities for the most independent variables (the root variable) when its values belonged to each of the above defined sub-intervals.

To randomly populate the file, we can now randomly assign values to the conditional probability from the joint and marginal distributions as follows: If, as per the assumed tree-based dependence, variable V_m, depends on variable V_n, this means we have to define a set of probabilities, $\{P_{nm}\}$, when the value of V_n, say v_{in}, belongs

to any defined sub-interval I_j given that the value of variable V_m, say v_{im} belongs to any sub-interval I_l. Thus,

$$P_{mn} = Pr(v_{in} \in I_j |\ v_{im} \in I_l), \tag{6.10}$$

where i represents the index of the record in the micro-data file and assumes values in $\{1, 2, \ldots, n\}$. The indices j and l represent the indices of the sub-interval where the random variable falls, and which are the result of dividing the entire domain into 10 sub-intervals. Finally, the indices n and m represent the specific dimensions in the micro-data file, and are in the range $\{1, \ldots, d\}$, $n \neq m$.

The above procedure was implemented for all pairwise combinations of random variables associated with the micro-data file.

6.3.2 Results

The experiments conducted were of five categories: In the first set of experiments the intention was primarily focused on testing whether the best dependence tree can be learned (or rather, inferred) from the continuous micro-data file, and if it sufficiently reflected the dependence model. In the second set of experiments, the goal was to investigate whether the algorithm is able to infer the dependence model between the random variables when additional information about the dependency between the sibling in the tree is available. In the third set of experiments, the goal was primarily to validate our strategy for determining the subset of variables (from the entire set of variables) to micro-aggregate the micro-data file, and to study its effect on the value of the IL. The fourth set of experiments was designed to determine the most suitable metric to calculate the edge weights of the fully-connected graph so as to minimize the required computation time and maximize the accuracy of estimating the dependence model and its effect on the value of the IL. Finally, since we are working with continuous vectors, the last set of experiments focused on understanding the effect

of assuming normality (*i.e,* the relevance of the Central Limit Theorem [61]) on the data set in calculating the edges weights.

- **Experiment Sets 1:**

  The first set of experiments was done on two types of data sets: Simulated data sets with a known structure of the best dependence tree which is to be inferred by the learning algorithm, and the real data sets possessing an unknown dependence model between the variables. It is worth mentioning that we could not approximate the dependence information of the multi-variate data set in its current form due to the inaccurate estimation for the joint and marginal probability distributions for continuous variables. This is a consequence of having a large domain space with only few records (sometimes only one or two) for each region of the corresponding random variable. Consequently, most of the *estimated* marginal and joint probability values were close to zero. Clearly, in these cases, the *estimated* probabilities will not reflect the actual dependence relationship between any corresponding variables.

  In order to overcome this challenging problem that prevents us from utilizing the dependence information, we were forced to reduce the domain space by categorizing the micro-data file as follows: We first scanned the micro-data file to specify the domain space of each variable in the file, and then divided it into a number of sub-intervals sharing the same width. After that, we achieved a categorization phase by replacing the values belonging to a certain sub-interval in each variable by the corresponding category/code. For example, in the case of the simulated data sets, all the variables shared the same domain space between 1 and 1000, which was divided into 10 subintervals, as explained earlier. Consequently, all values belong to the $[1, 100]$ interval were replaced by 1, all values belong to the $[101, 200]$ interval were replaced by 2 and so on. The above procedure was repeated for all the variables so as to generate the categorical micro-data file.

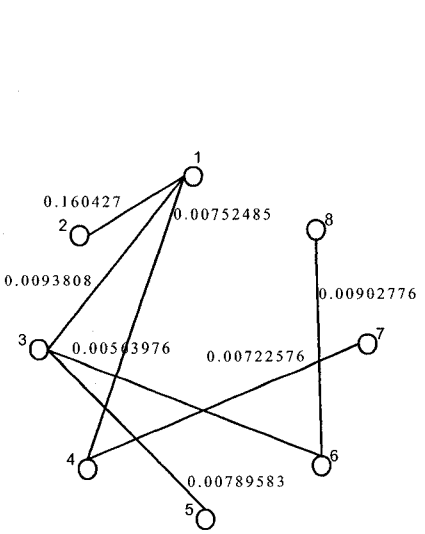
  From the above discussion, it is clearly shown that "width" parameter plays a predominant role in controlling the degree of smoothing and estimating the best

dependence tree. Our experiments indicated that assigning a suitable value to the width parameter guaranteed the convergence of the MST to the true underlying (unknown) structure of the best dependence tree. The most important point that one has to be aware of in a practical scenario is that a larger value for the width parameter implies a lower variance and a higher bias, because we are essentially assuming a constant value within the sub-interval. Generally speaking, the value of the width parameter should be large enough to generate a sufficient number of sub-intervals from the defined domain space to guarantee a satisfactory level of smoothing. The actual value used is specified in the respective experimental results.

Consider the tree structure given by Sim_1, Sim_2, and Sim_3 as given in Figure 6.4. Approximating the dependence information of the simulated data sets based on the structure of the MST obtained using the $EMIM$ metric succeeded in locating the real structure when the width parameter was set to the values 50, 100, and 150 for Sim_1, 70, 100, and 120 for Sim_2, and 90, 100 and 110 for Sim_3. Figure 6.5 shows the edge weights and the value of I_x for each simulated data set when the value of width was equal to 100. Figures 6.6, 6.7 and 6.8 show different snapshots of the convergence to the final structure of the dependence model for various sample sizes for Sim_1, Sim_2 and Sim_3, respectively, when the value of the width parameter was set to 100.

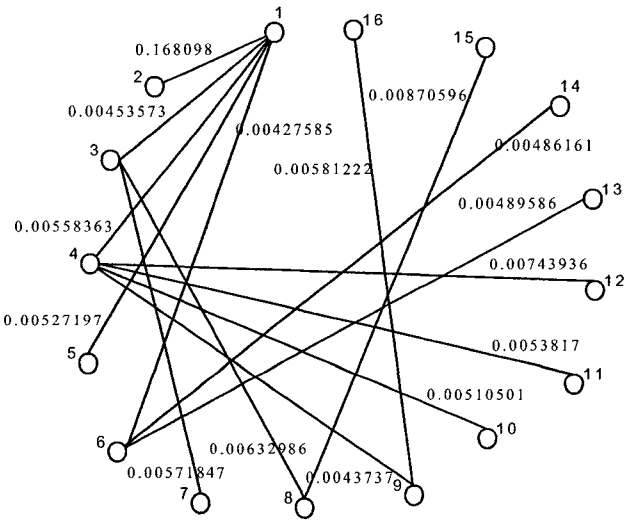
Approximating the dependence information for the real data sets was a little more "tricky", because of the unknown structure for the best dependence tree. Changing the value of the width parameter has an effect on the structure of the best dependence tree to which the algorithm converged. Figures 6.9 and 6.10 clearly show different structures for the best dependence tree by changing the value of the width for the Tarragona and Census data sets, respectively.

- **Experiment Sets 2:**
  The second set of experiments involves the so-called *Sibling-related* Model. The aim here was to see if the algorithms possessed the ability to infer the structure of the dependence model between the random variables if additional information



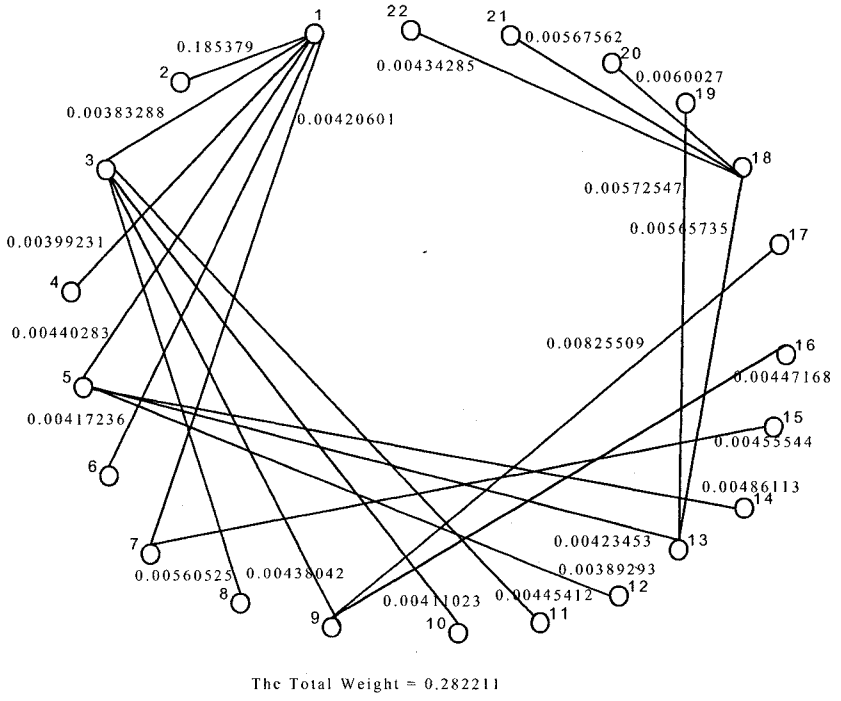
The Total Weight = 0.207122

(a) *Sim_1*



The Total Weight = 0.246389

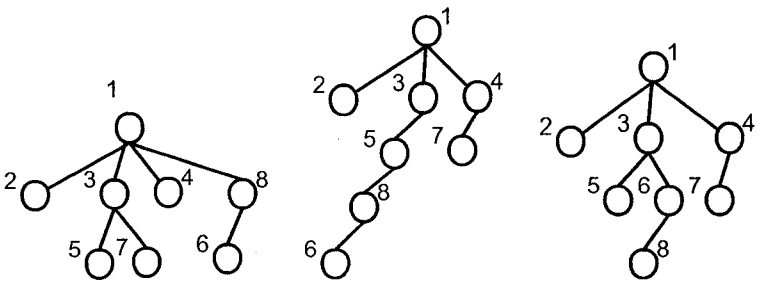
(b) *Sim_2*



The Total Weight = 0.282211

(c) *Sim_3*

Figure 6.5: The best dependence tree for the simulated data sets obtained by using the *EMIM* metric.



(a) 1000 (b) 2500 (c) 4000

Figure 6.6: The "inferred" dependence tree for the *Sim_1* binary data set as the number of samples increases. The width parameter was set to 100.

about the dependency between the siblings in the tree is available. The results that we have obtained are quite amazing.

To be more specific, we consider the possibility that after the structure of the underlying tree is determined, the *probability* values between the siblings in the structural tree are related. For example, thus, if a particular node had index i and its children were nodes j and k, the probabilities that could be *independently* set were:

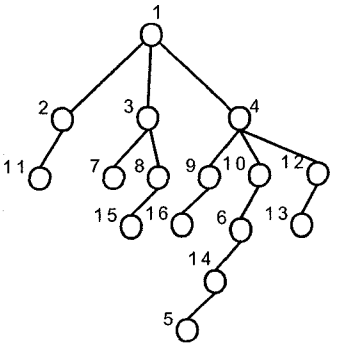
$$Pr[x_j = 0|x_i = 0]$$

$$Pr[x_j = 0|x_i = 1]$$

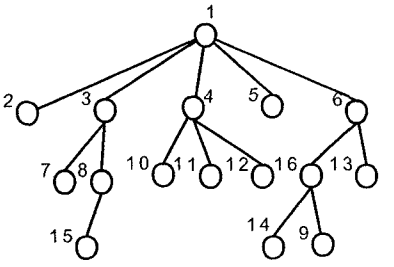
.

Since the probabilities of the siblings were thus determined, the values of $Pr[x_k = 0|x_i = 0]$ and $Pr[x_k = 0|x_i = 1]$ were then set to be $1 - Pr[x_j = 0|x_i = 0]$ and $1 - Pr[x_j = 0|x_i = 1]$ respectively. Further, observe that a result of these assignments, the probabilities, $Pr[x_j = 1|x_i = 0]$, $Pr[x_j = 1|x_i = 1]$, $Pr[x_k = 1|x_i = 0]$ and $Pr[x_k = 1|x_i = 1]$ were automatically assigned, since the sum of these quantities and the values of their counterparts, is unity.

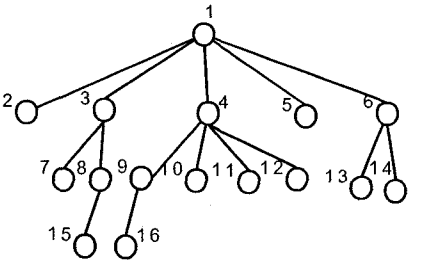
The question we were interested in investigating was to see if our strategy for



(a) 1000

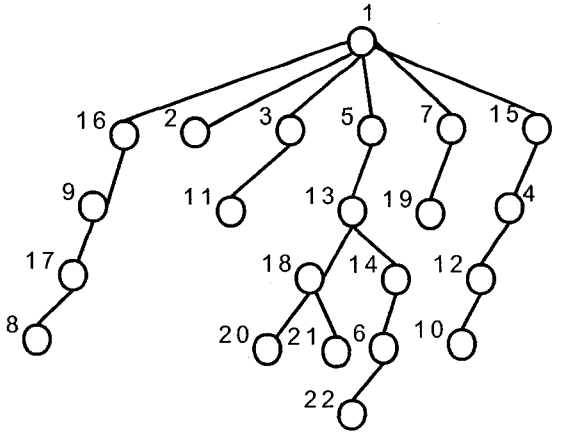


(b) 2500

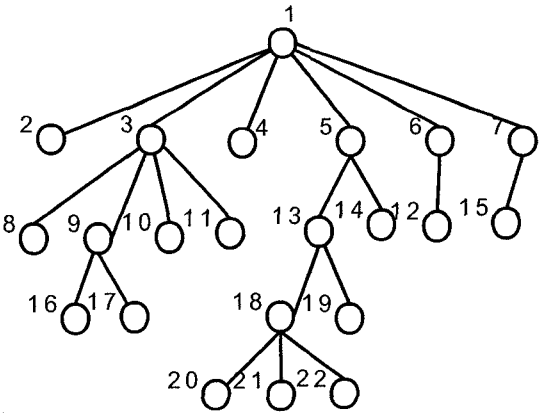


(c) 4000

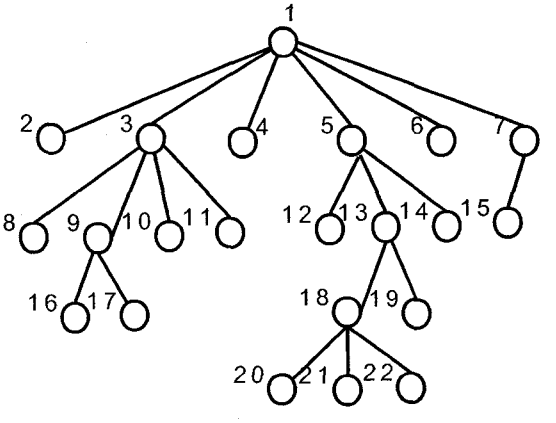
Figure 6.7: The "inferred" dependence tree for the Sim_2 binary data set as the number of samples increases. The width parameter was set to 100.



(a) 1000

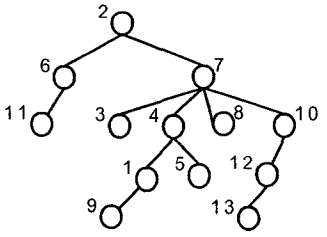


(b) 2500

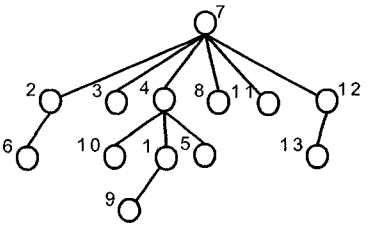


(c) 4000

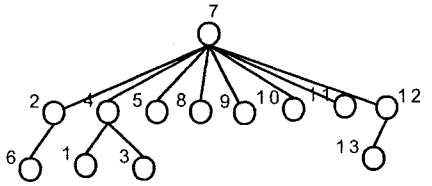
Figure 6.8: The "inferred" dependence tree for the *Sim_3* binary data set as the number of samples increases. The width parameter was set to 100.



(a) Width = 50,000

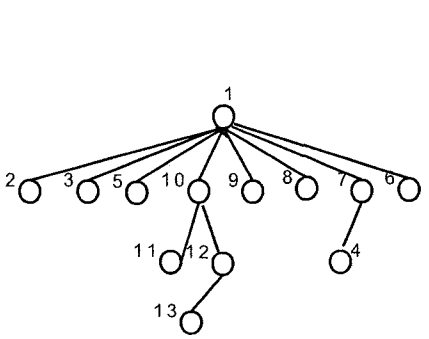


(b) Width = 100,000

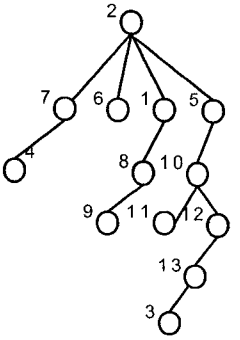


(c) Width = 150,000

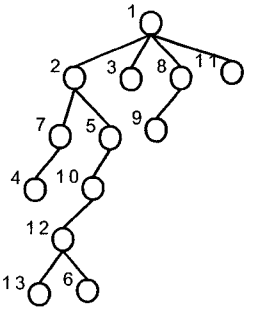
Figure 6.9: The best dependence tree for the Tarragona data Set obtained by using the $EMIM$ metric with various values of the width parameter.



(a) Width = 1,000



(b) Width = 5,000



(c) Width = 10,000

Figure 6.10: The best dependence tree for the Census data set obtained by using the $EMIM$ metric with various values of the width parameter.

learning the dependence tree using the MST on the constructed fully-connected graph (where the edges weights are calculated using the $EMIM$ or the χ^2 metric) was able to converge to the true (unknown) dependence tree even if this sibling relationship was not known. The answer was always in the affirmative.

By way of example, consider two binomial data sets with 6 variables. Both of them share the same dependency model between the variables, as shown in Figure 6.11. The only difference between these two data sets is that the values of the probabilities used to generate the random variables in the true tree structure – which in one case *was* sibling-related, and in the other *was not* sibling-related. Tables 6.2 and 6.3 show the values of the random probabilities which were used in generating each variable in the data set. Observe that in the first data set these values are related, while they are independent in the second data set. Figure show different snapshots of the convergence to the dependence model as the number of samples is increased.

The actual trees learnt for the data sets, as the number of samples processed increased, are given in Figures 6.12 and 6.13 respectively (reported at snapshots 50, 150 and 5,000). The decrease in the $EMIM$ and chi^2 metrics with time are plotted in Figure 6.14. Observe that the final inferred tree in both cases is exactly the unknown tree – which, again, was correctly inferred, and that the values of both the metrics ultimately converged to the lowest possible values. Thus we conclude that the relationship between the probabilities of generation of the sibling random variables, was not able to "confuse" the algorithm in learning the unknown structure.

It should be mentioned, though, that in the cases in which the sibling probabilities were related, the learning was faster – which we believe is quite remarkable.

All these figures assure that although we have exactly identical dependence model, the random probability values between the variables in addition of having dependent or independent probabilities between the sibling variables play predominant role in determining the rate of convergence and the number of records which are used to converge to the real model.

Table 6.2: The probability values used in generating the corresponding random variables when the corresponding probabilities for the sibling nodes in the structural dependence tree are related.

Probability	value
Prob(x1=0)	0.40
Prob(x1=1)	0.60
Prob(x2=0|x1=0)	0.30
Prob(x2=0|x1=1)	0.10
Prob(x3=0|x1=0)	0.70
Prob(x3=0|x1=1)	0.90
Prob(x4=0|x1=0)	0.20
Prob(x4=0|x1=1)	0.60
Prob(x5=0|x3=0)	0.80
Prob(x5=0|x3=1)	0.40
Prob(x6=0|x3=0)	0.15
Prob(x6=0|x3=1)	0.76

Table 6.3: The probability values used in generating the corresponding random variables when the corresponding probabilities for the sibling nodes in the structural dependence tree are unrelated.

Probability	value
Prob(x1=0)	0.40
Prob(x1=1)	0.60
Prob(x2=0|x1=0)	0.30
Prob(x2=0|x1=1)	0.10
Prob(x3=0|x1=0)	0.60
Prob(x3=0|x1=1)	0.70
Prob(x4=0|x1=0)	0.20
Prob(x4=0|x1=1)	0.60
Prob(x5=0|x3=0)	0.40
Prob(x5=0|x3=1)	0.50
Prob(x6=0|x3=0)	0.15
Prob(x6=0|x3=1)	0.76

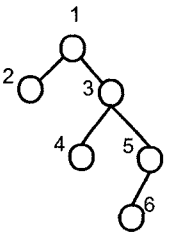
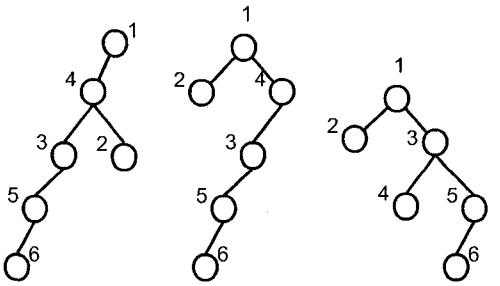
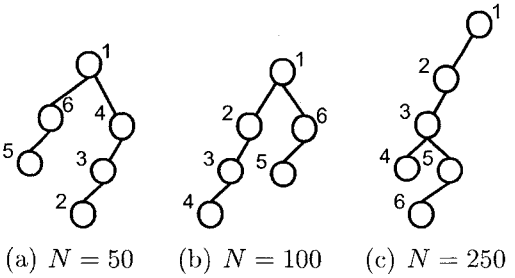


Figure 6.11: The best dependence tree for a binomial data sets with $5,000$ records and 6 variables.



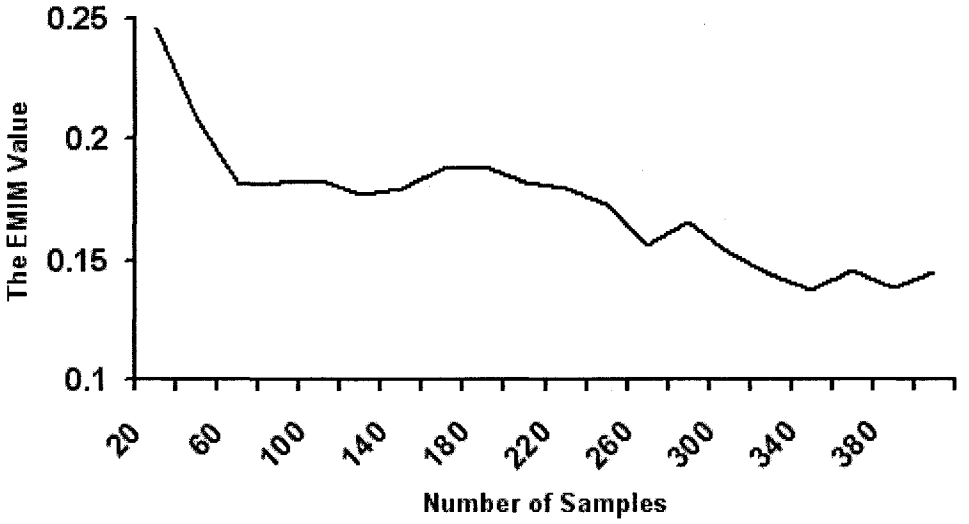
(a) $N = 50$ (b) $N = 70$ (c) $N = 80$

Figure 6.12: The "inferred" dependence tree for the binary data set as the number of samples increases. In this case, the probabilities between the sibling random variables are *related.*

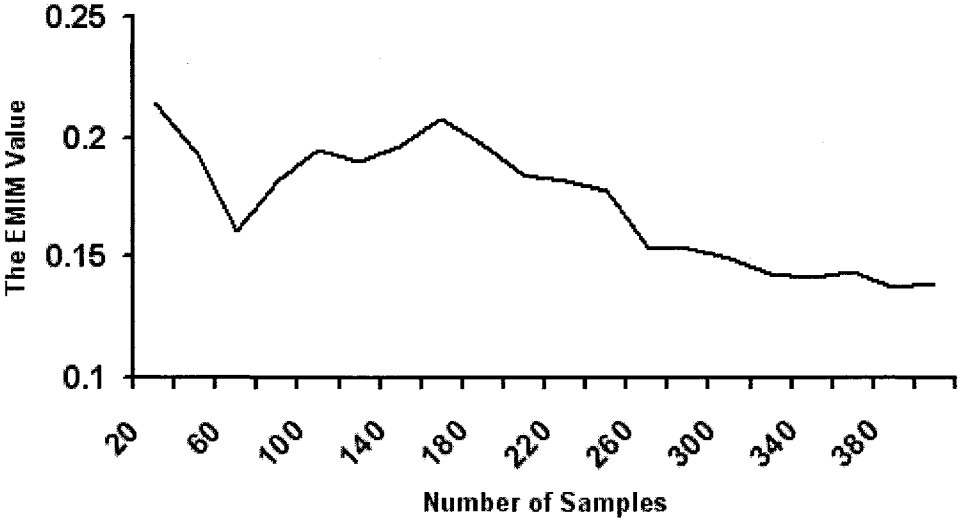


(a) $N = 50$ (b) $N = 100$ (c) $N = 250$

Figure 6.13: The "inferred" dependence tree for the binary data set as the number of samples increases. In this case, the probabilities between the sibling random variables are *unrelated.*



(a)



(b)

Figure 6.14: The convergence of the corresponding metric for the $Set-Up4$ data sets by using (a) the $EMIM$ metric to calculate the edges weights. In (a) the probabilities between the siblings are *related*, and in (b) these probabilities between the siblings are *unrelated*.

- **Experiment Sets 3:**

  The third set of experiments verified our conjecture that it was expedient to use the sub-set of the variables obtained (from the best dependence tree) by projecting the micro-data file into 3, 4 or 5 variables before invoking the multi-variate micro-aggregation process.

  Since an MAT seeks to reduce the loss in the data utility, it must be pointed out here that the value of the IL depends on the sub-set of variables used to micro-aggregate the multi-variate data file. As mentioned earlier, to infer the best sub-set of variables to be used in the micro-aggregation, we have to go through all the different projection possibilities. The results (Table 6.4) show that the estimation of the percentage value of the IL for various data sets obtained by projecting the entire data set into specified number of variables prior to invoking the $MDAV$ method. The value of the IL was bounded between the minimum value (in the fourth column) that was obtained by using the variable indices addressed in the third column, and the maximum value (in the sixth column) that was obtained by using the indices addressed in the fifth column. The last column in Table 6.4 represents the average value of the IL over all the different combinations of projected variables in the micro-data file.

  The most interesting observation was that the minimum value of the IL obtained by using 3, 4 or 5 projected variables in the Tarragona and Census data sets were exactly the same. This implies using the same "most independent variables", which in turn, preserve the same high value for the variance. Therefore, in the case of real-life data sets, we recommend projecting the entire micro-data file using 3 variables, since using a larger number of variables to project the micro-data file requires more time without leading to significant reduction in the IL value.

  Practically, due to the exponential number of combinations, we could not cover the entire solution space so as to reach to the best sub-set of the variables to be used in the micro-aggregation[3]. As opposed to this, by involving only

[3]On our processor, it took up to a few hours or even days depending on the dimensionality and

Table 6.4: The value of the IL obtained by using the $MDAV$ multi-variate MAT after projecting various data sets into the specific number of variables.

Data Set	No. of projected variables	No. of possible combinations	Indices of the variables used to obtain the min. value of IL	The min. value of IL	Indices of the variables used to obtain the max. value of IL	The max. value of IL	The average value of IL
Tarragona	1	13	10	37.6374	8	48.1006	43.1017
	2	78	11,13	24.7415	5,8	45.6925	31.6609
	3	286	2,3,10	20.7141	5,6,11	34.1569	25.1587
	4	715	2,3,10,11	20.7141	5,6,11,12	34.1569	25.4997
	5	1287	2,3,10,11,12	20.7141	5,6,11,12,13	34.1569	25.6141
Census	1	13	10	38.2133	1	62.9093	45.79787
	2	78	4,13	22.5795	1,8	55.634	31.618
	3	286	7,8,10	15.6043	1,8,9	45.815	21.2046
	4	715	7,8,10,11	15.6043	1,8,9,10	45.815	22.0308
	5	1287	7,8,10,11,12	15.6043	1,8,9,10,11	45.815	22.8299
Sim_1	1	8	2	56.6216	3	59.8342	58.6944
	2	28	1,8	46.8576	3,5	51.3179	49.4823
	3	56	2,4,7	37.6486	4,6,7	42.2961	39.7095
	4	70	2,4,7,8	37.6486	4,6,7,8	42.2961	39.5901
	5	56	1,3,5,6,7	37.6522	3,4,6,7,8	42.1577	39.7102
Sim_2	1	16	2	61.6118	11	63.0976	62.7308
	2	120	1,16	56.6698	9,13	59.0037	58.2026
	3	560	1,8,11	51.6551	6,8,9	54.3367	53.3249
	4	1820	1,8,11,12	51.6551	6,8,9,10	54.3367	53.18
Sim_3	1	22	2	62.5849	7	64.0993	63.7251
	2	231	2,22	59.0211	8,11	60.9375	60.4842
	3	1540	2,7,13	55.5274	4,6,14	57.7998	56.9827

the vertices that have the maximum number of I/O edges in the connected undirected graph to micro-aggregate the micro-data file, we were able to obtain an acceptable value of the IL close to its lower bound, and which is always (in all the cases) superior to the average value. Thus, such an automated strategy for projecting the multi-variate data sets will reduce the solution space to be searched which, in turn, reduces the computation time required to test the candidate variables, and to choose the best sub-set from them.

Tables 6.5 and 6.6 shows the percentage value of the IL obtained by using our strategy in projecting the micro-data file into sub-sets of sizes 3 and 4, respectively, prior to invoking the $MDAV$ method. When the Census data set was projected onto a number of variables prior to the micro-aggregation, the minimum values of the IL were equal to 17.47% when the width value was equal to $1,000$ and the number of variables was set to 3 or 4, to 16.23% when the width value was equal to $5,000$ and the number of variables was equal to 3 or 4. The value of the minimum IL was equal to 18.29% and to 17.70% when the width value was equal to $10,000$ and the projection was onto 3 and 4 variables, respectively. It is worth mentioning that the values obtained were quite close to the lower bound of the IL, *i.e.,* 15.60%, as shown in Table 6.4, besides being superior to the average values over all the different combinations (*i.e.,* 21.20% and 22.03% for 3 and 4 variables, respectively). Similar results were obtained for the Tarragona data set when the minimum value of the IL using 3 or 4 variables was equal to 24.13% by setting the width value to $50,000$ or $100,000$. But, it was equal to 25.05% when the width was $150,000$. Again, these values were closer to the lower bound of the IL which was 20.71%, and were superior to the average value which was close 25.5%. In Tarragona data set, the minimum values of the IL, when the width value was set to $50,000$, $100,000$ and $150,000$, were equal to 24.13%, 24.13% and 25.04%, respectively. The values obtained were quite close to the lower bound of the IL, *i.e.,* 20.71%, as shown in Table 6.4, besides being superior to the average values over all the different combinations (*i.e.,* 25.16%). Finally, we would like to state that the

cardinality of the data set, to exhaustively search the entire space.

simulated data set yielded similar results to those of the real data sets where the minimum values of the IL were equal to 38.11% for $Sim1$, 51.95% for $Sim2$ and 55.82% for $Sim3$. These values were quite close to the lower bound of the IL which were equal to 37.64% for $Sim1$, 51.65% for $Sim2$ and 55.52% for $Sim3$, respectively.

Table 6.5: The value of the IL obtained by using the $MDAV$ multi-variate MAT after projecting various data sets using 3 variables by using the $EMIM$ metric to calculate the edge weights in the connected undirected graph.

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	5	7,4,1	24.1333
			7,4,10	24.1881
			7,4,2	25.0465
			7,4,12	25.6574
			7,4,6	25.6826
	100,000	3	7,4,1	24.1333
			7,4,2	25.0465
			7,4,12	25.6574
	150,000	2	7,4,2	25.0465
			7,4,12	25.6574
Census	1,000	2	1,10,7	17.4700
			1,10,12	25.3632
	5,000	6	2,10,8	16.2332
			2,10,5	17.3421
			2,10,1	17.7012
			2,10,13	21.0694
			2,10,7	21.1128
			2,10,12	21.5828
	10,000	1	1,2,12	18.2996
Sim_1	100	2	1,3,4	51.9684
			1,3,6	52.1126
Sim_2	100	2	1,3,4	51.9684
			1,3,6	52.1126
Sim_3	100	2	1,3,5	56.1318
			1,3,18	55.8246

- **Experiment Sets 4:**

  The fourth set of experiments compares the $EMIM$ and χ^2 metrics in calculating the edge weights in the connected undirected graph. Generally speaking,

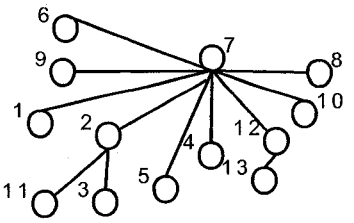
Table 6.6: The value of the IL obtained by using the $MDAV$ multi-variate MAT after projecting various data sets using 4 variables by using the $EMIM$ metric to calculate the edge weights in the connected undirected graph.

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	10	7,4,1,10	24.1333
			7,4,1,12	24.1333
			7,4,10,12	24.1881
			7,4,1,6	24.2114
			7,4,1,2	24.9648
			7,4,2,10	25.0465
			7,4,2,12	25.0465
			7,4,6,10	25.6826
			7,4,6,12	25.6826
			7,4,2,6	26.0992
	100,000	3	7,4,1,12	24.1333
			7,4,1,2	24.9648
			7,4,2,12	25.0465
	150,000	1	7,4,2,12	25.0465
Census	1,000	1	1,10,7,12	17.4700
	5,000	15	2,10,8,12	16.2322
			2,10,8,13	16.2322
			2,10,5,8	17.0012
			2,10,7,8	17.0012
			2,10,5,12	17.3421
			2,10,5,13	17.3421
			2,10,1,12	17.7012
			2,10,1,13	17.7012
			2,10,1,5	19.4846
			2,10,7,12	21.1128
			2,10,7,13	21.1128
			2,10,12,13	21.5828
			2,10,1,8	21.9116
			2,10,7,5	23.0105
			2,10,1,7	26.4757
	10,000	4	1,2,12,10	17.7012
			1,2,12,5	19.4846
			1,2,12,8	21.9116
			1,2,12,7	26.4757
Sim_1	100	1	1,3,4,6	38.1105
Sim_2	100	1	1,3,4,6	51.9684
Sim_3	100	1	1,3,5,18	56.1318

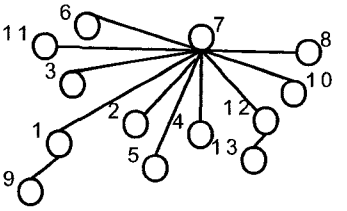
the χ^2 is faster in leading to a convergence to the best dependence tree than the $EMIM$ metric since it required a smaller number of observations or records to converge. It is worth mentioning, though, that both metrics converged to the same true structure of the dependence model for the simulated data sets by setting the value of the width parameter to 100. The scenario is completely different for the real data sets, as seen in Figures 6.15 and 6.16 which display different structures for the best dependence tree for the Tarragona and Census data sets, respectively, using various values for the width parameter. Table 6.7 shows the value of the IL obtained by invoking the $MDAV$ method after projecting various data sets into three variables by using the χ^2 metric to calculate the edges weights in the connected undirected graph. In the simulated sets, the χ^2 metric led to the same value of the IL which was obtained by using the $EMIM$ metric because they converged to the same dependence tree, implying that they used the same set of variables to micro-aggregate the micro-data file. As opposed to this, in the real data sets, the χ^2 converged to a different "best" dependence tree compared to the one obtained by using the $EMIM$ metric, thus leading to a different value of the IL. In general, the value of the IL obtained by using the χ^2 metric was lower than the corresponding value obtained by using the $EMIM$ metric for the Census data sets, but it was higher than the value obtained by using the $EMIM$ metric in Tarragona data set. Table 6.7 shows that the values of the IL for the Tarragona data set, when the width value was set to $50,000$, $100,000$ and $150,000$, were equal to 25.1%, 24.8% and 25.7%, respectively, and for the Census data set the minimum values of the IL were equal to 17.47% when the width value was set to $1,000$, 16.23% when the width value was set to $5,000$, and to 18.16% when the width value was set to $10,000$. In general, the χ^2-based solution space was superior to the $EMIM$-based solution.

- **Experiment Sets 5:**

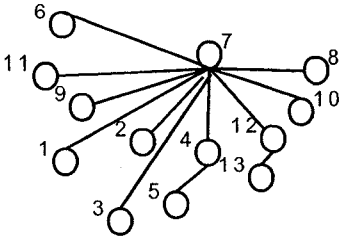
  The distribution of the average of a set of random variables tends to be Normal, even when the distribution of the individual random variable is decidedly



(a) Width = 50,000

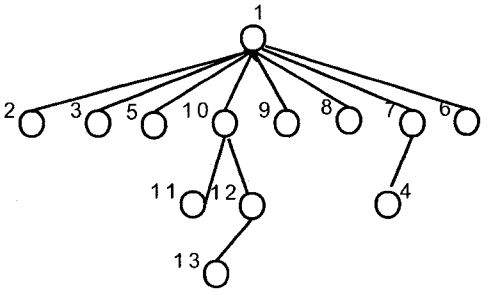


(b) Width = 100,000

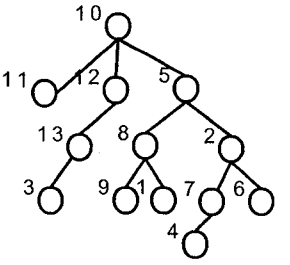


(c) Width = 150,000

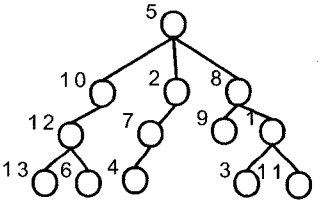
Figure 6.15: The best dependence tree for the Tarragona data set obtained by using the χ^2 metric with various values of the width parameter.



(a) Width = 1,000



(b) Width = 5,000



(c) Width = 10,000

Figure 6.16: The best dependence tree for the Census data set obtained by using the χ^2 metric with various values of the width parameter.

Table 6.7: The value of the IL obtained by using the $MDAV$ multi-variate MAT after projecting various data sets using 3 variables by using the χ^2 metric to calculate the edge weights in the connected undirected graph.

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	1	7,2,12	25.0923
	100,000	1	7,1,12	24.8137
	150,000	1	7,4,12	25.6574
Census	1,000	2	1,10,7	17.4700
			1,10,12	25.3632
	5,000	4	10,2,8	16.2322
			5,8,2	17.0012
			10,5,2	17.3421
			10,5,8	22.5430
	10,000	4	1,5,8	18.1627
			1,8,12	18.8725
			1,5,12	20.2778
			5,8,12	21.9324
Sim_1	100	2	1,3,4	51.9684
			1,3,6	52.1126
Sim_2	100	2	1,3,4	51.9684
			1,3,6	52.1126
Sim_3	100	2	1,3,5	56.1318
			1,3,18	55.8246

non-Normal. This is a consequence of the Central Limit Theorem, which is the foundation for many statistical procedures, because the distribution of the phenomenon under study does not necessarily have to be Normal. Therefore, the last set of experiments assumes the Normality of the micro-data file to quickly compute the first and second order marginals, and to thus lead to the MST for computing the best dependence tree. Subsequently, we applied our strategy to choose the subset of random variables to project the file before invoking the $MDAV$ method.

Under an assumption of normality, the edge weight, w_{ij}, between two variables V_i and V_j in the connected undirected graph can be calculated by [136]:

$$w_{ij} = -\frac{1}{2} log(1 - \rho_{ij}^2) \tag{6.11}$$

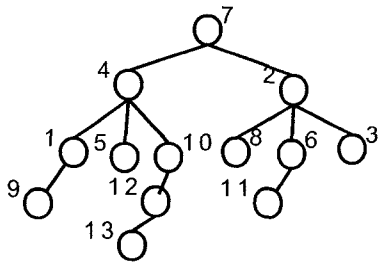
where ρ_{ij}^2 represents the correlation coefficient between the two variables V_i and V_j in the graph.

The beauty of estimating the dependence model assuming normality is that it does not depend on any parametric value. Therefore, it leads to a unique MST if the edges weight are unique. Figure 6.17 shows the best dependence tree for the simulated and real data sets. It is worth mentioning that using the correlation between two random variables in calculating the edges weights of the graph does not lead to convergence to the "true" underlying dependence model in the case of the simulated data sets. However, generally the overall process yielded a value of IL close to the minimum value of the IL after projecting the entire data set into 3 variables although the search space was greater than the search space that resulted from using the χ^2 or the $EMIM$ metrics. The minimum value of the IL was equal to 23.10% for Tarragona data set, 16.34% for Census data set, 37.8% for $Sim1$, 51.96% for $Sim2$, and 55.64% for $Sim3$.

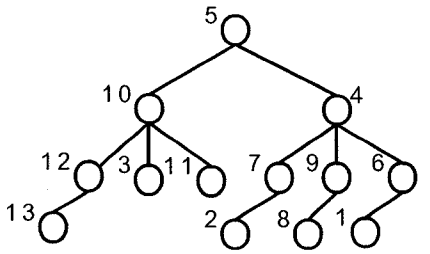
Finally, we conclude by stating that each method of calculating the edges weights has its own advantages and disadvantages. We believe that, in practice, the user is the only one who is capable of deciding which is the most suitable metric for the specific

Table 6.8: The value of the IL obtained by using the $MDAV$ multi-variate MAT after projecting various data sets into 3 variables assuming normality.

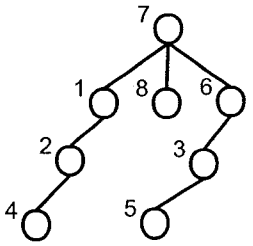
Data set	No. of possibilities	Variable indices	IL
Tarragona	5	2,4,10	23.1068
		2,4,1	24.9648
		2,4,7	25.0465
		2,4,12	25.9818
		2,4,6	26.0992
Census	5	4,10,5	16.3416
		4,10,9	16.8352
		4,10,6	19.6712
		4,10,7	20.2959
		4,10,12	20.9725
Sim_1	6	7,2,3	37.8147
		7,1,3	37.8676
		7,1,6	38.1610
		7,2,6	38.1704
		7,3,6	41.4843
		7,1,2	42.0634
Sim_2	3	4,1,3	51.9684
		4,1,12	52.0159
		4,3,12	53.9267
Sim_3	10	20,1,5	55.6419
		20,2,13	55.6631
		20,2,5	55.8020
		20,2,3	55.8722
		20,1,3	55.9010
		20,1,13	55.9474
		20,5,13	57.1943
		20,3,5	57.3903
		20,3,13	57.4192
		20,1,2	57.4770



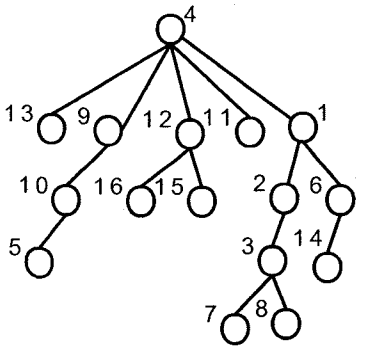
(a) *Tarragona*



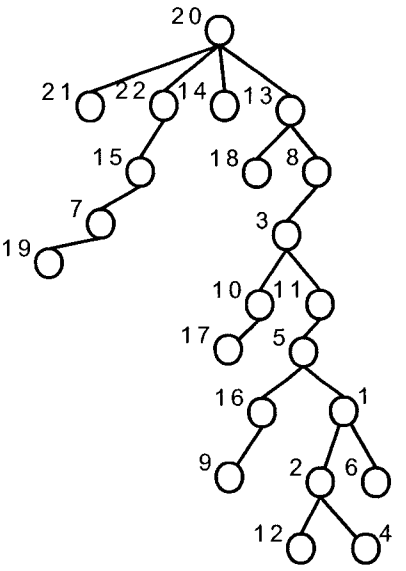
(b) *Census*



(c) *Sim_1*



(d) *Sim_2*



(e) *Sim_3*

Figure 6.17: The best dependence tree for the real and simulated data sets assuming normality.

data sets. Table 6.9 summarizes the characteristics of each metric in calculating the edge weights of the connected undirected graph.

Table 6.9: Characteristics of the $EMIM$, χ^2 and Correlation metrics in calculating the edges weights of the connected undirected graph.

	EMIM	χ^2	*Correlation*
Width parameter	Sensitive	Sensitive	Not sensitive
No. of Combinations in search space	Medium	Small	Large
Convergence to the best dependence tree structure	Converge	Converge	Does not always converge
Convergence speed	Slower than χ^2 metric	Slower than assuming Normality	Faster than both metrics

6.4 Conclusions

In this chapter we have shown how the information about the structure of the dependence between the variables in the micro-data file can be used as a fundamental indicator before invoking any MAT. By using this information, we have proposed a new automated scheme as a pre-processing phase to determine the number and the identity of the variables that are to be used to micro-aggregate the micro-data file. This is achieved by constructing a connected undirected graph whose nodes represent the random variables in the micro-data file, edges represent the statistically dependencies, and the edges weights are computed either by using the $EMIM$, χ^2 or the correlation values when a normality constraint is assumed. The experimental results show that such a methodology involving projecting the multi-variate data sets reduces the solution space, which further directly reduces the computation time required to search the entire space combinatorially. In spite of this, this methodology leads to a solution whose IL values are close to the minimum value of the IL that can be obtained by exhaustively searching over the entire search space.

Chapter 7

Conclusion and Future Research

7.1 Overall Contributions

With the increasing ability of users to access data and information, privacy and security have become primary concerns. Simultaneously, the need to access detailed statistical information is becoming mandatory for various governmental and private companies. This Thesis dealt with the issue of security and the efficient utility of data in the protection of micro-data files in statistical agencies. We specifically considered the problem of preventing the disclosure of confidential information concerning particular individuals without significantly harming the utility of the data being protected. In this context, we argued that the DR and IL, and a composite newly-devised metric are among the most important metrics for protecting the statistical data.

The research in this Doctoral Thesis focused on the family of $MATs$ which include recent well-known techniques in the area of SDC for micro-data. The rationale behind these techniques is the confidentiality rules of dissemination that allow publication if, and only if, each record in the published file corresponds to a group of a size equal to k or more individuals, where no individual in the original file dominates the group excessively. In general, data aggregation is not only the central concept in

data security and privacy, but also in many applications in computer science, such as artificial intelligence and data mining.

This section summarizes the work presented in this Thesis, and gives the overall conclusions of each chapter, and the lists of areas where further research can be conducted. In all brevity we mention that while Chapter 1 introduced the overall Thesis, and listed its main motivation and objectives, Chapter 2 embarked on a fairly comprehensive survey of the field of SDC techniques in general and, more specifically, on the family of $MATs$. In addition to covering the different approaches to measure the IL and DR, Chapter 2 described the existing methods available to resolve their conflicting goals.

The main contributions of the Thesis involve:

1. Enhancing the Heuristic k-Ward MAT.
2. Using Fixed Structure LA to Solve the MAP.
3. Using NN Methods to Solve the MAP.
4. Utilizing Dependence-Based Information for MATs.

We shall visit each of these in the next section.

7.2 Enhancing the Heuristic k-Ward MAT

7.2.1 Contributions

This chapter presented an enhancement to the k-ward MAT with respect to the required computational time. This was achieved by, first of all, minimizing the computational burden involved in evaluating the between-class distance matrix, and also by recursively partitioning the entire data set before invoking the standard k-Ward

method. The latter optimization partitioned the data set using either a fixed threshold or a dynamic one. In this chapter we showed that invoking these two optimization strategies simultaneously could reduce the computational time up to 90%. The various optimization strategies were applied to uni-variate data sets, and also on multi-variate data sets projected using the first principal component or the sum of Z-scores.

7.2.2 Future work

We foresee the following avenues for future research.

1. The question of investigating how these methods can be used for *multi-variate* online statistical databases is open.

2. Clearly, a shortest path algorithm will lead to an optimal solution. We submit that the philosophy that we have introduced in Chapter 3 is also valid for shortest-path-based algorithms. Indeed, the question of whether the shortest path for the graph can be approximately computed *quickly* using recursive invocations of shortest paths for *its* sub-graphs is a very interesting problem. Such a study remains open.

3. We believe that our recursive strategy can be rendered applicable to a "multi-variate k-Ward" heuristic that does not rely on such a uni-variate projection. Although we are currently considering how our matrix optimizing operations can be applied to the scheme proposed in [45], it does not seem to be straightforward. This work is still open.

4. It is plausible that our proposed time-based optimization could induce some statistical bias in the way the data is aggregated. It would be interesting to see what impact this has with respect to the privacy issue. However, we add in passing, that such a bias-based analysis will be data and distribution dependent and, is thus, far from trivial.

5. Problems related to statistical inference in databases have been solved with a variety of tools [2, 12, 35, 36, 37, 97, 114]. Incorporating these tools into our present framework would help to describe the limitations of our approach and/or the types of data sets that are, or are not, well suited for such an aggregation technique. Interestingly, this could involve the extension of our mechanism for discrete or ordinal data elements. All of these are open problems, warranting research "projects" in their own right!

7.3 Using Fixed Structure LA to Solve the MAP

7.3.1 Contribution

This chapter proposed the use of a sophisticated learning algorithm, namely the $OMMA$, to solve the MAP and optimize the MAT. The method was based on the notion of specifying various levels of uncertainties for pairs of individuals belonging to the same group. The $OMMA$ increased or decreased these uncertainty levels depending on whether a known similar pair is "correctly" or "incorrectly" clustered by the algorithm. The new algorithm was compared extensively to another algorithm that was developed and implemented in the μ-Argus software package ($MDAV$). It yielded the same results as the latter for the uni-variate case, and showed improvement for multi-variate micro-aggregation. The improvement in the value of the IL measure was about 10% on real data sets, and up to 13% on simulated data sets. Preliminary tests showed that the proposed strategy scales well with respect to the size of the data set, the dimensionality, and the group size. By defining a composite measure involving the IL and DR, we proved that proposed strategy also obtains a minimum score value when compared to the $MDAV$ method. This indicates that the $OMMA$ technique is probably the best MAT not only from the IL perspective, but also from the viewpoint of a measure which was a combination of the IL and DR.

7.3.2 Future Work

We foresee the following avenues for future research.

1. It will be interesting to see how we can extend the $OMMA$ for the data-oriented micro-aggregation, where the group size, n_i, satisfies $k \leq n_i < 2k$.

2. The question of studying whether the minimum group size constraint can be treated as a "soft constraint" associated with a certain preference level is open.

3. The problem of incorporating other families of learning automata, especially those of continuous variable structure types, with a fixed-size or variable-size group constraints, is also unsolved.

4. The question of applying the concepts of hierarchical $OMMA$ for a large data set is also an interesting avenue for future work.

5. Finally, it would be good if some studies were conducted to compare the $OMMA$ to the $MDAV$ with respect to the computational time for large real data sets, typically containing more than $100,000$ records.

7.4 Using NN Methods to Solve the MAP.

7.4.1 Contribution

This chapter provided a practical methodology (for micro-aggregating the micro-data file) that is based on the theory of NNs and their novel clustering properties. Our newly-proposed scheme, referred to as the $IAMAT$, introduced the Association and the Interaction between records as a new metric for measuring the similarity. The $IAMAT$ strategy preserved the meaning of the closeness between the records so as to coalesce them into groups, and to also maximize the actual Interaction among the records inside each group. In general, incorporating the concepts of the Interaction

among the records and their mutual Association, had been shown to be advantageous to solving the MAP, because they have the effect of minimizing the IL. Our experiments clearly showed a significant improvement to the state-of-the-art $MDAV$. The improvement in the value of the IL measure was about 8% on real data sets and up to 14% on simulated data sets. This indicated that the resulting clusteres results in better grouping compared to the classical Euclidean distance approach. Preliminary tests showed that the new strategy scales well with respect to the cardinality of the data set, the dimensionality, and the group size. Again, the method also yields a superior SI value when compared to the $MDAV$ method. This indicated that the $IAMAT$ is probably one of the best $MATs$ not only from the IL perspective, but also from the viewpoint of a measure which is a combination of the IL and DR.

7.4.2 Future Work

We foresee two avenues for future research.

1. It would be fascinating to study how we could extend the $IAMAT$ towards data-oriented MAT, where the group size, n_i, satisfies $k \leq n_i < 2k$.

2. The question of investigating the effect of having a dynamic value of α on the compactness of each group and on the value of the IL remains unsolved. We believe that we can compute the dynamic value of α in any iteration based on the remaining unassigned records at the end of every iteration in the clustering phase. As opposed to this, the fixed value is computed once before invoking the clustering phase, and is based on the entire data set.

7.5 Utilizing Dependence-Based Information for MATs

7.5.1 Contribution

This chapter discussed how the presence and structure of dependency between a set of random variables in the micro-data file is valuable information that can be used as a fundamental indicator before invoking any MAT. We proposed a new automated scheme as a pre-processing phase to determine the number and the identity of the variables that are to be used in the micro-aggregation phase. This is achieved by constructing a completely connected, undirected graph whose nodes represent the random variables in the micro-data file, whose edges represent the statistically dependencies, and whose edge weights are computed by using either an information theoretic, χ^2 or correlation-based measure. The advantageous and disadvantages of these respective measures were investigated. The experimental results showed that such a methodology involving projecting the multi-variate data sets reduces the solution space, which further directly reduces the computation time required to search the entire space combinatorially. Additionally, this methodology leads to a solution whose IL values were close to the minimum value of the IL obtained by *exhaustively* searching over the entire search space. In general, this chapter presented an entirely new approach to tackling a tough problem for which no prior solution was available.

7.5.2 Future Work

We foresee three avenues for future research.

1. We believe that our newly automated scheme can be rendered applicable to *any* known MAT. Therefore, it will be interesting to see how it can be applied to the $OMMA$, and the $IAMAT$.

2. An investigation on the effect of this methodology on the value of the DR would also be interesting.

3. As argued above, the use of dependence information can help to reduce the required computational time in the aggregation stage, so as to effectively measure the distance between the records. We believe that capturing such dependence information is not only important to $MATs$, but also to other SDC techniques, such as data swapping. However, since data swapping annihilates multi-variate relationships - such as the regression and the correlations between two or more variables [96] - the question of considering how the swapping can be used on a *selected* set of dependent variables without disturbing or removing any mutual relationships, remains open.

7.6 Future Research Avenues

From the literature we have reviewed, it appears as if there is a scarcity of good *practical* methods for $MATs$ which are based on perturbative techniques. We do not imply that there are no methods available, or that none of them perform well. On the contrary, as mentioned in Chapter 2, numerous methods have been proposed and studied, especially in the recent past, where novel schemes for specific applications have been developed. But based on the fact that this field is relatively new, none of these methods have attained to such a degree of maturity so as to grant it the right to be called a "standard benchmark technique" worthy to be imbedded as a built-in function and labeled as being the "optimal" strategy. However, some of the available methods are very promising, and, generally speaking, using an adaptive clustering method in an MAT represents a very positive approach. This is because it will reduce the area of the search space significantly by enforcing the group size constraint [15, 79, 130, 138]. Additionally, combining data mining tools with the MAT can be very beneficial since a resulting scheme can yield a good trade-off between the DR and the IL criteria [59, 131, 137].

Several overall research directions for further investigation arise from this Thesis. They are discussed below.

- The need to further merge the areas of LA and micro-aggregation so as to devise a novel MAT, which can provide variable-size groups for a multi-variate data set, is open. We know that the main challenge in micro-aggregation is to design good heuristics which minimize the IL and the DR. This can be achieved by providing a variable-group size which, in turn, implies a lower loss in the data utility because it generates more homogeneous groups. This, of-course, also leads to the preservation of the natural data aggregation by allowing the group sizes to be between k and $2k-1$. In addition, a variable-group size will result in decreasing the likelihood of the information being compromised. However, such a variable-group size constraint would directly imply a more complex implementation than the schemes which constrain a fixed-size, as explained in Chapter 4.

  The primary disadvantage of the current $OMMA$ scheme is that the micro-data file is to be equi- partitioned. With a little insight it is not trivial to translate the fixed- size constraint into the variable-size constraint. Hence, this improvement will decrease the IL and the DR. Generalizing the $OMMA$ to handle the variable group size constraint will increase the difficulty of the problem, since we have no formal method of determining how to characterize the optimal solution. The main concept utilized in the $OMMA$ version was the limit imposed on the number of records in each group. The purpose of setting this limit was to prevent having any interleaving groups. But if variable-group size is used, this limit is no longer valid since each class may or may not have exactly the same number of groups. In other words, in the current version we assumed a knowledge of the number of the groups. In contrast, in a future study, we would seek a solution, which requires the number of groups to be bounded between $n/(2k-1)$ and n/k.

  Due to the fact that each group may have a different number of records (*i.e.*, between k and $2k-1$), the religious invocation of the $OMMA$ is not applicable. With a little insight, it is easy to see that the scenario encountered for a reward does not pose any explicit problem (see Section 4.3). This is because enforcing

the reward condition does not alter the basic constitution of the groups. On the other hand, the problems encountered on enforcing a penalty are numerous which deserve further investigation.

- Numerous government organizations, such as Statistics Canada, aim to release useful statistical summaries for public use and for analysis, without violating the confidentiality requirements. An extremely fascinating avenue for future work is to focus on applying our newly-developed techniques to such real-life data sets. The main goal of such studies will be to test the applicability of the applied methodology in a real scenario, and to assess their ability to overcome the relevant limitations and handicaps of the currently-used methods. Besides this, it would be beneficial to analyze the effect of the new micro-aggregation schemes from the perspectives of regression, dependency, and linear models. Such an investigation would also attempt to extensively study the effect of the micro-aggregation process on the IL, DR, and the SI.

- The wavelet transformation possesses the characteristic of providing spatial and frequency domain information, which play a predominant rule in many applications such as image compression, watermarking, and clustering. An integration between the fields involving wavelet transformation and micro-aggregation would prove to be extremely rewarding. To the best of our knowledge, we are not aware of any MAT that utilizes the characteristics of wavelets during the micro-aggregation process. We believe that such a transformation can guide the process of micro-aggregation by operating in the transformed feature space, and by thus identifying clusters at different levels of the transformed space. The goal of such a research endeavor would be to understand how this integration can be achieved while simultaneously preserving almost the same computational time and minimizing the loss in the data utility and the risk of disclosing the confidentiality.

- The processes of updating, deleting, or inserting individual records in the micro-data file causes a serious problem in the problem of micro-aggregation. This, in turn, implies a large burden on the time and storage required to re-compute

the micro-aggregated file after the change. The question of devising $MATs$ in the presence of time varying data files remain open, and would definitely lead to many future research projects and theses.

Bibliography

[1] M. Adachi and K. Aihara. Associative Dynamics in a Chaotic Neural Network. *Neural Networks*, 10(1):83–98, 1997.

[2] N. Adam and J. Wortmann. Security-Control Methods for Statistical Databases: A Comparative Study. *ACM Computing Surveys*, 21(4):515–556, 1989.

[3] M. Agache and B. Oommen. Generalized Pursuit Learning Schemes: New Families of Continuous and Discretized Learning Automata. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 32(6):738–749, 2002.

[4] C. Aggarwal and P. Yu. *Privacy-Preserving Data Mining.* Springer, 2008.

[5] C. Aggarwal and P. Yu. *Privacy-Preserving Data Mining: Models and Algorithms (Advances in Database Systems).* Springer Science and Business Media L.L.C., 2008.

[6] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *In Proceedings of ACM SIGMOD*, pages 207–216, USA: Washington DC, 1993.

[7] R. Agrawal, H. Mannila, H. Srikant, R. Toivonen, and I. Verkamo. Fast Discovery of Association Rules. *In Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

[8] N. Anwar. Micro-Aggregation - The Small Aggregates Method. Technical report, Luxembourg: Eurostat, 1993.

[9] F .Bacao, V .Lobo, and M. Painho. Self-organizing Maps as Substitutes for K-Means Clustering. In *International Conference on Computational Science*, pages 476–483, 2005.

[10] Y. Baeyens and D. Defays. Estimation of Variance Loss Following Microaggregation by the Individual Ranking Method. In *Proceedings of Statistical Data Protection'98*, pages 101–108, Luxembourg: Office for Official Publications of the European Communities, 1999.

[11] L. Bechet and P. Loosveldt. Development of Micro-aggregation Software. Technical report, Ariane II,275 route d'Arlon, L-8011 Strassen, Luxembourg, 2000.

[12] L. Beck. A Security Mechanism for Statistical Database. *ACM Trans. Database Syst.*, 5(3):316–3338, 1980.

[13] R. Brand. Microdata Protection through Noise Addition. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 97–116, London, UK, 2002. Springer-Verlag.

[14] R. Brand, J. Domingo-Ferrer, and J. Mateo-Sanz. Reference Data Sets to Test and Compare SDC Methods for Protection of Numerical Microdata. Technical report, CASC PROJECT, Computational Aspects of Statistical Confidentiality, 2002.

[15] P. Brucker. On the Complexity of Clustering Problems. In R. Hehn, B. Korte, and W. Oettli, editors, *Optimization and Operations Research*, pages 45–54. Berlin: Springer-Verlag, 1977.

[16] S. Castano, M. Fugini, G. Martella, and P. Samarati. *Database security.* Addison-Wesley, 1994.

[17] C. Chow and C. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. Information Theory*, 14(11):462–467, 1968.

[18] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms.* MIT Press, McGraw-Hill, 1990.

[19] G. Crises. Additive Noise for Microdata Privacy Protection in Statistical Databases. Technical report, 2004.

[20] G. Crises. An Introduction to Microdata Protection for Database Privacy. Technical report, 2004.

[21] G. Crises. Information Loss Measures for Microdata in Database Privacy Protection. Technical report, 2004.

[22] G. Crises. Microaggregation for Privacy Protection in Statistical Databases. Technical report, 2004.

[23] G. Crises. Microdata Disclosure Risk in Database Privacy Protection. Technical report, 2004.

[24] G. Crises. Non-Perturbative Methods for Microdata Privacy in Statistical Databases. Technical report, 2004.

[25] G. Crises. Perturbative Masking for Microdata Privacy Protection in Statistical Databases. Technical report, 2004.

[26] G. Crises. Synthetic Microdata Generation for Database Privacy Protection. Technical report, 2004.

[27] G. Crises. Trading off Information Loss and Disclosure Risk in Database Privacy Protection. Technical report, 2004.

[28] M. Cuppen. *Secure Data Perturbation in Statistical Disclosure Control.* PhD thesis, Statistics Netherlands, 2000.

[29] R. Dandekar, M. Cohen, and N. Kirkendall. Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 117–125, London, UK, 2002. Springer-Verlag.

[30] C. Date. *An Introduction to Database Systems.* Addison-Wesely, 2000.

[31] D. Defays. Protecting Microdata by Microaggregation: the Experience in Eurostat. *Questiio*, 21:221–231, 1997.

[32] D. Defays and M. Anwar. Masking Micro-data Using Micro-Aggregation. *Journal of Official Statistics*, 14(4):449–461, 1998.

[33] D. Defays and N. Anwar. Micro-Aggregation: A Generic Method. In *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, pages 69–78, Luxembourg: Office for Official Publications of the European Communities, 1995.

[34] D. Defays and P. Nanopoulos. Panels of Enterprises and Confidentiality: the Small Aggregates Method. In *Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, Ottawa: Statistics Canada, 1993.

[35] D. Denning. Secure Statistical Databases with Random Sample Queries. *ACM Trans. Database Syst.*, 5(3):291–315, 1980.

[36] D. Denning and P. Denning. Data Security. *ACM Comput. Surv.*, 11(3):227–249, 1979.

[37] D. Denning and P. Denning. The Tracker: A Threat to Statistical Database Security. *ACM Trans. Database Syst.*, 4(1):76–96, 1979.

[38] J. Domingo-Ferrer. Statistical Disclosure Control in Catalonia and the CRISES Group. Technical report, 1999.

[39] J. Domingo-Ferrer. Microaggregation: Achieving k-Anonymity with Quasi-Optimal Data Quality. In *Proceedings of Q2006: European Conference on Qulity in Survey Statistics*, 2006.

[40] J. Domingo-Ferrer. Microaggregation for Database and Location Privacy. volume 4032, pages 106–116, 2006.

[41] J. Domingo-Ferrer. Microaggregation for Database and Location Privacy. In *Next Generation Information Technologies and Systems-NGITS'2006*, pages 106–116, 2006.

[42] J. Domingo-Ferrer, A. Martínez-Ballesté, J. Mateo-Sanz, and F. Sebé. Efficient Multivariate Data-Oriented Micro-Aggregation. *The International Journal on Very Large DataBases, (VLDB)*, 15(4):355–369, Sep. 2006.

[43] J. Domingo-Ferrer and J. Mateo-Sanz. On Resampling for Statistical Confidentiality in Contingency Tables. *Computers and Mathematics with Applications*, 38:13–32, 1999.

[44] J. Domingo-Ferrer and J. Mateo-Sanz. An Empirical Comparison of SDC Methods for Continuous Micro-data in Terms of Information Loss and Disclosure Risk. In *Joint ECE/Eurostat Work Session on Statistical Data Confidntiality, Conference of European Satisticians*, Statistical Commision and Economic Commision for Europe, 2001.

[45] J. Domingo-Ferrer and J. Mateo-Sanz. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[46] J. Domingo-Ferrer, J. Mateo-Sanz, A. Oganian, V. Torra, and A. Torres. On The Security of Microaggregation with Individual Ranking: Analytical Attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):477–491, 2002.

[47] J. Domingo-Ferrer, J. Mateo-Sanz, and F. Sebe. Information Loss in Continuous Hybrid Microdata: Subdomain-Level Probabilistic Measures. In *Soft Computing for Information Retrieval on the Web: Models and Applications, Studies in Fuzziness*, pages 287–298, 2006.

[48] J. Domingo-Ferrer, J. Mateo-Sanz, and V. Torra. Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk. In *Proc. of*

ETK-NTTSŠ2001, volume 2, pages 807–825, Luxembourg: Office for Official Publications of the European Communities, 2001.

[49] J. Domingo-Ferrer and F. Sebé. Optimal multivariate 2-microaggregation for microdata protection: a 2-approximation. In *Privacy Statistical Databases*, pages 129–138, Italy: Rome, 2006.

[50] J. Domingo-Ferrer, F. Sebé, and J. Castellà-Roca. On the Security of Noise Addition for Privacy in Statistical Databases. pages 149–161, Spain: Barcelona, 2004. Berlin: Springer-Verlag.

[51] J. Domingo-Ferrer, F. Sebe, and A. Solanas. A Polynomial-Time Approximation to Optimal Multivariate Microaggregation. *Computers and Mathematics with Applications*, 55:714–732, 2008.

[52] J. Domingo-Ferrer, A. Solanas, and A. Martínez-Balleste. Privacy in Statistical Databases: k-Anonymity Through Microaggregation. In *Proceedings of IEEE Granular Computing 2006*, pages 774–777, 2006.

[53] J. Domingo-Ferrer and V. Torra. A Quantitative Comparison of Disclosure Control Methods for Microdata. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 113–134, Amesterdam: North-Holland, 2002. Berlin: Springer-Verlag.

[54] J. Domingo-Ferrer and V. Torra. Aggregation Techniques for Statistical confidentiality. In *Aggregation operators: new trends and applications*, pages 260–271, Germany: Heidelberg, 2002. Physica-Verlag GmbH.

[55] J. Domingo-Ferrer and V. Torra. Disclosure Control Methods and Information Loss for Microdata. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 93–112, Amsterdam: North-Holland, 2002.

[56] J. Domingo-Ferrer and V. Torra. Extending Microaggregation Procedures Using Defuzzification Methods for Categorical Variables. In *IEEE Intelligent Systems'2002*, pages 44–49, Bulgaria: Varna, 2002.

[57] J. Domingo-Ferrer and V. Torra. Fuzzy Microaggregation for Microdata Protection. *Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII)*, 7(2):153–159, 2003.

[58] J. Domingo-Ferrer and V. Torra. Median-Based aggregation operators for Prototype Construction in Ordinal Scales. *International Journal of Intelligent Systems*, 18(6):633–655, 2003.

[59] J. Domingo-Ferrer and V. Torra. On the Connections Between Statistical Disclosure Control for Microdata and Some Artificial Intelligence Tools. *Information Sciences*, 151:153–170, 2003.

[60] J. Domingo-Ferrer and V. Torra. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[61] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* A Wiley-Interscience Publication, 2000.

[62] A. El-Hamdouchi and P. Willet. Hierarchic Document Clustering Using Ward's Method. In *In proceedings of the Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 149–156, 1986.

[63] M. Elliot. Integrating File and Record Level Disclosure Risk Assessment. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 126–134, London, UK, 2002. Springer-Verlag.

[64] E. Fayyoumi and B. Oommen. A Fixed Structure Learning Automaton Micro-Aggregation Technique for Secure Statistical Databases. In *Privacy Statistical Databases*, pages 114–128, Italy: Rome, 2006.

[65] E. Fayyoumi and B. Oommen. On Optimizing the k-Ward Micro-Aggregation Technique for Secure Statistical Databases. In *11th Austratasian Conference on Information Security and Privacy Proceeding*, pages 324–335, Australia: Melbourne, 2006.

[66] E. Feige and H. Watts. Protection of Privacy Through Microaggregation. In *Data bases, Computers and the Social Sciences*, pages 261–272. Wiley, 1970.

[67] F. Felso, J. Theeuwes, and G. Wagner. Disclosure Limitation Methods in Use: Results of A Survey. In *Confidentiality, Disclosure and Data Access*, pages 17–42, Amsterdam: North-Holland, 2001.

[68] L. Feng, T. Dillon, H. Weigana, and E. Chang. An XML-Enabled Association Rule Framework. In *In Proc. of DEXA'03*, pages 88–97, Prague: Czech Republic, 2003.

[69] E. Fernandez, R. Summers, and C. Wood. *Databases Security and Integrity.* Addison-Wesley, 1980.

[70] W. Fuller. Masking Procedures for Microdata Disclosure Limitation. *Journal of Official Statistics*, 9(2):383–406, 1993.

[71] W. Gale, S. Das, and C. Yu. Improvements to an Algorithm for Equipartitioning. *IEEE Trans. Comput.*, 39(5):706–710, 1990.

[72] S. Giessing and A. Hundepool. The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality. Technical report, 2003.

[73] J. Gouweleeuw, P. Kooiman, L. Willenborg, and P. Wolf. Post Randomization for Statistical Disclosure Control: Theory and Implementation. Technical report, 1997.

[74] S. Hansen and S. Mukherjee. A Polynomial Algorithm for Univariate Optimal Microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1043–1044, 2003.

[75] G. Horn and B. Oommen. A Fixed-Structure Learning Automaton Solution to the Stochastic Static Mapping Problem. In *19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05)*, pages 297–304, 2005.

[76] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Nordholt, G. Seri, and P. Wolf. *Handbook on Statistical Disclosure Control.* a CENtre of EXcellence for Statistical Disclosure Control CENEX SDC, 2006.

[77] A. Hundepool, A. Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P. Wolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. *M-ARGUS Version 4.0 Software and User's Manual*, 2004.

[78] A. Hundepool, L. Wetering, L. Gemerden, A. Wessels, M. Fischetti, J. Salazar, and A. Caprara. *t-ARGUS User's Manual*, 1998.

[79] A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[80] W. Jonge. Compromising statistical Databases Responding to Queries About Means. *ACM Trans. Database Syst.*, 8(1):60–80, 1983.

[81] K. Kenthapadi. DIMACS Working Group on Challenges for Cryptographers in Health Data Privacy. Technical report, 2004.

[82] J. Kim and W. Winkler. Masking Microdata Files. In *Proceedings of the Section on Survey Research Methods*, pages 114–119, 1995.

[83] P. Kooiman, L. Willenborg, and J. Gouweleeuw. PRAM: A Method for Disclosure Limitation of Microdata. Technical report, 1998.

[84] M. Laszlo and S. Mukherjee. Minimum Spanning Tree Partitioning Algorithm for Microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.

[85] Y. Li, S. Zhu, L. Wang, and S. Jajodia. A Privacy-Enhanced Microaggregation Method. In *FoIKS '02: Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems*, pages 148–159, London, UK, 2002. Springer-Verlag.

[86] S. Mandujano. Inference attacks to statistical databases: Data suppression, concealing controls and other security trends. Technical report, Purdue University, Department of Computer Science, 2001.

[87] M. Markey, J. Lo, G. Tourassi, and C. Floyd Jr. Self-Organizing Map for Cluster Analysis of A Breast Cancer Database. *Artificial Intelligence in Medicine*, 27:113–127, 2003.

[88] A. Martinez-Balleste, A. Solanas, J. Domingo-Ferrer, and J. Mateo-Sanz. A Genetic Approach to Multivariate Microaggregation for Database Privacy. In *ICDE Workshops*, pages 180–185, 2007.

[89] M. Mas. Statistical Data Protection Techniques. Technical report, Eustat: Euskal Estatistika Erakundea,Instituto Vasco De Estadistica, 2006.

[90] J. Mateo, J. Domingo-Ferrer, F. Sebé, A. Martinez-Balleste, A.Torres, and N.Macia. New Microaggregation Algorithms Software Documentation and Related Papers. Technical report, CASC PROJECT, Computational Aspects of Statistical Confidentiality, 2002.

[91] J. Mateo-Sanz and J. Domingo-Ferrer. A Comparative Study of Microaggregation Methods. *Questiio*, 22(3):511–526, 1998.

[92] J. Mateo-Sanz and J. Domingo-Ferrer. A Method for Data-Oriented Multivariate Microaggregation. In *Proceedings of Statistical Data Protection'98*, pages 89–99, Luxembourg: Office for Official Publications of the European Communities, 1999.

[93] J. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé. Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. *Data Mining and Knowledge Discovery*, 11(2):181–193, 2005.

[94] J. Mateo-Sanz, F. Sebé, and J. Domingo-Ferrer. Outlier Protection in Continuous Microdata Masking. pages 201–215, Spain: Barcelona, 2004. Berlin: Springer-Verlag.

[95] S. Misra and B. Oommen. Dynamic Algorithms for the Shortest Path Routing Problem: Learning Automata-Based Solutions. *IEEE Trans Syst Man Cybern B Cybern*, 35(6):1179–1192, 2005.

[96] R. Moore. Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets. Technical report, Statistical Research Division Report Series, Washington D. C., 1996.

[97] K. Muralidhar and R. Sarathy. Security of Random Data Perturbation Methods. *ACM Trans. Database Syst.*, 24(4):487–493, 1999.

[98] K. Narendra and M. Thathachar. *Learning Automata: An Introduction.* New Jersey: Prentice-Hall, 1989.

[99] K. Narendra, E. Wright, and L. Mason. Application of Learning Automata to Telephone Traffic Routing and Control. *IEEE Trans Syst Man Cybern B Cybern*, SMC-7(11):785–792, 1977.

[100] J. Nin and V. Torra. Modeling Projections in Microaggregation. In *proceedings of the* 12^{th} *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System (IPMU)*, pages 138–145, 2008.

[101] E. Nordholt. Statistical Disclosure Control from A Users' Point of View. Technical report, Statistics Netherlands, 2001.

[102] A. Oganian. *Security and Information Loss in Statistical Data Protection.* PhD thesis, University URV Catalunya, 2002.

[103] A. Oganian and J. Domingo-Ferrer. On The Complexity of Optimal Microaggregation for Statistical Disclosure Control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18(4):345–354, 2001.

[104] A. Oganian and F. Karr. Combinations of SDC Methods for Microdata Protection. In *Privacy Statistical Databases*, pages 102–113, Italy: Rome, 2006.

[105] B. Oommen and S. Croix. String Taxonomy Using Learning Automata. *IEEETSMC: IEEE Transactions on Systems, Man, and Cybernetics*, 27(2):354–365, 1997.

[106] B. Oommen and E. Fayyoumi. An AI-Based Causal Strategy for Securing Statistical Databases Using Micro-Aggregation. In *21st Austratasian Joint Conference on Artificial Intelligence, To appear.*

[107] B. Oommen and E. Fayyoumi. A Novel Method for Micro-Aggregation in Secure Statistical Databases Using Association and Interaction. In *Information and Communications Security, 9th International Conference on Information and Communications Security, LNCS 4861, Springer Verlag*, pages 126–140, 2007.

[108] B. Oommen and E. Fayyoumi. Enhancing Micro-Aggregation Technique by Utilizing Dependence-Based Information in Secure Statistical Databases. In *Information Security and Privacy, 13th Austratasian Conference on Information Security and Privacy Proceeding, LNCS 5107, Springer Verlag*, pages 404–418, 2008.

[109] B. Oommen and C. Fothergill. Fast Learning Automaton-Based Image Examination and Retrieval. *The Computer Journal*, 36(6):542–553, 1993.

[110] B. Oommen and J. Lanctot. Discretized Pursuit Learning Automata. *Systems, Man and Cybernetics, IEEE Transactions on*, 20:931–938, 1990.

[111] B. Oommen and D. Ma. Fast Object Partitioning Using Stochastic Learning Automata. In *SIGIR '87: Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–122, New York, NY, USA, 1987. ACM Press.

[112] B. Oommen and D. Ma. Deterministic Learning Automata Solutions to the Equipartitioning Problem. *IEEE Transaction Computer*, 37(1):2–13, 1988.

[113] B. Oommen and T. Roberts. Continuous Learning Automata Solutions to the Capacity Assignment Problem. *IEEE Transactions on Computers*, 49(6):608–620, 2000.

[114] M. Palley and J. Simonoff. The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases. *ACM Trans. Database Syst.*, 12(4):593–608, 1987.

[115] J. Panaretos and N. Tzyvidis. *Aspects of Estimation Procedures at Eurostat with Some Emphasis on Over-Space Harmonization.* In HERCMA 2001 CONFERENCE, 2001.

[116] J. Park, S. Chen, and P. Yu. Using A Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Transactions on Knowledge and Data Engineering*, 9:813–826, 1997.

[117] B. Reddy. Aspects of Database Security and Program Security. Technical report, Kanwal Rekhi School of Information Technology, Indian Institute of Technology-Bombay, 2003.

[118] S. Sahni. *Data Structure, Algorithms, and Applications in C++.* McGraw-Hill, 1998.

[119] J. Sanchez, J. Urrutia, and E. Ripoll. Test Report on Multivariate Microaggregation in m-Argus 3.2. Technical report, CASC PROJECT, Computational Aspects of Statistical Confidentiality, 2003.

[120] J. Sanchez, J. Urrutia, and E. Ripoll. Trade-Off between Disclosure Risk and Information Loss Using Multivariate Microaggregation: A Case Study on Business Data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004 Proceedings*, pages 307–322, Spain: Barcelon, 2004. Berlin: Springer-Verlag.

[121] G. Sande. Methods for Data Directed Micro-aggregation in One Dimension. In *Proceedings of NTTS and ETK: New Techniques and Technologies for Statistics and Exchange for Technology and Know-how*, Crete: 18-22 June.

[122] G. Sande. Exact and Approximate Methods for Data Directed Microaggregation in One or More Dimensions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):459–476, 2002.

[123] J. Schatz. Survey of Techniques for Securing Statistical Databases. Technical report, University of California at Davis, 1997.

[124] F. Sebé, J. Domingo-Ferrer, J. Mateo-Sanz, and V. Torra. Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 163–171, London, UK, 2002. Springer-Verlag.

[125] S. Skiena. *The Algorithm Design Manual.* Springer-Verlag, 1998.

[126] A. Solanas and A. Martíínez-Ballestéé. V-MDAV: A Multivariate Microaggregation With Variable Group Size. In *17th COMPSTAT Symposium of the IASC*, Rome, 2006.

[127] A. Solanas, A. Martíínez-Ballestéé, J. Domingo-Ferrer, and J. Mateo-Sanz. A 2d-Tree-Based Blocking Method for Microaggregating Very Large Data Sets. In *The First International Conference on Availability, Reliability and Security. The International Dependability Conference Bridging Theory and Practice*, 2006.

[128] A. Solanas, A. Martinez-Balleste, J. Mateo-Sanz, and J. Domingo-Ferrer. Multivariate Microaggregation Based Genetic Algorithms. In *3rd International IEEE Conference on Intelligent Systems*, pages 65–70, 2006.

[129] G. Sullivan and W. Fuller. The Use of Measurement Error To Avoid Disclosure. In *The Section on Survey Research Methods*, pages 802–807, 1989.

[130] D. Tasoulis and M. Vrahatis. Unsupervised Clustering on Dynamic Databases. *Pattern Recognition Letters*, 26:2116–2127, 2005.

[131] V. Torra. Microaggregation for Categorical Variables: A Median Based Approach. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004 Proceedings*, pages 162–174, Spain: Barcelon, 2004. Berlin: Springer-Verlag.

[132] V. Torra, J. Abowd, and J. Domingo-Ferrer. Using mahalanobis distance-based record linkage for disclosure risk assessment. In *Privacy in Statistical Databases-PSD 2006*, pages 233–242, 2006.

[133] V. Torra and J. Domingo-Ferrer. Towards Fuzzy C-Means Based Microaggregation. In P. Grzegorzewski, O. Hryniewicz, and M. Gil, editors, *Advances in Soft Computing: Soft Methods in Probability, Statistics and Data Analysis*, pages 289–294, Germany: Heidelberg, 2002. Physica-Verlag.

[134] V. Torra and S. Miyamoto. Evaluating Fuzzy Clustering Algorithms for Microdata Protection. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004 Proceedings*, pages 175–186, Spain: Barcelona, 2004. Berlin: Springer-Verlag.

[135] R. Valiveti and B. Oommen. On Using the Chi-Squared Metric for Determining Stochastic Dependence. *Pattern Recognition*, 25(11):1389–1400, 1992.

[136] R. Valiveti and B. Oommen. Determining Stochastic Dependence for Normally Distributed Vectors Using the Chi-squared Metric. *Pattern Recognition*, 26(6):975–987, 1993.

[137] A. Valls, V. Torra, and J. Domingo-Ferrer. Aggregation Methods to Evaluate Multiple Protected Versions of the Same Confidential Data Set. In P. Grzegorzewski, O. Hryniewicz, and M. Gil, editors, *Soft Methods in Probability, Statistics and Data Analysis (Warsaw, 2002)*, pages 355–362, Germany: Heidelberg, 2002. Physica-Verlag.

[138] M. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides. The New k-Windows Algorithm for Improving the k-Means Clustering Algorithm. *Journal of Complexity*, 18:375–391, 2002.

[139] J. Ward. Hierarchical Grouping to Optimize an Objective Function. *J. American Statistical Association*, 58:236–245, 1963.

[140] T. Wende. Different Grades of Statistical Disclosure Control Correlated with German Statistics Law. pages 336–342, Spain: Barcelona, 2004. Berlin: Springer-Verlag.

[141] L. Willenborg and T. Waal. Statistical Disclosure Control in Practice. *Lecture Notes in Statistics 111*, 1996.

[142] L. Willenborg and T. Waal. *Elements of Statistical Disclosure Control.* New York: Springer-Verlag, 2001. ILL Number: 2132712.

[143] R. Wilson and P. Rosen. Protecting Data through Perturbation Techniques: The Impact on Knowledge Discovery in Databases. *Journal of Database Management*, 14(2):14–26, 2003.

[144] W. Winkler. Re-identification Methods for Masked Microdata. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004*, pages 216–223, Spain: Barcelona, 2004. Berlin: Springer-Verlag.

[145] P. Wolf, J. Gouweleeuw, P. Kooiman, and L. Willenborg. Reflections on PRAM. Technical report, 1997.

[146] W. Yancey, W. Winkler, and R. Creecy. Disclosure Risk Assessment in Perturbative Microdata Protection. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 135–152, London, UK, 2002. Springer-Verlag.

[147] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

[148] Y. Yao, L. Chen, and Y. Chen. Associative Clustering for Clusters of Arbitrary Distribution Shapes. *Neural Processing Letters*, 14:169–177, 2001.

[149] Y. Yao, L. Chen, A. Goh, and A.Wong. Clustering Gene Data via Associative Clustering Neural Network. In *the 9th International Conference on Neural Information Processing (ICONIP'02)*, volume 5, pages 2228–2232, 2002.

[150] A. Zaslavsky and N. Horton. Balancing Disclosure Risk Against the Loss of Nonpublication. *Journal of Official Statistics*, pages 411–419, 1998.