CHAPTER 1

INTRODUCTION

1.1 THE AIM AND OBJECTIVES

The impulsive escalation of new data mining techniques has augmented privacy risks leaps and bounds. This is indubitably probable, because now it is plausible to strongly coalesce and interrogate gigantic data stores accessible on the web, in the midst of fumble of prior mysterious out of sight patterns. Privacy issues also augment its convolution because of new upcoming technologies; which are linking enormous number of reciprocally strange and erratic people, to make a worldwide economy. This scenario is of serious apprehension challenging consideration. The excerpt of Earl Warren (1974), “The fantastic advances in the field of electronic communication constitute a greater danger to the PRIVACY of the individual,” further enlightens the claim for momentousness of privacy.

The principal stipulation wished-for privacy is a high-quality data fortification modus operandi known as data mystification techniques. The aim of this thesis is to design, develop and deploy,
'Mystification of Data for Optimized and Efficacious Privacy Preserving Data Mining.'

The thesis discusses the privacy issues that are likely probable to control the utility of data mining projects. At its nucleus, the significance of Privacy Preserving Data Mining (PPDM) is plagiaristic not only from its flamboyance to haul out imperative knowledge, but also from its resiliency to molestation. PPDM is a discipline whose desire is to authorize delivery transmits of respondent data while preserving respondent privacy. It introduces solutions to problems where the question is how to get hold of data mining results without violating privacy.

In the proposition of novel solutions there are two vital things to be highlighted. The first one is Privacy of users and personal data within strenuous environments and implicit communities. The second one is Information Security as it relates to privacy and the information resources provided in the same environments. This thesis aspires to contribute to the solution of a specific problem, namely, the problem of sharing sensitive data. Developing new, improvising existing algorithms and techniques for PPDM is ventured at.

1.2 MOTIVATION

In the midst of the implausible detonation of data garnering, data dissemination, internet technologies and the manifestation of susceptible applications, security issues in knowledge discovery have
reached the pinnacle. The chic data analysis and mining techniques have produced a sharp awareness of their potential, compromising privacy, and has posed challenges of controlling data access in accordance with privacy policies. The escalating exploitation done by them have reached the zenith, and have surfaced as a central and ubiquitous problem. Knowledge is pre-eminence, but as humble users of the most modern technologies we are pitted with possessions that may even make us paranoid concerning usage of a computer. This thesis will bring to the surface, the modus operandi needed to shield privacy.

The task of PPDM force’s a central thrust on establishing a world with robust data security, where knowledge users persist to profit from data without compromising the data privacy.

The objective of privacy-preserving data mining is to liberate a dataset that researchers can study without being able to identify sensitive information about any individuals in the data (with high probability). The PPDM should perform well at needed levels during times of both crisis and normal operations.

1.3 PRIVACY PRESERVING DATA MINING

PPDM is accomplishing valid data mining results devoid of the wisdom of the core data values. In recent years, PPDM as an area of study and development of accurate models without access to the original data has been receiving attentiveness in the research society and beyond.
The technological tempo in PPDM has enabled us to cross the gateway from where one can stride out into the twenty first century with confidence and poise. The perpetual necessity is escalating to protect privileged information and it has to be used for research purposes. There has been recent proliferation in PPDM techniques, which is apparent by its global emergence. Motivated by the increasing number of successful privacy-preserving techniques, in this thesis the attempt is to discover the best amongst them.

1.3.1 Key Directions in PPDM Study

In this research work of, ‘Mystification of Data,’ liberal concern is taken to excogitate, optimize and aggrandize PPDM along diverse vista. Efforts to improvise the techniques at least one for each panorama is endeavored.

(a) Privacy-Preserving Data Publishing and releasing techniques tend to swot up how diverse alteration methods can be used in juxtaposition with classical data mining methods such as clustering. At this juncture the striking problem of, ‘the theoretical challenges in high dimensionality’ needs mention. Real data sets are by and large immensely high dimensional which makes the process of privacy preservation tremendously complicated both from a computational and effectiveness point of view. Despite the fact that an optimal solution is NP-hard, the technique is not even effectual with escalating dimensionality, since the data can typically be
pooled with either public or background information to reveal the identity of the underlying record owners.

(b) **Altering the results of Data Mining (DM) applications to preserve privacy** is a field of privacy in which the results of DM algorithms such as Association Rule Mining (ARM) are altered in order to preserve the privacy of the data.

(c) In **Query Auditing** the results of queries are either modified or restricted.

(d) **Cryptographic Methods for Distributed Privacy** uses a variety of cryptographic modus operandi in order to communicate among the diverse sites, so that secure function computation is possible without revealing sensitive information.

1.3.2 **Taxonomy of PPDM Techniques**

The three types of PPDM modus operandi are as shown in Figure 1.1. In the first type privacy is preserved by Statistical Disclosure Limitation (SDL) of data; in the second type (Cryptography-based techniques like secure multiparty computation) of privacy, data is manipulated so that the mining result is not affected or minimally affected; and in the third type (Data Befuddlement) the data is obscured
(i.e. perturbation, blocking, aggregation or merging, swapping, sampling etc.), so that the mining result will preserve certain privacy.

Figure 1.1 Taxonomy of PPDM Techniques

In this thesis, work is done to enhance and optimize all the three types of PPDM Techniques individually and in amalgamation.

1.3.3 Privacy Preservation: Problem Definition

In this hypothesis, the predicament embarked upon is mystifying a database into a fresh one that makes sensitive information obscure, while preserving the general patterns and trends from the original database. The data is spread across personal data to customers procuring behaviour, financial, medical, marketing, and insurance liability information and sensitive patterns.
The mystification applied to the database transpires prior or posterior to the partaking of data for mining as can be seen in Figure 1.2.

![Figure 1.2 The Framework for PPDM Systems: High-Level](image)

The first and foremost thing spotlighted on PPDM, markedly in the milieu of the mining tasks are: a) association rules which portray interesting relationships amid items grouped together; b) clustering which is concerned with grouping objects into classes of similar objects; and c) classification which is concerned with categorising data.

1.3.4 The Equipoise

PPDM swots up modus operandi for meeting the potentially contradictory goals of respecting entity rights and allowing lawful organizations to assemble and excavate colossal data sets.
1.4 THESIS STATEMENT

In this work, the practicability of achieving PPDM by data mystification is scrutinized. The core of the thesis declaration of this research is offered as follows:

Privacy preservation in data mining, by data mystification, is to some degree effectively equipoisable. A noteworthy quantity of information as of the input data set may possibly be mystified with the intention that the privacy echelon of the data mining course of action may be augmented.

This research demonstrates empirically and hypothetically the expediency and practicability of achieving PPDM. Specially, it is publicized that symmetry between privacy preservation and knowledge discovery can be accomplished while addressing information fortification in association rule mining, privacy preservation in classification and privacy preservation in clustering.

A new modus operandi is suggested in this thesis to consent to the PPDM. In view of the fact that it exploits the (enhanced, optimized and efficacious) existing techniques and novel techniques it is pertinent to baptize the modus operandi as data mystification (Indumathi et al 2008 o). In this research the name mystification is utilized, because
mystification is the act or an instance of making obscure or mysterious (i.e. the emphasis is more on the increase in privacy without decreasing data utility). In the proposed system, each user (the data provider) first mystifies the personal data, and then forwards to a middle place (the data collector), such that the data collector cannot originate the candid information about a user's private information. However, the data camouflaging system should still be able to allow the data collector to conduct collaborative knowledge mining from the disguised data. The basic idea of mystification is that the middle place can only know the scope of the data, and such scope is wide enough to safeguard users' privacy. Even though information from each individual user is muddled, if the number of users is significantly large, the aggregate information of these users can be ball parked with decent precision. For those computations, it can still engender meaningful results without knowing the exact values of individual data items since the required aggregate information can be estimated from the jumbled data.

The mystification techniques with collaborative mystification algorithms, when combined, can accomplish a decent scale of precision for the privacy-preserving collaborative mystification is theorized (Indumathi et al 2007 b). To verify this hypothesis, the mystification technique for a collaborative mystification algorithm is implemented. Sequences of experiments illustrate how precise the research results are. The predictions premeditated based on original data are then compared with the predictions on mystification data.
1.5 RESEARCH METHODOLOGY

This research was conducted in a phased approach, as follows:

Assessment of preceding efforts: In the primary phase the active privacy preserving techniques (PPT) in the literature are scrutinized, studied and investigated. The existing solutions for PPDM are compendiously thrashed out in Chapter 2.

Formation and design of innovative PPT: In the secondary phase an efficacious PPDM is targeted by aggrandizing the existing PPT and thought-out innovative Mystification Techniques (MYST). PPT’s discussed here covers the main aspects of Clustering, Classification and ARM.

Implementation and test: In the third phase innovative Mystification (MYS) methods are implemented and tested on real datasets to examine and study their performance (with the intent of improvising their efficacy).

Evaluation: In the fourth phase the MYS methods are evaluated and compared. An assessment framework is specified and implemented as a common platform to establish the suitable technique for every kind of application. The impact and result of the data mining task with and without the alteration of the MYS methods were evaluated. The
innovatively propounded MYS methods are compared with the published results found in the corresponding literature. The efficiency of the proposed algorithms is evaluated. Metrics from literatures are used to quantify the privacy level provided by the MYS methods, to measure information loss and the quality of the data mining results in the transformed datasets.

**Dissemination:** The research results were disseminated through the deference of papers to peer reviewed conferences and journals [1 to 37]. The PPMYS methods were fine tuned with the help of the criticism received through the peers.

1.6 **ORGANISATION OF THE THESIS**

The remnants of this thesis are organized as follows:

Chapter 2 outlines the survey of literature of the topics associated with this work. The existing works are analyzed for their advantages and limitations.

Chapter 3 describes the overall architecture of the framework proposed in this thesis that deals with Mystification of Data for Optimized and Efficacious Privacy Preserving Data Mining.

Chapter 4 expounds the mystification techniques which decrease the granularity of representation or limit access to resources in
order to increase the privacy. The majority techniques for privacy computations use some form of mystification on the data in order to perform the privacy preservation.

Chapter 5 explicates the development of PPDM Ontologies, a PPDM e-learning system and an Ontology-based multi-agent model. It also explains the pioneered ontology-based multi-agent model framework system premeditated to aid tyro users, researchers, and experts in retrieving relevant facts regarding information privacy; facilitates reclaim of PPDM domain knowledge by integrating various existing ontologies that portray portions of a colossal realm or reuse a general ontology and expand it.

Chapter 6 portrays the enhanced Modus Operandi for Selection of the best Privacy Preserving Mystification technique for any domain and depicts the intelligent mystification selector.

Chapter 7 construes the modifications made in the PPDM Standards for efficacious development and deployment of methodological solutions and will let vendors and developers to construct unyielding advances in the upcoming of PPDM. The Privacy Preserving – Systems Development Life Cycle’s mainstay is to integrate the most capable privacy investigation results and tools, including the proposed sole contributions and ideas, into a systems design methodology.
Chapter 8 analyses and discusses the results obtained from this research work in which the performance comparison of the various PPT are carried out and presented. Moreover, we analyze the performance of optimized Mystification techniques against the un-optimized Mystification techniques are also analyzed.

Finally, the conclusions on this work, some possible provocations left to reconnoiter and Future research penchants are discussed in Chapter 9.