Volume 6 • Issue 5 • October 2012

Editor-in-Chief
Dr. Chen-Chi Shing

INTERNATIONAL JOURNAL OF

COMPUTER SCIENCE AND SECURITY (IJCSS)

ISSN : 1985-1553

Publication Frequency: 6 Issues / Year

CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

VOLUME 6, ISSUE 5, 2012

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-1553

International Journal of Computer Science and Security is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCSS Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

Book: Volume 6, Issue 5, October 2012
Publishing Date: 25 - 10- 2012
ISSN (Online): 1985 -1553

IJCSS Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2012

EDITORIAL PREFACE

This is fifth issue of volume six of the International Journal of Computer Science and Security (IJCSS). IJCSS is an International refereed journal for publication of current research in computer science and computer security technologies. IJCSS publishes research papers dealing primarily with the technological aspects of computer science in general and computer security in particular. Publications of IJCSS are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJCSS are databases, electronic commerce, multimedia, bioinformatics, signal processing, image processing, access control, computer security, cryptography, communications and data security, etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 7, 2013, IJCSS will be appearing with more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJCSS is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJCSS as one of the top International journal in computer science and security, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Computer science and security fields.

IJCSS editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCSS provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computer Science and Security (IJCSS)

EDITORIAL BOARD

Assistant Professor Vishal Bharti
Maharishi Dayanand University
India

Dr. Parvinder Singh
University of Sc. & Tech
India

Assistant Professor Vishal Bharti
Maharishi Dayanand University,
India

TABLE OF CONTENTS

Volume 6, Issue 5, October 2012

Pages

Data Mining And Visualization of Large Databases

AbdulRahman R. Alazmi raphthorne@yahoo.com
College of Petroleum and Engineering
Kuwait University
Kuwait

AbdulAziz R. Alazmi fortinbras222@hotmail.com
College of Petroleum and Engineering
Kuwait University
Kuwait

Abstract

Data Mining and Visualization are tools that are used in databases to further analyse and understand the stored data. Data mining and visualization are knowledge discovery tools used to find hidden patterns and to visualize the data distribution. In the paper, we shall illustrate how data mining and visualization are used in large databases to find patterns and traits hidden within. In large databases where data is both large and seemingly random, mining and visualization help to find the trends found in such large sets. We shall look at the developments of data mining and visualization and what kind of application fields usage of such tools. Finally, we shall touch upon the future developments and newer trends in data mining and visualization being experimented for future use.

Keywords: *Applications of Data Mining, Business Intelligence, Data Mining, Data Visualization, Database Systems.*

1. INTRODUCTION

Since the inception of information storage, the ability to sift through and analyze huge amounts of information was a dream sought out for in many ways and through different ways. With the advent of electronic and magnetic data storage, rational databases emerged as one of the efficient and widely used method to store data. Data stored in such large databases are not always comprehendible by humans, it needed to be filtered and analyzed first. Stored records are raw amounts of data poor in information, not only is it large and seamlessly irrelevant but also continuously increasing, updating and changing [1].Here is where data mining and visualization comes into the picture. Data mining and visualizations are knowledge discovery tools [2] used for autonomous analysis of data stored in large sets in many different ways. Large data sets of data cannot possibly be analyzed manually; mining tools and visualization provide automated means to comprehend such data sets. Data mining is defined as the automated process of finding patterns, relationships, and trends in the data set. On the other hand, data visualization is the process of visually representing the data set in a meaningful and comprehendible manner [3]. In fig. 1, the figure shows what Data Mining is and is not.

Data mining is a knowledge discovery process; it is the analysis step of knowledge discovery in databases or KDD for short. As an interdisciplinary field of computer science, it involves techniques from fields such as artificial intelligence AI, machine learning, probability and statistics theory, and business intelligence. As in actual mining, where useful substance is mined out of large deposits hidden deep with mine. Data mining mines meaningful and hidden patterns, and it's highly related to mathematical statistics. Though utilizing pattern recognition techniques, AI techniques, and even socio-economic aspects are taken into consideration. Data mining is used in today's ever-growing databases to achieve business superiority, finding genome sequences, automated decision making, monitoring and diagnosing engineering processes, and for drug

discovery and diagnosis in medical and health care [4]. Data Mining, as with other Business Intelligence tools, efficiency is affected by the Data Warehousing solution used [5] [6].

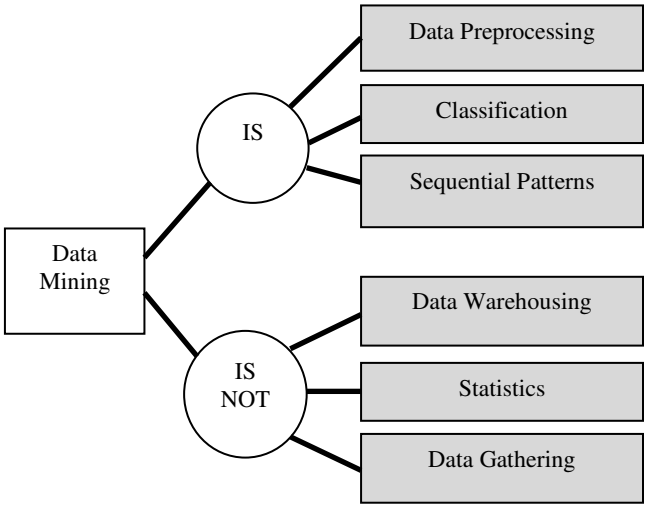

FIGURE 1: What is Data Mining?

Data Visualization is a data mining application considered as an information-modeling paradigm, in which seamlessly random data may be represented in an appealing graphical manner. Visualization of collected data can be found as early as the middle of the 19th century, where Dr. John Snow made a map of central London, pinpointing the locations of the possible sources of the cholera and its victims. Thus allowed for the detection of the hidden relation of the alleged sources of cholera (the water pumps) and its victims, and helped in ridding of the disease [7], other examples are also given in [8]. Visualization can be divided in to seven main subfield according to Frits [9], visualization algorithms, volume visualization, information visualization, multi-resolution methods, modeling techniques, and architecture and interaction techniques. Human beings understand and comprehend graphics more easily than numbers and letters. Human brains can interpret graphs, charts, icons and models quicker than numbers in tables; this is in contrast to computers, where numerical representation is perceived more efficiently. For example, a pie chart showing the classification of a university student will be understood quicker than the same data represented in a table, as Fig. 2 shows. Visualization of such data helps the human brain figure out and perceives such knowledge hidden in the data. The goal of data visualization is to not only summarize the large dataset, but also provide a better way of exploring the knowledge hidden and waiting to be found there automatically and autonomously. Visualization of datasets helps in explicitly showing proximity, enclosure, similarity, connection, and continuity. Analysis of data through visualization is further divided into two main categories, the Exploratory Data Analysis EDA, and Qualitative Data Analysis QDA; we shall see both analysis types in the paper.

Data Visualization and Data Mining can be used together, or in sequence, whether Data Visualization first or Data Mining first [10] [11]. It's worth noting that data mining and visualization today are available on most platforms. Cost for data mining applications range from high-end DBMS, and huge processing power costing in millions to business class systems costing in several hundreds of dollars. In addition, it is worth mentioning that data mining and visualization efficiency depends on the type of DBMS, and the processing power available for the application.

In this paper, we shall review data mining and visualization in light of their usage in large databases. We shall see the current trends, main tools and application of such technologies and have a look at the latest and possible future uses for data mining and visualization. In the next section we shall demonstrate and give a background on the developments that led to the modern data mining and visualization technologies we came to know today. In section III, we will talk

about the tasks and techniques used today and available in the market for data mining and visualization. In sections IV and V, we shall see the applications were data mining and data visualization is used. In section VI, we examine some of the tools used; in section VII we shall see latest developments and future uses of such technologies. In section VIII, we see some of the challenges. Finally, we conclude the paper in the conclusion and references.

2. BACKGROUND

The notion of automated discovery tool in a large data set has prevailed in the development of data storage technologies. Tracing the roots of data mining to the early days of mathematical regression and probability theories in the eighteenth century, we can see that mathematical models such as regression and Bayesian theories provided means of analyzing large data sets effectively. With electronic computers taking the exclusive position for data storage in the twentieth century, early commercial computers quickly over took manual and other means of data storage. By the 1950s, early high level languages were developed; this development dramatically changed how humans interact with computers. Computerized data storage was not only used for storage but also for querying. Further on after the advancements in both hardware and software, rational database systems RDBS were developed. Structured Query Languages SQLs were used for semi-automatic acquisition of knowledge through querying the data storage, although tedious programming and substantial efforts have to be done.

By the late 1970s and through the 1980s, developments in computer networking, data storage and software had led to the break-through developments of the famous *online-analytical processing* OLAP techniques. Further developments in databases such as multidimensional and spatial database, and the dramatic cost reduction in data storage have lead the way for complex databases with sizes ranging in terabytes to petabytes ever growing and 24 hours online connected. Such developments led to the development of more complex algorithms derived from both AI and neural networks to efficiently search the database to automatically and autonomously acquire knowledge, more efficient and complex than OLAP. In the early 1990s, we can safely say that early data mining and visualization tools were developed.

As modern RDBMS dominate the market, data mining tools are developed to search such solutions. RDBMS usually store the data in the form of bytes or *Binary Data Objects* blobs; this makes the data mining of such datasets even more elusive and harder. Data mining was known by many names including knowledge extraction, information discovery, data archaeology, and data pattern processing. The term data mining however is the most popular term in the database field, since then it was also incorporated both the AI and machine-learning fields as well [12]. With further developments in data mining tools happening today, and the huge increase in processing power and decrease of hardware and network costs, accessible and efficient data mining can be achieved at a moderate costs. The business sector, scientific research, and health care are the dominant users of such data mining and visualization tools. Standards such as the European Cross Industry Standard Process for Data Mining CRISP-DM were developed to create a cross platform compatible data mining interface to cope with the increasing demand for data mining across many different applications and fields. Today, data mining software packages imply complex algorithms and techniques for searching, pattern recognition, and forecasting complex global stock exchange markets changes. Oracle, IBM and Microsoft are the most prominent providers of commercial data mining software. Such advanced and available intelligence software have influenced and played a major role in reshaping the security practices and techniques applied by international intelligence agencies such as FBI and CIA.

Data visualization is an emerging field, developed to counter the ever-increasing growth of databases in both size and complexity. Developed from the statistical, probabilistic and data representation fields to make sense of large quantitative data sets found in databases. As with data mining, data visualization techniques began as mathematical tools that summarize large datasets into a single representation or values. Mathematical models included time-series graphs, cartography, and fitting equation [13]. Even before computers were available, visualization

operation on data was done [14], such as Francis Galton's weather maps fating back to 1980's [15]. Today complex techniques are used in visualization. Visualizations are mainly used for business and scientific research applications. Usually data visualizations, unlike data mining, work on raw data such as numbers or letters as in names [16]; this makes the visualization process consume both time and energy. Such a problem is frequently faced with large DBMS.

According to D. Keim in [11], Visualization techniques can achieve several ends that include visual bases for data hypotheses, evidence for or against a trend in the data, and/or data models for demonstration purposes. Data visualization is used to visualize and present the data set, test hypothesis, or explore dataset freely. Visualization of data also helps in communication and easily emphasizes trends otherwise buried within the dataset. As stated by Friedman [17] "*main goal of data visualization is to communicate information clearly and effectively through graphical means*".

Data Visualization techniques can be used to *pre-process* data before Data mining techniques is used. These include segmenting, sub setting, and aggregation techniques. Visualization techniques of Data can be categorized into three categories, which are *Data Visualization, Distortion, and Interactive Techniques*. The first type include among others Geometric, Graph-Based, and Icon-Based. This type is seen in the form of Histograms, Scatter plots, and Shape Coding. Distortion types include the use of Perspective Wall, and Hyper-box, the latter being techniques used for multivariate datasets [18]. The latter type, the Interactive techniques can be in the form of Projections, Zooming, and Detail on Demand. There still exist among researchers of Data Mining and machine learning fields the need to incorporate and embed Data Visualization tools into Data Mining tools.

3. RELATED WORK

In [19] the author reviews several interactive visualization techniques that are used in the context of data mining. The paper also retrospectively defines visualization techniques in the world of data mining; these can be defined as expressing data sets to discover trends, for *exploration*, or can visualize the workings of complex data mining processes, for *comprehension*. The paper focuses on data visualization, while in our survey we shall review both data mining and data visualization and their integration as one field.

Authors C. Romero and S. Ventura of [20] give a survey data mining techniques in the field of education. Not just in e-learning but also in traditional class rooms. Data mining can help in improving educational courses through knowledge discovery of facts in the past history of a specific course. These include: feedback for the educators such as effectiveness of content, students' classifications, and mistakes in the teaching process, feedback for the students such as suggesting helpful educational content available for them. The paper surveys data mining and a few data visualization techniques used in education such as classification, text mining, sequential patterns and visualization. In our survey, data mining and visualization techniques, trends, and application will be discussed not only for education but for a wider range of fields.

In [21], the paper reviews the history of Knowledge Discovery and Data Mining KDDM process, its definitions, models, and standards. The survey suggests a need for the standardization of the KDDM methodology rather than its somewhat haphazard usage in industry. While effective models are used, they are however separate in form and methodology. This can affect the field of KDDM as it matures by making ambiguous and redundant set of models and techniques. Data mining being a step in the KDDM process helps in understanding processes and gives input for decision support systems. Both KDD and KDDM are related, while the latter is not only concerned with databases, but other sources of data. KDDM models range from industrial to academic, each having several different steps. Important steps are data extraction, preparation, mining, and evaluation. The survey compares several KDDM models, while in our survey we do not take the whole KDD or KDDM process in the picture; we focus on data mining and visualization alone. Data mining is mostly a step in the middle of any KDD or KDDM model, and it is a pivotal one with many dimensions and factors.

Other body of work usually surveys a specific field in the data mining and data visualization techniques, such as [22] for web mining, [23] for data visualization in bioinformatics, and [24] for data mining in e-commerce. In this survey, we take a broader view in many fields and trends.

4. TASKS AND TECHNIQUES

Data mining and data visualization were developed from mathematical methods of pattern recognition and probabilistic theories to deal with unstructured, time varying, and fuzzy data in huge amounts. Such techniques allowed for finding correlations, relations and assertions. We shall touch upon some of the main tasks associated with data mining and visualization and the techniques to achieve such tasks that are popular in the field of data mining and visualization in the following paragraphs.

Data mining tasks include *Classification, Association Rules, Clustering, Anomaly detection, Summarization, Regression, and Sequential Patterns,* M. Sousa et al [3]. Visualization tasks include *statistical modeling, regression modeling, information abstraction*, *mindmaps*, and usually *data presentation* in other forms like graphs, maps, and histograms. All data mining and visualization techniques and algorithms relay on three main steps, model representation, then evaluation of the model, finally model search to identify patterns [25]. Model representation depends on the dataset itself, most datasets require certain models, clustering usually is effective with demographical datasets. Second is model evaluation, where the model used is evaluated to make sure it matches the nature of the dataset. Finally the model search, it's done after evaluation of the model is verified, it extracts the knowledge we need from the dataset.

Classification is the process classifying sets of data based on common attributes. Classification is considered a classical data mining techniques as it's highly related to statistics method used before data mining was conceived. This classification help divide the large dataset into further smaller and correlated datasets. Such classification is the basis for further analysis, as classifications divide the datasets into smaller correlated groups; this is called consolidation of data. Then we have *association rules*, it's the process of testing or when implying a set of hypotheses are made against a certain data set. These hypotheses are called *rules.* After the verification of the plausibility of such rules, or associations, then we subject the dataset against such association rules. Associations rules can find hidden links between otherwise unrelated data; the *beer-diaper* links used in *market baskets* is an example for such associated rules. Market baskets are defined as items usually bought together; such an analysis is used heavily in *marker research*.

Clustering is the technique of grouping of several objects unto groups of similar attributes in order to simplify large, complex sets. Clustering is a learning technique and therefore it has no correct answer. Clustering can be hierarchical and non-hierarchical. Hierarchical clustering clusters groups of data in size (can be from small to large or vice versa), and it comes in two flavors, Agglomerative and Divisive. The first clusters each record alone, and then merges clusters together. The second, does the opposite, it starts with one full cluster and then subdivides the cluster. The non-hierarchical clustering has two flavors as well; the difference here is that no hierarchic clusters are used. The first type is the single pass methods, where the database is scanned once to create the cluster. The second type is the relocation method, where records are relocated from one cluster to another for optimization. Several passes against the database may be used, as opposed to the single pass methods [26].

Sequential Patterns the use of sequential pattern algorithms on sets of sequential data (e.g. bills made on the same month). The goal is to find a trend or pattern that happen in sequence. Rule induction task is used to find hidden if-then rules in the dataset. These rules are based on statics analysis and probabilistic models. Derived if-then rules are further used in analyzing the dataset in the future.

Data mining techniques are varied and interdisciplinary, since they come from varied fields. Neural Networks are techniques frequently used in data mining. These techniques are from the field of Artificial Intelligence AI. Neural Networks link different attributes through vectors intelligently; it has considerable training time when compared to other techniques, and has little confidence intervals that depend on the number of neighbors. Also AI derived techniques tend to be more sophisticated and show human like-intelligence in finding hidden correlations.

Nearest neighbor technique is another classical technique to classify records of data based on their resemblance or closeness to a specific record. This technique is used to compare newer or updated records to a pivotal or historical important record; it tries to mimic the human comparison process. Decision trees are techniques from *machine learning* field. When compared to neural networks, *decision trees* are much faster in performance, due to less computational overhead. Decision trees algorithms are greedy algorithms that divide the cases or classified groups in the training set of data until no more cases in the dataset can be logically or ontologically divided. Their drawback is their need of large datasets to provide efficient results. Different kinds of decision trees exist, we mention two kinds for example. Classification and Regression Trees CARTs, these trees split the data set into 2 way splits for decision making. Other type of decision trees is Chi Square Automatic Interaction Detection CHAID. CHAID trees on the other hand create splits in the dataset using the Chi square tests, creating multi-way splits in the dataset.

Moving on to data visualization, techniques for visualization vary depending on the type, usually they are classified as query independent techniques, and query dependent techniques. Query independent techniques directly visualize data set without any assertions. On the other hand query dependent techniques will visualize depending on a query specified prior. We shall look at techniques of both classes.

In D. Keim's work in [27] the authors presented a novel technique called pixel-oriented visualization techniques. Pixel oriented techniques are mappings of the data values into a 2D or 3D map of colored pixels depending on the value of the data. The colored maps give immediate and precise information on the trend or the average values of the dataset [28]. Pixel oriented techniques are further divided into query dependent and independent pixel techniques. The query dependent pixel oriented techniques tend to form a map of the current trend of the data. Usually this is not very useful as most of the time the data values or the colored map is not very easy to read out. On the other hand, the query dependent pixel oriented techniques are more effective, as the finished colored graphs indicates how the data is scattered or varied around the queried data set or target values.

Other visualization techniques are the geometric projection techniques. These techniques are sometimes summed under the projection techniques. These techniques find efferent or convenient 'projections' of the data in multiple dimensions. In addition, these techniques are used chiefly in EDA, as most of the time multiple projections are done for further exploring the dataset. Since the search space is very large in terms of multi-dimensional datasets, exploration can prove to be very difficult. Systems developed specifically for geometric projection pursuits found in [29], automatically find such convenient and interesting projections more easily and effectively.

5. VISUALIZATION OF LARGE DATABASES

As discussed previously in this paper, data visualization is an important application that helps to convey knowledge mined *graphically*. As human beings, we are more familiar with drawings, icons, and graphs then we are with numbers and tables. Raw numerical data or even alphanumerical data can be represented in a map; chart, bars, pie chart, or even a histogram to visually identified and convey important trends and correlations visually. Data mining in large databases is still a difficult task; due to the fact of the huge amount of raw data need to be processed. A turnaround is to partition the data into sets, and tackle them individually. This makes supporting tools such as *Visualization Tools* needed. Data visualization is considered with

two kind of analysis, first, Exploratory Data analysis EDA and model visualization [30], second is the Qualitative Data Analysis QDA.

By EDA, it is meant the careful exploration of the data set graphically to identify a pattern, a recurring trend or behavior that connects different views or visualization. EDA helps to identify patterns without preconceived knowledge, hypothesis, or suggested models used on the data set. Model visualization is the use of predefined models, such as XY charts, 3D plots, or box plots to model the data. Usually visualization of data plays on the key idea that human beings are more capable in analyzing and understanding graphs than digits and letters. Figure 2 include a very simple, yet effective example of tabular versus visualized data sets. Visualizations such as Venn diagrams and clustering help the observers see grouping and partitions in a dataset more easily than rows of alphanumerical records. QDA on the other hand, is the analysis of non-numerical data. QDA is considered with database containing images, text, links, or other kinds of data that is not numerical or alphanumerical.

Table 1. University A Students Classified

Academic Year 20xx/20xx	Males	Females
Freshmen	220	120
Sophomores	245	312
Juniors	389	279
Seniors	295	320
Graduate	112	98

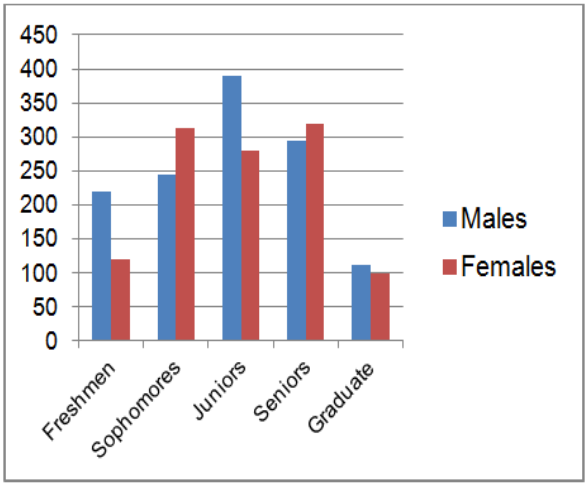

FIGURE 2 (a): Histogram Representation

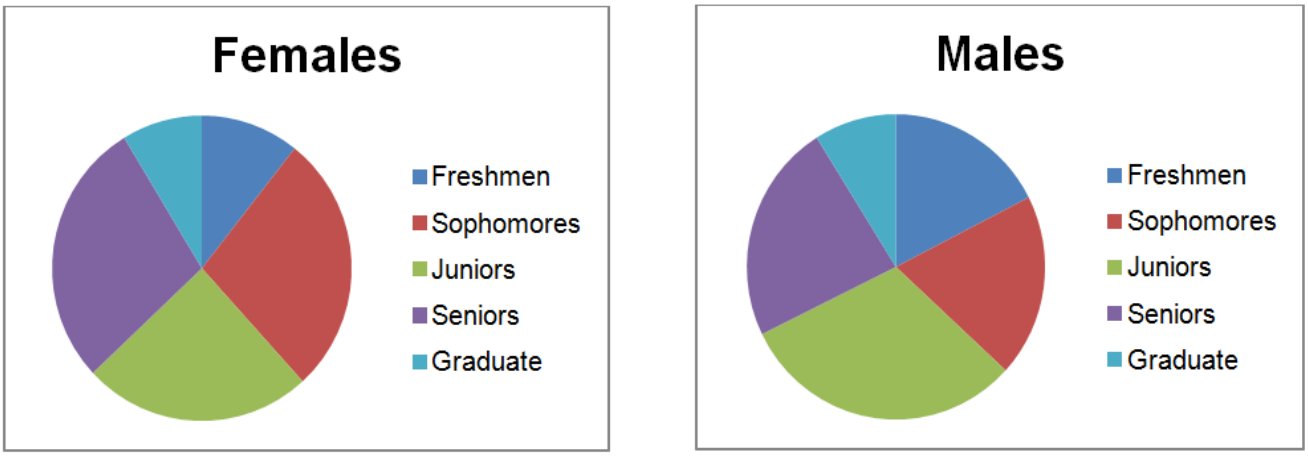

FIGURE 2 (b): Pie-Chart Representation

FIGURE 2: Different Data Representation

Prominent visualization techniques used to visualize large datasets is charting [31]. Charts or namely pie charts are the most common form of data visualization. Pie charts are both easy to understand and an elegant and fast delivery method. As most people are familiar with pie charts, they convey information relatively in a fast and direct way, see Fig2. Large database consists of millions of data records that are updated frequently, an example of such databases are Geographical Information Systems GISs. Visualization is used in GISs to visualize dataset.

Making understanding more visually and less tedious of the gathered satellite data over the course of time, weather maps and contour maps with colored regions depending on different fauna and flora. This is usually the case with real-time DBMS, where the visualization must support some kind of animation over time [32]. The time intervals of such sequenced data play a role in making the visualization more realistic and effective. A well-known example of GIS application is Google Earth and Google Maps. Google Earth and Google Maps have both changed how we deal with GISs; it is a free, online, fully visualized GIS. It is provided to the masses an interactive and visually appealing GIS system. Most people do not know that Google Map is just another GIS system visualizing a large real-time GIS system. Data visualization is used widely in computer networking as in data network traffic plotting, as we shall see later in this paper [33], Market segmentations, Anomaly detection [34], and Manufacturing [35] are among the best domains were data visualization provides tangible results.

6. APPLICATIONS OF DATA MINING

Applications of data mining vary, depending on the nature of the data to be mined. Since its inception data mining was used in various other fields. The classical application of data mining encompasses statistical and probabilistic applications. These classical applications included for example, population census studies, biosphere analysis, and marine life and oceanography. As a prominent use for data mining is data visualization, we have touched upon this application in the previous section; we focused on visualization since its importance as an application for data mining. We have selected other important applications of data mining used widely today. Many of these applications have branched out to become separate but related fields.

6.1 Spatial Data Mining

Spatial databases are databases that have unique data; this data is about space and geometry, such as the coordinates of earth, maps, and satellite data. This data is in the form of geological or geographical data. Such databases are extremely large and the data seems for the most part unrelated, and without any signs or correlations. Data mining is a natural candidate to find logic and make sense of such data. Data visualization, which was discussed earlier, is another tool of data mining heavily used in spatial databases.

Spatial databases are also used in geographical for marketing, traffic control and analysis, and GIS systems [36]. Economic geographers use such spatial data to acquire global market information such as customers' demography, manage inventory, and have logistic advantages. Another interesting feature of data visualization of large databases is that the visualization also finds relation among the non-spatial data in the database, such as local maxima and minima.

Algorithms used in spatial databases data mining and visualizations fall into neighborhood graph algorithms [37]. Beside algorithms, machine-learning techniques are also used for geometric clustering, since we have so much data, largely in 3D or topological in 2D [38]. Applications of data mining in spatial databases are mostly statistical such as the Global autocorrelation. Global autocorrelation is, basically, the calculations of the average, variance and mean of the special data [39]. On the other hand, Density analysis is an EDA process in which visualization can show at a glance we the data values are mostly concentrated, such as plotting on a map or the globe of the earth. Famous applications that rely heavily on spatial data mining include Microsoft Bing Maps and Google Earth and Google Maps. Such applications offer up-to-date information, with search capabilities allowed for end users, such as finding names, streets, and locations.

6.2 Business Intelligence

Data mining help business intelligence in many ways and for that, it is one of the fundamental tools of business intelligence. Business intelligence's (BI) goals are to gain a competitive advantage over competitors, increase productivity and effectiveness of current business operations, and to maintain a balance and control of risk management. Business intelligence is a usual task of any Enterprise Resource Planning ERP solution [40]. Business-intelligence mine habits and trends of customers' data stored as records through internet cookies and sales

profiles. This mining helps in discovering the customers' segmentations, and demographics. Data mining provides market basket analysis; items purchased together are identified and in turn bundled and advertised together. Anomalies can be also caught using intelligent mining tools; such tools mine the transactions and try to extract anomalies. Anomalies may be deliberate, such as fraudulent transactions or they could be unintentional, a glitch or bug in the program or just an odd transaction that may never be presented again in the entire database. Fraudulent transactions are caught due to their recurring characteristics, such as credit card theft, identity thefts or account hackings.

Global economies today around the world are information driven, known as knowledge-based economies [41]. Business intelligence is one of the top proponents and drivers for the development of technologies of data mining. Most data mining tools in the market today are integrated in enterprises tools such as Enterprise Resource Planning tools, and Customer Relationship Management tools. Top applications of business intelligence include market research, risk management, Market baskets, and fraud and anomaly detection. Automated business intelligence through data mining support is being used by modern enterprises in decision-making, and drive knowledge based decision rather than human imitation based decisions.

Business intelligence achieves what is known as a competitive advantage. A competitive advantage is defined as the advantage that other rivals lack, the specialty or secret skill others lack. One of the main reasons to acquire such an advantage as a competitive advantage is the competitive pressure. According to [42], competitive pressure is degree of pressure that companies feel from rivals in the market and possible new entrants. For gaining a competitive advantage, enterprises develop market research groups that analyze through data mining large datasets. Market research finds what products dominate the market, and the hidden elements that set such products from others in the market. As an example, media networks use data mining in their market research to set the common factors between audience and the program's scheduled slot. Large media groups used to hire human experts to schedule their programs slots, now the use of fully automated data mining tools for scheduling is the common trend. Results were equivalent or better than the human manual scheduling [43]. Data mining tools also discover the market's baskets, as mentioned earlier, market basket are associations of certain products that are highly likely bought together. The retail industry is dominated by market baskets predictions, giant retailers such as Wal-Mart, Costco, and K-Mart, are among the main adopters of such business intelligence achieved by data mining.

6.3 Text Mining

Another widely used application of data mining is text mining. Text mining deals with textual data rather than records stored in a regular database. It is defined as an automated discovery of hidden patterns, knowledge, or unknown information from textual data [44]. Most of data found on the World Wide Web WWW is text, after distilling the multimedia elements; most of knowledge out there is text. Text mining utilizes different techniques and methodologies, mostly linguistic and grammatical techniques, such as the Natural Language Processing NLP. Techniques of text mining originated from computational linguistics, statistics, and machine learning, such techniques were developed to make machines, specifically speaking computers, understand human language.

Text mining mines large sums of documents and articles stored in a database or even fetched from the web. The way that text mining works is very complex, were NLP algorithms try to parse sentences and matching verbs with nouns to make logical connections between all of the elements in a single sentence. Computers today are not able to understand human languages directly, not without complex AI and NLP algorithms, so mining text is still considered a daunting task, which consumes resources. Text mining to be effective, it involves a training period for the text-mining tool to comprehend the hidden and recurring patterns and relations. The process of textually mining documents involve both steps, first the linguistically analysis then the semantically analysis of the plain text. After scrutinizing plain text, mining then finally can relate

nouns and verbs, mining out some hidden traits found in the text, traits such as the frequency of use of some verbs, entity extractions such as the main characters, and possible summarizations of long documents. Text mining is used in business applications, scientific research, and in medical and biological research [45]. TM is very useful in finding and matching proteins' names and acronyms, and finding hidden relations between millions of documents.

6.4 Web Mining

With the revolution of the Internet that have changed how databases are used, this revolution brought the term of web mining. Web mining is considered as a subfield of data mining, it's regarded as the vital web technology that is used heavily to monitor and regulate web traffic. Web mining is further divided into three main sub groups, web content mining, web structure mining, and web usage mining [46]. Web content mining is the mining of content found on the web, this include metadata, multimedia, hyperlinks and text. Web structure mining is considered with the semantics and hyperlinks making up a website or a network. Web structure mining are usually is used by search engines to 'crawl' the web and find all possible links forming a network. Web usage mining is considered with the traffic patterns in the World Wide Web WWW. Most of the data is mined from the web servers and web proxies. Web servers log most traffic, such logs are the data needed to construct an overview map of the traffic coming and going to that web site.
Web mining is used in Information Retrieval IR systems, such as search engines. Web mining is also used in web trafficking measures, were traffic is traced and monitored. But for the most times, web mining is used for business intelligence [47], as it can search the web with all its fuzziness to retrieve business oriented information from the web.

7. TOOLS

Data mining tools are basically software packages, whether integrated packages or individual packages. These sophisticated software tools often require special data analysts. Such analysts are trained to use such tools, as data mining itself is not a straightforward process. It is worth mentioning that data mining tools need a substantial investment in hardware and software, as well as human resources. Deployment of data mining tools and packages is also an overwhelming task, in size and management, as it needs careful planning and management. In the next paragraphs, we shall look into some of the used data mining tools and data visualization tools.

7.1 Data Mining Tools

Data mining tools are also called *siftware*, for the sole reason that they 'sift' through the dataset. Data mining tools varies depending on level of their sophistication and projected level of accuracy. In 2008, the global market for business intelligence software, data mining centric software, reached over 7.8 billion USD, a vast amount. IBM *SPSS* is an example of business intelligence software package [48]; it is integrated data mining software with diverse business intelligence capabilities. IBM also provides online services for web mining, these services are called *Surfaid* Analytics; they provide sophisticated tools for web mining [49]. Other data mining with business intelligence capabilities is *Oracle Data Mining* [50], a part of the company's flagship RDBMS software suite. SAS also offers its *SAS Enterprise Miner* [51], as a part of its enterprise solutions. SAP, a world-renowned business solution provider, offers world known ERP solutions along with providing other mining tools software that can be integrated into their ERP solutions. Other software companies include Microsoft; it offers *SQL Server Analysis Services*, a platform dependent solution integrated in Microsoft SQL platform for Microsoft Windows Server. Microsoft also offers a less sophisticated product, namely the *PowerPivot*, a mining tool for small and middle size enterprises, with limitations and ease of use to match with its nature of use. Open source mining tools exist; they include the *Waikato Environment for Knowledge Analysis* or WEKA [52].

With the huge decline of the costs of both storing and acquiring data, through utilizing mining tools to mine web, documents, or the use of data acquisition tools such as RFID tag readers and imaging devices, data mining tools are being adapted more rapidly and incorporated into almost every business tools in the market today.

7.2 Data Visualization tools

For Data Visualization tools, we have checked IBM's *Parallel Visual Explorer*. This software package is used for market analysis, oil exploration, engineering and aerospace applications, and agriculture to name a few.

For medical fields, Parallel Visual Explorer is used to analyze various effects of treatments on the immune system. It helps in visualizing many different diverse effects on the patients' immune system [53]. For manufacturing, this tool helps in monitoring the processing parameters. Process parameters are vital for effective streamlined production. For agricultural usages, this tool helps in determining which seed to plant by analyzing the soil parameters with taking in consideration the weather conditions. Finally, Parallel Visual Explorer is also used for market research such as providing visual aids to help market analyst find customers trends, habits, and buying sprees.

An interesting visualization tool is Cave5D [54]. It's a data visualization tool developed by the university of Wisconsin-Madison. The inventors of this tool are Glen Wheless, Cathy Lascara, from the center for Pacific Oceanography, with Bill Hibbard and Brian Paul back in 1994. This software ran as a package for the Vis5D software. Cave5 provides interactive 3D, time variable visualization of dataset in a virtual environment. Cave5D integrates Vis5D's libraries and framework; it uses its graphical libraries to model the dataset. An image showing Cave5D in actual usage is shown in Fig. 3.

Vis5D [55] is a visualization system used for 3D animated simulation of weather and geological data. Vis5D uses 5D arrays that contain the time sequences of the 3D spatial datasets. Vis5D was incorporated into Cave5D through its extendable PLI libraries. Cave5D and Vis5D have their limitations as only relatively medium to small datasets can be visualized and animated at the same time.

FIGURE 3: Cave5D

In [56] the paper offer a system that offers a simple interface that overcomes the difficulties faced by other visualization tools. The system utilizes tree structures to visualize the data. Its interface allows the users to zoom in on data set as well as dynamic branching. Navigation controls are also given, to allow for smooth switching in and out of the dataset trees. Visualizations can be in pie charts, scatterplots, and histograms among others. This proposed system was compared to *Polaris* of [57].

FlowScan is network traffic flow reporting and visualizing tool proposed by Dave Plonka in [33]. FlowScan is a collection of software that includes flow collection engine, a database, and a

visualization tool. At 2000, FlowScan is an early indication of the need for visualization for data, especially for prolific network data.

Tools such as *Spotfire* and *XGobi* provide the user with predefined query visualization tools. These tools also have interactive functionalities such as zooming and brushing, which enable for finer graining the results. Academic tools such as Visage and Polaris offer similar functionality, but with custom block building query tools. Polaris, which is a visual query declarative language, has been extended to the Tableau software. Polaris offers Gant charts, scatter plots, maps, and tables.

Visualization tools are abundant. These tools range from internet network visualization, music information network, social network tagging, and web feeds visualization tools. Internet visualization tools are abundant over the internet, these include Mapping the Blogsphere, Websites as Graphs, and Opte Project. These tools offer to visualize the network from a single computer as neural networks. Music information tools such as Tuneglue, MusicMap allow the user to have a visual map of the artist of their choice and the other related artists, bands, and musical movements that influence the target of the search. Fidg't, TwittEarth, and Flickr Related Tag Browser all offer visualized social networking information. The first, offer you to Flickr and Last.FM tags to compare them to your network tagging activities. The second tool correlates a map of the world and the tweets made from twitter arising from their geographical locations. The third of this kind, offers the search through a visualized map of tags and their related tags. Other visualization tools such as Visualizing Information Flow in Science allows for a visualized view of citations used throughout scientific journals and are used to evaluate them [58].

8. FUTURE TRENDS

Future trends for data mining lie in the hands of innovation and scientific breakthrough. As data mining is both a difficult problem, and a relatively new problem that incorporates many interdisciplinary fields. We shall see some new trends that will shape the way that data mining will be used in the upcoming future. Visualization tools are also witnessing a rise, credited to the newer technologies in human-computer interactions.

8.1 Cloud Computing Based Data Mining

A relatively new trend in utilizing and benefiting from data mining tools for middle-sized and small enterprises, incapable of supporting a full-fledged data mining solution, is *cloud computing* based data mining tools [59]. Because small and middle-sized enterprises usually lack the infrastructure and budget available for large enterprises, they tend to try this new cost effective trend. Cloud computing promises to provide data mining tools benefits at relatively lower costs form such small or middle sized enterprises. Cloud computing provides web data warehousing facilities, were the actual data warehouse [60] [61] application is outsourced and accessed entirely through the World Wide Web. Cloud based data mining also provides sophisticated mining analysis of the dataset, comparable to actual data mining software, as the enterprise specifies and demands.

Aside from lowering the costs of the data mining software tools infrastructure, cloud based mining also provides expertise that is not available in such middle-sized and small enterprises. Most cloud based data mining providers tend to have data experts, data analysis, and a broader experience with data mining then their clientele. Usually start-ups or entrepreneur level enterprises lack not only the financial resources but also the human resources and expertise in the Information Technology IT field, not to mention in the data analysis field.

The *Infrastructure-As-Service IAS* helps middle-sized and start-up enterprises to be rid from the burden of software, hardware, and human resources management costs. It also helps in reducing the already limited budget. The main downfalls of cloud computing based data mining are the dependency and privacy issues that occur from the fact that another party that the enterprise have to agree to store its data on its machines and data warehouses facilities. Such issues are the main reason that limit and turn off large capable enterprises from going with cloud based data mining solution. These enterprises, large enough and have huge IT resources, can set up their

own data mining solutions instead of taking the much less needed risks. *Dependency* is another problem, it means that the whole service depends on the other party, not the enterprise itself, meaning that the enterprise is pretty much tied up with what the service provider has to offer, huge switching costs. The privacy concerns arise from the fact that the enterprise's data is technically not under its control or even possession, the other party has it, it utilize its resources to give results and analysis. The privacy concern entails the misuse of the data, mostly causing confidentiality risks.

8.2 Data Conditioning Tools

Data conditioning is currently a technique that is not only meant for data mining. It is used for intelligent routing, privacy and protection as well as for data mining. As data grows today in unprecedented rate, the need to clean up the huge piles of data is necessary. Reports suggest that more than 80% of enterprises data are unstructured and fuzzy data [62]. The other goal of data conditioning is to elevate or at least minimize the interference of IT people. This would quicken the BI step, and in turn make it ubiquitous for the end-users, whether business or science users.

The key technique used for data conditioning for data mining is data warehousing. Data warehousing is used for organizing such unstructured data, it's the middleware that transfer data from the transactional database into a structured, aggregated warehouse [63]. Data warehousing is tasked with data extractions transformations, and load, this is known as the ETL process were the data is modified to be stored in the warehouse. Data in the data warehouse is not like its previous form were it was in the original database, it's an aggregated more cleaned version.

Usually data mining processes are done on the data stored in the data warehouse as it has already cleaned and formatted for the analysis tool, which will mine useful knowledge. The quality of the mined knowledge depends heavily on the data warehouse design and model used. Finally, we can deduce that data mining efficiency as well as quality is highly affected by the level of structures and aggregation found in the data warehouse [64].

8.3 Human like Intelligence

The goal of today's data mining tools is to reach human experts level, in terms of accuracy and innovation. The promise of such intelligence lies in incorporating more AI techniques into data mining tools. This newfound intelligence will help incorporate data mining into fields that was not usual for such mining to occur. Technically the data mining is one of the main uses of AI algorithms commercially available today among other data-mining related fields [65].

Such intelligence incorporation has led to fraud detection mining tools, summarization, predictive analysis, and information retrieval tasks to name a few. IBM's SPSS, statistical modeling software, usages many AI techniques, incorporating machine learning also. Data mining seems to be the most prominent frontier were AI is currently thriving. In addition, a new technique rising in the field of AI in data mining is soft computing. *Soft computing* is considered with computing techniques that tolerate and exploit imprecision, uncertainty, approximation and reasoning [66]. This new and promising technique allows for traceability, robustness, and close resemblance, forming the new term of Machine IQ. *Fuzzy logic* also is a contributor to the advancement of newer more intelligent data mining techniques.

8.4 Interactive Visualization

The trend for the visualization tools is being more and more interactive with the user [67]. This is due to the advances in User Interfaces (UI) designs, from graphical interfaces, voice recognition, to touch sensitive displays. This trend of visualization graphs is called *advanced* visualization as opposed to the olden types of *static* graphs such as pie charts, histograms and scatter plots. While the interactive -advanced- visualization tools do have limits such as the need of a multimedia medium such a monitor of a computer, laptop, or a tablet, they are still have the edge of being able to show more complex structures through zooming in and out, 3D rotation, and/or changes in datasets by enabling user input. These types of interactive tools can also be

embedded into systems and websites, due to their nature of being targeted toward end users and able to have multiple outputs.

Interactive visualization must also keep their level of details to a tolerable degree, because some tools might go as far as to require programming of languages or structures, to evaluate datasets. While this may be acceptable for scientists and researchers, however, among business users it is unwelcomed. On the other hand, performance is another parameter that will appreciate among the latter, but might not be a key aspect to the former group. All groups of users welcome the level of accessibility of such interactive tools, such as changing the colors, font sizes, and font types among other features and configurations that allow any user with any level of visual media perception to use such charts and figures.

Live data feed is another factor contributing to the popularity of data visualization, especially interactive data visualizations. Hot in the data feed categories are the customers' reaction to the business, decision makers would highly appreciate the visualization of their large data sets of their customers' reaction, live and interactive. This is true especially in the case companies that have electronic data, such as websites.

9. CHALLENGES

Data Mining and Data Visualization is usually more effective if the data on which to be mined are conditioned beforehand. Future directions show the usage of visualizations output as inputs for Data Mining through the tight integration of implementing visual and pattern recognition algorithms in Data Mining functions themselves. Selecting a data mining algorithm can also be challenging. The user must select an algorithm that would represent the set of data accurately; a method to evaluate the representation; and a search criterion [68].

9.1 Challenges in Data Mining

Currently data mining, in the form we know it today, has not really achieved the potential of what was expected, envisioned in the late 1980's or early 1990's. The vision of becoming a mainstream application, it's widely used but to a degree still limited, data mining hasn't reached that vision. Challenges come in many forms, mainly in three categories, technical, legal, and ethical challenges, all of which they hinder adaption of data mining as a common practice. We shall examine some of these challenges that hinder the further development of data mining in the next few paragraphs.

Technically, data mining is not an application widely adapted by enterprises. It did not reach the level of a common desktop application, still. Although this was the intended goal envision for data mining. It was intended to grow until it reaches the desktop level. The technical issues, such as the huge and elaborate hardware and software infrastructure, are a usual suspect, because data mining requires substantial resources to be deployed and careful thinking and planning, to be effective. Usually the cost of a typical data mining tool in millions, as such is evident in integrating these tools into full ERP systems.

Other than costs, technical issues reside in the human resources as well; data mining require expert data analysts. These analysts will design and perform tasks on the data mining tool. Finally, the technical challenges can also manifest themselves in the limitations we have today in the current tools. Most data mining tools are not extendable, or easily upgradable or adaptable to other applications. These tools are hardwired into using a set of models based on certain best known methodologies. For example, a data measurement for business intelligence is hardly ever useful for a medical application. In addition, the limitations of today's tools are such that they cannot really replace, although we have good progress in this direction, the human element.

Ethical challenges plaguing data mining originate from the public concerns about their personal data found on the Internet. The privacy issues stems from the fact that mining can link, find, and relates the public profiles, personal preferences and possibly private data such as emails and

photos. The initial goal of data mining tools is not to identify such individuals; to counter act this, most dominated data are anonymized before it is published into the public internet. However, still huge concerns are raised around how enterprises may want to exploit the individuals' data for such mining purposes.

Other than privacy concerns around data mining application for business, governments are utilizing data mining tools for its own security and national security purposes. Such governmental security agencies are sifting through public data to locate certain wanted individuals, possible terrorists or other convicts. These uses, along with the business uses of data mining tools; made public awareness of their legal and privacy implications more evident in programs like *Total Information Awareness* program [69]. The Total Information Awareness Program was a secret program sponsored by the Pentagon; it was aimed at national security and the possible identification of terrorists. It used mining tools to mine private individuals' records on a massive scale. Public awareness against the exploitation of individuals' privacy and private data forced the congress in 2001 to stop the funding for this program. Legal and regulatory acts were issued to address these ethical concerns, acts like the *Health Insurance Portability and Accountability Acts* HIPAA, in the United States stated by the congress. The HIPAA act requires a prior consent from individuals regarding the use of their information and the notification of the purpose will their information will be used. Another ethical issue in raised by data mining tools is that they made *Globalization* far easier. Globalization has dire consequences on emerging economies, emerging businesses that can never compete with international top companies utilizing data mining.

Legislatively data mining has resulted in new levels of transparency in the free market globally. This is due to the vast data decimation across the internet willingly or unwillingly. *Wikileaks* for example, have had a hand in decimating much documentation otherwise thought to be secrets. The term *data quality* [70] is a relatively recent term, refers to authentic, complete, and accurate data and that the source of this data is legally liable for its authenticity of its quality. Governments, international legislative and professional organizations have made standardizations and regulations regarding the quality of published data from companies and other agencies. According to Will Hedfield case study in [59], more than 25% of critical data in leading organizations' databases are inaccurate and incomplete.

9.2 Challenges in Data Visualization

Challenges in visualizing large databases arise from the fact that an intuitive interface is as important and critical as any part of the visualization tool. Allowing the user to have a full view of the database, and then allow the user to zoom in on a piece of information, as more zooming is allowed, navigation tools to allow the user to change the path of zooming-in different directions, which can greatly complicate the querying and negatively decrease its performance. Suggested solution include the use of cubes to visualize the data hierarchy levels, however data cubes grow more complex far more quickly with the growth of the data set, and making the visualization huge, complex, and unintuitive. Another problem in visualizing databases is when the database itself is large in size, the form, tool, or graph used to visualize the data mined cannot be anything but cluttered or difficult to grasp.

Challenges facing Visualization tools are the limitations of the current human-computer interaction frontiers. For example, in the past decade, touch sensitive screen, although available, were limited in functionality, whether because of the software, or hardware itself. This had led to limitations of their use, but on the turn of its end, the last decade witnessed a rise in the touch sensitive screens with integrated touch sensitive functionalities such as the slates, pads, and tablets available in the market, ranging from several developers. This allows for more and better interaction levels with the visualization graphs such as the zooming, especially dynamic branching in zooming in on 3D objects [71]. Other potential of these human-computer interaction see the use of holograms, and 3D screens [72] [73] [74]. Each of which has its set of challenges as well, such as availability, cost, and reliability since most of these technologies are offered by limited vendors.

Other hazards still exist in the form of designing such interfaces to cope with these levels of interactions [75], [76]. For example, double clicking and dragging objects should still be in use instead of demanding the user to delve into levels and levels of menus and submenus, or the use of difficult finger swipes or error prone voice commands.

Data feeds to such tools are also a challenge; can a visualization tool offer interaction and real time modeling of data feeds? Most visualizations offer interaction but they are fed with linear sets of data, not online data which is refreshed at rates that may reach each second.

Unstructured data, which can be more than 70% - 80% of an organization data [77], also offer difficulties for visualization tools as well. Unstructured data comes in many forms; one form is text in a document. A document can be in any form, data and information are interleaved together, and must be extracted in order to infer facts. Data Mining tools and computational intelligence tools can be used to structure these data, in the form of an index. Even though the Data Mining tools have formatted the data, it is still difficult to visualize. For example it may be represented as a neural network, which is still hard for the user to navigate through, in and out, because of its branching paths, which are scattered in multiple directions.

10. CONCLUSION

Data mining is a vast, yet an emerging, computer science field. Widely vast and encompassing many other subfields such as *web mining* and *text mining*, and overlaps with fields such as such as *text mining, machine learning, fuzzy logic, probabilistic reasoning,* and *computational intelligence*. Data mining and Visualization have developed a lot since its inception from hundreds of years ago. Many application of data mining have gotten a huge adaption and user base. Google, the internet giant is one of the main adaptors of data mining. Data visualization today is helping to solve many engineering and scientific problems in ways that were unimaginable before, such as Map-Reduce algorithms.

Future trends in data mining and visualization are becoming more apparent with every new introduction of newer data mining solutions and data visualizations tools. Most of such advancements are based on AI, such tools aims at human like intelligence. The aim is of is to replace the human factor in decision-making. Data conditioning is a promising solution to the amounts of data that is on the rise, which is of need to be mined. For visualization tools, interactivity is the new wave, allowing users to touch, rotate, and select how to view data sets on its wake.

Future work includes more investigation on the current challenges facing the development and wide spread use of data mining and visualization techniques. The current emerging automated data conditioning tools that provide a more effective dataset to be processed by the data mining tools had an enormous impact on how data mining and visualization tools are designed. With those emerging tools in mind, data mining and visualization tools can achieve more than accurate results than before. Also more work should be done in terms of the current ethical issues associated with data mining and visualization techniques, namely the anonymization problem associated with the privacy concerns of the general public. How to ethically sift through data records without harming or breaching others privacy.

Today, most leading enterprises and organizations depend heavily on such tools for decision support, and business intelligence. Finally, it is clear now how data mining and visualization tools are essential in the knowledge discovery process, and they have an enormous impact on businesses and the research facilities. While both are separate and have their respectful principals and methods, their integration is eminent for the benefit of both fields.

11. REFERENCES

[1] D. Alexander “Data Mining” Internet: http://www.laits.utexas.edu/~norman/BUS.FOR/course.mat/Alex/, [Mar. 11, 2012].

[2] B. Palace, "Data Mining," Internet: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm, spring 1996 [Feb. 25, 2012].

[3] M. Sousa, M. Mattoso, and N. Ebecken "Data mining: a database perspective," In Proc. *International Conference on Data Mining*, 1998, pp.413-431.

[4] G. Dennis Jr, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biology*, vol. 4, pp.3-14, August 2003.

[5] V. Friedman, "Data Visualization: Modern Approaches," Internet: http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches, Aug. 2, 2007 [Mar. 12, 2012].

[6] R. Mikut, and M. Reischl "Data mining tools" *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp.431-443 , September/October 2011.

[7] D. Tegarden, "Business Information Visualization," *Communications of AIS,* vol. 1, January 1999.

[8] S. Few, "Human Perception," Internet: http://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html, Sept. 16, 2010 [Mar. 16, 2012].

[9] F. Post, G. Nielson, and G. Bonneau "Data Visualization: the State of the Art," United States of America: Springer, 2002, pp.464.

[10] G. Grinstein, and B. Thuraisingham, "Data Mining and Data Visualization" in Proc. of *the IEEE Visualization '95 Workshop on Database Issues for Data Visualization*, October 1995, pp.54-56.

[11] D. Keim "Visual Techniques for Exploring Databases," *International Conference on Knowledge Discovery in Databases (KDD '97)*, California, USA, August 1997.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," in AI Magazine, *American Association for Artificial Intelligence AAAI*, vol. 17, pp. 37-54, Fall 1996.

[13] M. Friendly "A Brief History of Data Visualization," *Handbook of Computational Statistics: Data Visualization*, vol.2, pp. 15-56, 2008.

[14] E. Tufte, "The Visual Display of Quantitative Information," *Cheshire, CT: Graphics Press*, 1986, pp.200.

[15] S. Allen, "The Value of Many Eyes," Internet: www.interactiondesign.sva.edu/classes/datavisualization/updates, Jul. 29, 2010 [Apr. 1, 2012].

[16] P. Kochevar, "Database Management for Data Visualization," *Database Issues for Data Visualization*, vol.871, pp.107-117, 1994.

[17] V. Friedman, "Data Visualization and Infographics,"Internet: http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics, Jan. 14, 2008 [Jan. 14, 2008].

[18] B. Alpern, and L. Carter "Hyperbox," in Proc. of IEEE Conference on Visualization '91, October 1991, pp. 133-139.

[19] M. Ferreira de Oliveira, "From visual data exploration to visual data mining: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 378-394, July-September 2003.

[20] C. Romero, and S. Ventura "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol.33 (2007) pp. 135–146, 2007.

[21] L. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," *The Knowledge Engineering Review,* vol. 21, pp. 1- 24, March 2006.

[22] Q. Zhang and R. Segall, "Web Mining: A Survey of Current Research, Techniques, and Software," *International Journal of Information Technology & Decision Making,* vol.7, pp.683-720, December 2008.

[23] G. Pavlopoulos, A. Wegener, and R. Schneider, "A survey of visualization tools for biological network analysis," *BioData Mining,* vol.1, pp.12, November 2008.

[24] N. Raghavan, "Data mining in e-commerce: A survey," *SADHANA Academy Proceedings in Engineering Sciences*, vol.30, pp.275-289, April-June 2005.

[25] B. Gaddam, D. Ghosh, N. Ahmed, S. Donepudi, and V. Khadilkar, "Computational Intelligence in Data Mining," Internet: http://www.cs.lamar.edu/faculty/disrael/COSC5100/ComputationalIntelligenceInDataMining.pdf, [Apr. 1, 2012].

[26] A. Berson, S. Smith, and K. Thearling, "An Overview of Data Mining Techniques," Excerpted from the book *Building Data Mining Applications for CRM*, McGraw Hill: USA, 1999, pp.488.

[27] D. Keim "Pixel-oriented Visualization Techniques for Exploring Very Large Databases," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 58-77, March 1996.

[28] D. Keim, and H. Kriegel "Visualization Techniques for Mining Large Databases: A Comparison" *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, pp.923-938, December 1996.

[29] D. Asimov, "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Science & Statistical Computing,* vol. 6, pp. 128-143, 1985.

[30] M. Oliveira, and H. Levkowitz "From Visual Data Exploration to Visual Data Mining: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 378 - 394, July-September 2003.

[31] Information Management, "Charting information management how your business works," Internet: www.information-management.com/media/ui/mk2010.pdf, 2010 [Apr. 1, 2012].

[32] S. Casner "A Task-Analytic Approach to the Automated Design of Graphic Presentations," *ACM Transactions on Graphics*, vol.10, pp.111–151, April 1991.

[33] D. Plonka, "FlowScan - Network Traffic Flow Visualization and Reporting Tool," *14th Systems Administration Conference (LISA 2000),* New Orleans, Louisiana, USA, December 3– 8, 2000, pp. 305-317.

[34] M. Marwah, R. Sharma, R. Shih, C. Patel, V. Bhatia, M. Mekanapurath, R. Velumani, and S. Velayudhan, "Visualization and Knowledge Discovery in Sustainable Data Centers," Compute 2009 ACM Bangalore Chapter Compute, Bangalore, India, January 2009.

[35] MAIA Intelligence, "Business Intelligence in Manufacturing", 2009, Internet: www.maia-intelligence.com, 2008 [Apr. 1, 2012].

[36] M. Ester, H. Kriegel, and J. Sander "Spatial Data Mining: A Database Approach" Advances in Spatial Databases, vol. 1262, pp47-66, 1997.

[37] M. Erwig, and R. Gueting, "Explicit Graphs in a Functional Model for Spatial Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol.6, pp.787-803, October 1994.

[38] K. Zeitouni, "A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views," *Information Resources Management Association International Conference IRMA 2000, Data Warehousing and Mining, Anchorage*, Alaska. pp. 229-242

[39] R. Geary, "The Contiguity Ratio and Statistical Mapping," *Incorporated Statistician*, vol. 5, pp. 115-145, 1954.

[40] S. Chaudhuri, and V. Narasayya, "New Frontiers in Business Intelligence" *The 37th International Conference on Very Large Data Bases*, Seattle, Washington, pp.1502-1503.

[41] A. Mocanu, D. Litan, S. Olaru, and A. Munteanu "Information Systems in the Knowledge Based Economy" *WSEAS Transactions on Business and Economics*, vol. 7, pp.11-21, January 2010.

[42] T. Ramakrishnan, M. Jones, and A. Sidorova, "Factors influencing business intelligence (bi) data collection strategies: An empirical investigation," *Decision Support Systems*, vol. 52, pp. 486–496, January 2012.

[43] M. Fitzsimons, T. Khabaza, and C. Shearer, "The Application of Rule Induction and Neural Networks for Television Audience Prediction," *In Proceedings of ESOMAR/EMAC/AFM Symposium on Information Based Decision Making in Marketing*, Paris, November 1993, pp. 69-82.

[44] M. Hearst, "What Is Text Mining?" Internet: http://people.ischool.berkeley.edu/~hearst/text-mining.html, Oct. 17, 2003 Oct. 17, 2003 [May 2, 2012].

[45] K. Cohen KB, L. Hunter, "Getting Started in Text Mining," *Public Library of Science* PLOS, vol. 4, pp.20-22, January 2008.

[46] F. Facca, and P. Lanzi "Mining interesting knowledge from weblogs: a survey," *Data & Knowledge Engineering,* vol.53, pp. 225–241, 2005.

[47] A. Abraham, "Business Intelligence from Web Usage Mining," *Journal of Information & Knowledge Management*, vol. 2, pp. 375-390, 2003.

[48] IBM, "SPSS", Internet: http://www-01.ibm.com/support/docview.wss?uid=swg21506855, [Apr. 1, 2012].

[49] IBM, "SurfAid Analytics", Internet: http://surfaid.dfw.ibm.com, [Apr. 1, 2012].

[50] Oracle, "Oracle Data Miner 11g Release 2," Internet: http://www.oracle.com/technetwork/database/options/odm/dataminerworkflow-168677.html, Jan. 2012 [Apr. 1, 2012].

[51] SAS, "SAS Enterprise Miner," Internet: http://www.sas.com/technologies/analytics/datamining/mine, Sept. 2, 2010 [Apr. 15, 2012].

[52] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *Special Interest Group on Knowledge Discovery and Data Mining SIGKDD Explorer News*, vol. 11, pp. 10-18, June 2009.

[53] IBM, "IBM Parallel Visualizer," Internet: www.pdc.kth.se/training/Talks/SMP/.../ProgEnvCourse.htm, Sept. 22, 1998 [Apr. 15, 2012].

[54] Cave5D, "Cave5D Release 2.0," Internet: www.mcs.anl.gov/~mickelso/CAVE2.0.html, Aug. 5, 2011 [Apr. 17, 2012].

[55] W. Hibbard and D. Santek, "the Vis5D System for Easy Interactive Visualization", *Proceedings of IEEE Visualization,* pp 28-35, 1990.

[56] C. Stolte , D. Tang , and P. Hanrahan, "Multiscale Visualization Using Data Cubes," in Proc.*of the IEEE Symposium on Information Visualization (InfoVis'02)*, October 2002, pp. 28-29.

[57] C. Stolte , D. Tang , P. Hanrahan, "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp.52-65, January 2002.

[58] C. Chapman, "50 Great Examples of Data Visualization," Internet: http://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/, 2012 [Mar. 22, 2012].

[59] W. Hedfield "Case study: Jaeger uses data mining to reduce losses from crime and waste," Internet: www.computerweekly.com, 2009 [Apr. 1, 2012].

[60] Inmon W.H., "Building the Data Warehouse," Indiana, USA: J. Wiley&Sons, 1994. pp.576.

[61] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim, and A. Valencic, "Data Modeling Techniques for Data Warehousing," Internet: www.redbooks.ibm.com/redbooks/pdfs/sg242238.pdf, Feb. 1998 [Nov. 16, 2011].

[62] K. Lynn, "Search Fuels Business Intelligence for Decision Making," *TNR Global*, Available at http://www.tnrglobal.com/blog/tag/business-intelligence/, 2004-2012 [Mar. 20, 2012].

[63] L. Greenfield, "The Data Warehousing Information Center," Internet: http://www.dwinfocenter.org/against.html, 1995 [Mar. 8, 2012].

[64] J. Lawyer, and S. Chowdhury, "Best Practices in Data Warehousing to Support Business Initiatives and Needs," In Proc. *37th Annual Hawaii International Conference on System Sciences*, January 2004, pp.9.

[65] S. Badawi, "AI Computer Vision Blog," Internet: blog.samibadawi.com, Mar. 26, 2012 [Apr. 23, 2012].

[66] S.K. Pal, "Soft Data Mining, Computational Theory of Perceptions, and Rough-Fuzzy Approach", Information *Sciences (Special Issue on Soft Computing Data Mining)*, vol. 163, pp.5-12, 2004.

[67] The Global Community of Information Professionals, "What is Information management?" Internet: www.aiim.org/what-is-information-management , 2012 [Apr. 1, 2012]

[68] B. Gaddam, and S. Donepudi, "Computational intelligence," Internet: http://cs.lamar.edu/faculty/disrael/COSC5100/ComputationalIntelligenceInDataMining.pdf, 2005 [Mar. 20, 2012].

[69] K.A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *Columbia Science and Technology Law Review*, vol. 5, December 15, 2003, Available at: http://www.stlr.org/cite.cgi?volume=5&article=2
Retrieved on 1st of April 2012

[70] Carlos Rodríguez, Florian Daniel, Fabio Casati, Cinzia Cappiello "Toward Uncertain Business Intelligence: The Case of Key Indicators" *IEEE Internet Computing*, vol.14, pp.32-40, July-Aug. 2010.

[71] D. A. Keim, C. Panse, and M. Sips "Information Visualization: Scope, Techniques and Opportunities for Geovisualization" *Exploring Geovisualization*, pp.23-52, June 27, 2005. Available at http://bib.dbvis.de/uploadedFiles/124.pdf

[72] T. M. Lehtima□ki, K. Sa□a□skilahti, M. Kowiel, and T. J. Naughton, "Displaying Digital Holograms of Real-World Objects on a Mobile Device using Tilt-Based interaction" *9th Euro-American workshop on Information Optics (WIO)*, pp.1-3, July 2010.

[73] Zebra Imaging, "Motion Displays," Internet: http://www.zebraimaging.com/products/motion-displays/, 2010 [Apr. 9, 2012].

[74] Z Space, "The ZSpace Experience," Internet: http://zspace.com/about-zspace/, 2012 [Apr. 9, 2012].

[75] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough, "SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny," *Nucleic Acids Research*, vol. 37, pp.380-386, December 2009.

[76] Van den Berg, J. P. "A literature survey on planning and control of warehousing systems" *IIE Transactions*, vol. 31, pp.751–762, 1999.

[77] O. Grabova, J. Darmont, J. Chauchat, and I. Zolotaryova; "Business Intelligence for Small and Middle-Sized Enterprises," in the *Special Interest Group on Management of Data SIGMOD Record*, vol. 39, pp. 39-50, December 2010.

Analysis of the Iriscodes Bioencoding Scheme

Patrick Lacharme *patrick.lacharme@ensicaen.fr*
Universite de Caen Basse Normandie UMR 6072
GREYC F-14032, Caen, France
ENSICAEN UMR 6072 GREYC F-14050, Caen, France
CNRS UMR 6072 GREYC F-14032, Caen, France

Abstract

Cancelable biometrics is a technique used to enhance security and user privacy. These schemes are employed to generate multiple revocable data from the original biometric template. In this paper, the security of binary template transformations is evaluated, through a new transformation for iris templates, called bioencoding scheme. This transformation and its security is analyzed, using Boolean functions and non linear Boolean systems. A general discussion on binary template transformations is finally proposed.

Keywords: Cancelable Biometrics, Bioencoding Scheme, Iriscodes

1. INTRODUCTION

Biometric schemes are widely used for identification and authentication. Nevertheless, biometric techniques give rise to many security and privacy concerns, especially because biometric data cannot be revoked. The protection of these sensitive data is a major requirement for the deployment of biometric schemes. Cancelable biometrics is based on a randomized transformation for the generation of a biometric template, from the original biometric data. The additional random number (the *seed*) is used to diversify the template. This seed needs to be carefully stored in the biometric system, and is used during the verification phase. The verification procedure only applies on the transformed template.

Cancelable biometrics is first proposed by Ratha et al. for fingerprint authentication [1]. This concept is later developed on other biometric traits, as iris or face characteristics. This alternative allows the generation of a new biometric template, if the previous template is compromised, or if a new template is required for a new application. Security analysis of these schemes is realized using several properties, as required in [2], [3], [4], including mainly:

- Recognition performance: FAR and FRR of the biometric system do not decrease significantly with the template transformation.
- Non-invertibility: the original template cannot be recovered from a compromised template, even if the random seed is known.
- Unlinkability: the original template cannot be recovered from several compromised templates, even if the corresponding random seeds are known (correlation attack).

For non-invertibility and unlinkability, it should be impossible or computationnaly hard to recover the original template, with the knowledge of the seed. These two properties ensure the user's privacy with the protection of biometric data. Detailed reviews on cancelable biometrics and other biometric cryptosystems are proposed by Jain et al. in [5] and by Rathgeb and Uhl in [6].
A new cancelable biometrics scheme on iriscodes is recently presented by Ouda et al. in [7] [8], called *bioencoding scheme*. The iriscode is partitioned into separate blocks, and each block is treated separately with a pseudorandom sequence. Authors claimed that the performance of their scheme is good, based on experimental results for low block sizes. This scheme ensures diversity and is secure against invertibility. In a second paper, the authors investigated the

unlinkability property of their scheme and found vulnerabilities if several biocodes are compromised, in [9] [10]. Nevertheless, there is no computational evaluation in their analysis.

In this paper, binary template transformations for iriscodes are investigated. The bioencoding scheme and its security are analyzed and criticized. Then, a security proof against correlation attacks is given in relation to non linear Boolean systems. Nevertheless, this security proof is only usable if the block size is high. Thus, the correlation attack is practical for low block sizes. This paper concludes on a discussion of such binary transformations, directly applied on the iriscode.

2. THE BIOENCODING SCHEME

2.1 Iris Cancelable Biometrics

Iris biometrics is known for its very good performance, with low FAR and FRR rates [11] [12]. Iriscode is the most used representation of an iris biometric feature. The iriscode generation is described and improved by Daugman in [13] [14] [15], where a binary vector of 2048 bits is derived from an image of an iris. Hao et al. assume in [16] from 10 to 20 percent of error bits within an iriscode. Several effective constructions of fuzzy commitments schemes on iriscodes are presented with various error correcting codes as the Hadamard and the Reed-Solomon codes in [16], or a Reed-Muller based product code in [17]. Iris cancelable biometrics includes schemes proposed by Chong et al. [18] [19], Zuo et al. [20] or Pillai et al. [21]. More details on iris biometric cryptosystem and cancelable iris biometrics are given in [22]. All these schemes directly work on the iris feature, and not on the binary iriscode, except for the BIN_COMBO and BIN_SALT algorithms of [20]. In these two last schemes, authors propose to use a random secret as a secret permutation or a mask on iriscode. Clearly, diversity and non-invertibility are ensured if the random data is secret. Nevertheless, the iriscode is easily recovered if the secret data is compromised by an attacker in both schemes.

2.2 Description of the Bioencoding Scheme

The bioencoding scheme is a cancelable biometric scheme, with a binary template transformation, applied on the iriscode. In this paper, the iriscode is a n-bits vector. The bioencoding scheme divides the iriscode in n/m blocks of length m and applies a random transformation on each block (more precisely a pseudorandom transformation, defined by 2^m pseudorandom bits, generated from a m-bits random seed). Let S be the pseudorandom sequence of length 2^m which can be made public (or compromised). Let S[i] denotes the i-th bit of the binary sequence S. The transformation uses an address system defined by S and maps independently each of n/m blocks of the iriscode as follows. Each m-bits input block represents a number N between 0 and 2^m-1, and the corresponding output bit is S[N]. Finally, the biocode is composed of the n/m output bits, as illustred in Figure 1.

The bioencoding scheme can be described with random Boolean functions. Let f be a Boolean function with m variables, mapping $\{0,1\}^m$ to {0,1}. A Boolean function is called *balanced* if there are the same number of zeros and ones in its truth table. A Boolean function can also be described with its *algebraic normal form* (ANF), which is just a binary polynomial with m variables.

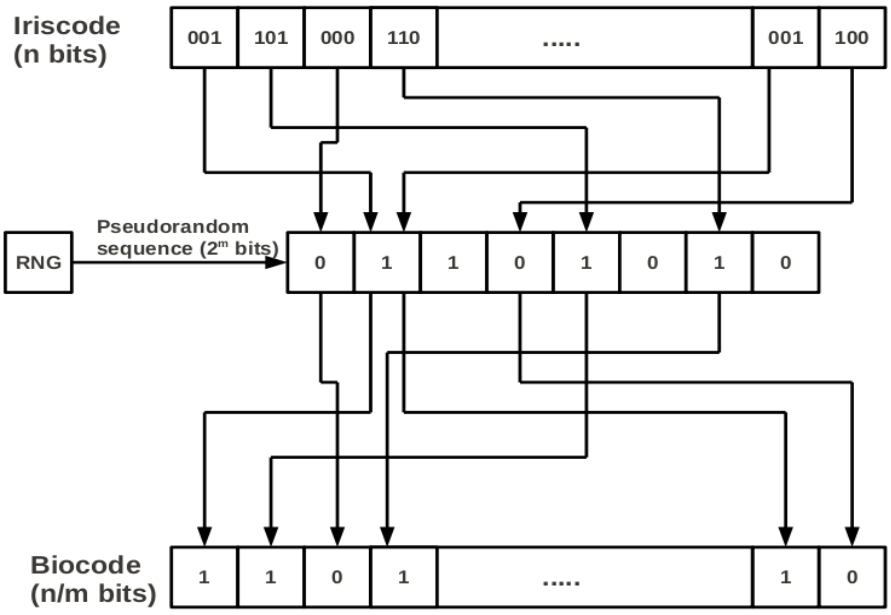

FIGURE 1: The bioencoding scheme.

For example, the balanced Boolean function f with three variables, defined by $f(x_1,x_2,x_3) = x_1+x_2+x_3+x_1.x_2 \bmod 2$ corresponds to the truth table described in Table 1. The generation of a random Boolean function with m variables requires 2^m random bits for the description of the truth table (or equivalently for coefficients of the ANF polynomial). Consequently, it is not possible to generate a random Boolean function with m variables if the number m is high. More details on Boolean functions can be found in [23].

x_1	x_2	x_3	f(x)
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

TABLE 1: Truth table of the Boolean function f

The proposed transformation takes m-bits vectors from the original n-bits iriscode and just applies a random Boolean function with m variables, corresponding to the sequence S in order to obtain one output bit. Thus, the function is applied to n/m blocks, and returns a n/m-bits biocode. The address system described by authors is not useful to understand or implement their scheme. For example, the previous Boolean function described in Table 1 corresponds to the pseudorandom sequence described in Figure 1. Thus, the revisited bioencoding scheme is presented below:

- Generate a pseudorandom Boolean function f with m variables from a random seed.
- Divide the n-bits iriscode in n/m blocks of m bits $x^{(0)}, x^{(1)},..., x^{(n/m-1)}$.
- Apply the Boolean function f to each block, such that $b^{(0)} = f(x^{(0)}),..., b^{(n/m-1)} = f(x^{(n/m-1)})$.

- Output the n/m-bits biocode $(b^{(0)},...,b^{(n/m-1)})$.

In the rest of this paper, the bioencoding scheme is described with the Boolean functions terminology. This scheme is only related with two parameters, the size n of the original binary template, divided in block of size m and the (public) Boolean function f, applied to each block.

2.3 Performance of the Bioencoding Scheme

Experimental results on the bioencoding scheme are described in [8] for very small values of m, using the CASIA iris database v_1 and later v_3. The Hamming distance d_{bio} between two biocodes, derived from two iriscodes, is clearly lower than the Hamming distance d_{iris} between the two original iriscodes. However, the length of the biocodes is divided by m. Consequently, the Hamming distance d_{bio} should verify $m.d_{bio} < d_{iris}$, in order to ensure the performance requirement. More precisely, the intra-class variability requires that f(x) + f(x') mod 2 is zero for pairs x, x' with low Hamming distance. For a Boolean function f, the *derivate* of f in a, denoted $D_a(f)$, is defined by the Boolean function $D_a(f)=f(x) + f(x + a \bmod 2) \bmod 2$. Therefore, intra-class variability requires that most of derivates of the Boolean function are zero for elements with low Hamming weight. Unfortunately, there is no general construction for such Boolean functions. Following [16], the percent of error bits in two genuine iriscodes is between 10 and 20. It is not possible to ensure a correct performance requirement, concerning the intra-class variability, with a random balanced Boolean function.

3. NON-INVERTIBILITY AND UNLINKABILITY

3.1 Non Invertibility

Template transformations are designed to produce biometric templates, from which it is computationally hard it is computationally hard to recover the original template, even with the knowledge of the seed [5]. The irreversibility of the bioencoding scheme is based on the compression rate of the Boolean function. If this function is balanced, there are 2^{m-1} inputs for each output bits, providing $2^{(m-1)n/m}$ possible inputs. Consequently, the original iriscode cannot be recovered from one compromised template by an impostor, having the knowledge of the Boolean function.

However, it must be computationally hard to find a biometric template, matching with the given template [24]. This criteria is different to the standard noninvertibility criteria. The construction of another preimage is related to the security of the scheme against spoofing atacks, as in [25] and [2] for the biohashing algorithm. For a given biocode and the knowledge of the Boolean function, it is easy to construct an iriscode which provides the same biocode (directly from the truth table). Consequently, the bioencoding scheme is not protected against spoofing attacks, if the Boolean function is known.

3.2 Unlinkability

The correlation attack proposed by Ouda et al. comes from a basic example with m = 3, where three compromised biocodes are sufficient to recover the original iriscode. This attack is described here with the Boolean functions terminology, using three Boolean functions f_1, f_2, f_3 defined in Table 2.

x_1	x_2	x_3	$f_1(x)$	$f_2(x)$	$f_3(x)$
0	0	0	0	0	1
0	0	1	1	1	0
0	1	0	1	0	0
0	1	1	0	1	1
1	0	0	1	1	1
1	0	1	0	1	0
1	1	0	1	0	1
1	1	1	0	0	0

TABLE 2: Truth table of the Boolean functions f_1, f_2, f_3

Authors consider an example where the first bits of three biocodes, defined by f_1, f_2 and f_3, are 0, 1 and 0. In Table 2, there is only one input x such that $f_1(x) = 0$, $f_2(x) = 1$ and $f_3(x) = 0$, then the first block of the orignal iriscode is x = 101. This attack is generalized for all m, by claiming that if m biocodes are compromised, then the original iriscode is recovered. However, the security of a system is generally related to the computational hardness of a given problem, as for NP problems. In this case, there are no description on the method to recover the original iriscode in general case, and the complexity of the attack is not estimated.

The correlation attack is revisited by a Boolean system, in order to evaluate the resistance of the scheme to this attack. Considering that m biocodes $b_1,..., b_m$ are compromised, with m Boolean functions $f_1,..., f_m$. Then, the correlation attack requires the resolution of n/m Boolean systems where the unknown is one of m-bits block of the original template. For example, let $x=(x_1,...,x_m)$ be the first block of the original iriscode and $b_j^{(0)}$ denotes the first bit of the j-th biocode. The following Boolean system must be solved:

$$f_1(x_1,... ,x_m)=b_1^{(0)}$$

...

$$f_m(x_1,... ,x_m)=b_m^{(0)}$$

A similar Boolean system exists for each bits of biocodes. If the Boolean functions $f_1,..., f_m$ are linear and linearly independent, then this system is easily invertible. But the probability that m random Boolean functions are linear is very low. Otherwise, if the Boolean functions are non linear, the resolution of this system is known as a NP problem, providing a security proof on the hardness to realize a correlation attack on this scheme. Nevertheless, if we want to use this NP problem, the number m can not be too small, implying a very high performance degradation.

3.3 Additional Modifications and Discussion

Additional modifications in [7] include a preliminary operation on the iriscode, involving a second random number, before the bioencoding transformation. The first proposition performs the bit-wise XOR between this random number and the iriscode. Nevertheless, the security of this scheme is only related to the secret of the random number. The second proposition uses a secret permutation on the n bits of the iriscode. This proposition is more interesting because the random permutation has not to be secret to ensure the security of biocodes. In this case, the correlation attack is determined by Boolean systems, where the n original bits are possibly involved. Consequently, this attack becomes computationally unfeasible, considering the size of n, even if all permutations are known. Moreover, the random permutation ensures the biocode diversity. Thus, the Boolean function has not to be random and can be determined to optimize the intra-

class variability. It is a strong improvment compared to the original scheme, where the Boolean function was generated at random in [8].

Unfortunately, the non-invertibility property is not verified. For a given biocode it is easy to reconstruct a new iriscode which is tranformed to the same biocode with the knowledge of the Boolean function and the permutation. A protection against spoofing attacks requires that the Boolean function (or the permutation) is not compromised. A similar vulnerability is realized by Nagar et al. for fingerprint and face biometrics in [2]. It is the reason why the non-invertibility property ensures that the construction of another preimage should be hard, as suggested in [24]. Consequently, tokenless template transformations should be very carefully designed, especially in the iriscode context, and the protection of the random token in an additional secure element is recommended for many applications.

4. CONCLUSION AND PERSPECTIVES

Binary template transformations are investigated in this paper, through the bioencoding scheme on iriscodes. This scheme is revisited with random Boolean functions and the performance of the transformation is analyzed. Thus, the bioencoding system appears to be a simple application of a random Boolean function on the original iriscode, realized block by block. The bioencoding scheme cannot ensure functional performance requirements for general block sizes. The protection of the original template is related to a Boolean system, possibly enhanced with a random permutation. However, this scheme is invertible because a lot of preimages can be reconstructed from a biocode.

The perspective of this work would be to provide a robust binary template transformation, ensuring a good intra-class variability and a strong preimage resistance in a tokenless environment. Another alternative for iris cancelable biometrics is to transform directly the iris feature, without iriscode transformation.

5. REFERENCES

[1] N. Ratha, J. Connell and R. M. Bolle. “Enhancing Security and Privacy in Biometrics biased Authentication Systems”, IBM Systems, vol 40, N 3, pp. 614-634, 2001.

[2] A. Nagar, K. Nandakumar and A. K. Jain. “Biometric template transformation: A security analysis”, SPIE, Electronic Imaging, Media Forensics and Security XII, 2010.

[3] D. Maio, D. Maltoni, A. Jain and S. Prabhakar. “ Handbook of fingerprint recognition”, Springer, 2009.

[4] R. Belguechi, E. Cherrier and C. Rosenberger. “How to Evaluate Transformation Based Cancelable Biometric Systems?”, IBPC 2012.

[5] A. K. Jain, K. Nandakumar and A. Nagar. “Biometric Template Security”, EURASIP J. Advances in Signal Processing, vol 8, N 2, pp. 1-17, 2008.

[6] C. Rathgeb and A. Uhl. “A Survey on Biometric Cryptosystems and Cancelable Biometrics”, EURASIP J. on Information Security, vol 3, 2011.

[7] O. Ouda, N. Tsumura and T. Nakaguchi. “Tokenless cancelable biometrics scheme for protecting iris codes”, ICPR'10, pp. 882-885, 2010.

[8] O. Ouda, N. Tsumura and T. Nakaguchi. “BioEncoding: A reliable tokenless cancelable biometrics scheme for protecting iriscodes”, IEICE Transaction on Information and Systems, vol E93-D, N 7, pp. 1878-1888, 2010.

[9] O. Ouda, N. Tsumura and T. Nakaguchi. “Securing BioEncoded iriscodes against Correlation Attacks”, IEEE Int. Conference on Communications, ICC'11, pp. 1-5, 2011.

[10] O. Ouda, N. Tsumura and T. Nakaguchi. “On the Security of BioEncoding Based Cancelable Biometrics”, IEICE Trans. on Information and Systems, vol E94-D, N 9, pp. 1768-1778, 2011.

[11] J. Daugman. “Probing the uniqueness and randomness of iris codes: results from 200 billion iris pair comparisons, IEEE, vol 94, N 11, pp. 1927-1935, 2006.

[12] K. W. Bowyer, K. Hollingsworth and P. J. Flynn. “Image understanding for iris biometrics: A survey”, Computer Vision and Image Understanding, vol 110, N 2, pp. 281-307, 2007.

[13] J. Daugman. “High confidence visual recognition of persons by a test of statistical independence”, IEEE Transactions on PAMI, vol 15, N 11, pp. 1148-1161, 1993.

[14] J. Daugman. “The importance of being random: Statistical principles of iris recognition”, Pattern Recognition, vol 36, N 2, pp. 279-291, 2003.

[15] J. Daugman, ”How iris recognition works”, IEEE Transactions on Circiuts and Systems for Video Technology, vol 14, N 1, pp. 21-30, 2004.

[16] F. Hao, R. Anderson and J. Daugman. “Combining crypto with biometrics effectively”, IEEE Transactions on Computers, vol 55, N 9, pp. 1081-1088, 2006.

[17] J. Bringer, H. Chabanne, G. Cohen, B. Kindarji and G. Zemor. “Optimal iris fuzzy sketches”, 1st IEEE Int. Conference on Biometrics: Theory, Applications and Systems, pp. 1-6, 2007.

[18] S. C. Chong, A. B. J. Teoh and D. C. L. Ngo. “High security iris verification system based on random secret integration”, Computer Vision and Image Understanding, vol 102, N 2, pp. 169-177, 2006.

[19] S. C. Chong, A. B. J. Teoh and D. C. L. Ngo. “Iris authentication using privatized advanced correlation filter”, International Conference in Biometrics (ICB), pp. 382-386, 2006.

[20] J. Zuo, N. K. Ratha and J. H. Connell. “Cancelable iris biometric”, Conference on Pattern Recognition (ICPR), pp. 1-4, 2008.

[21] J. K. Pillai, V. M. Patel, R. Chellappa and N. K. Ratha. “Secure and Robust Iris Recognition using Random Projections and Sparse Representations”, IEEE Transactions on pattern analysis and machine inteligence, vol 33, N 9, 2011.

[22] C. Rathgeb and A. Uhl. ” The State-of-the-Art in Iris Biometric Cryptosystems”, InTech, pp. 179-202, 2011.

[23] C. Carlet. “Boolean functions for cryptography and error correcting codes”, Cambridge Univ. Press, pp. 257-397, 2010.

[24] A. Nagar, K. Nandakumar and A. K. Jain. “MultiBiometric Cryptosystems Based on Feature-Level Fusion”, IEEE Transactions on Information Forensics and Security, vol 7, N 1, 2012.

[25] L. Nanni and A. Lumini. “Local binary patterns for a hybrid fingerprint matcher”, Pattern Recognition, vol 41, N 11, pp.3461-3466, 2008.

Smartphone Forensic Investigation Process Model

Archit Goel *writetoarchit@gmail.com*
Student, B Tech
Northern India Engineering College, GGSIPU
New Delhi, 110053, INDIA

Anurag Tyagi *tyagi_anurag@ymail.com*
Student, B Tech
Northern India Engineering College, GGSIPU
New Delhi, 110053, INDIA

Ankit Agarwal *cs.ankit11@gmail.com*
Asst Prof
Northern India Engineering College, GGSIPU
New Delhi, 110053, INDIA

Abstract

Law practitioners are in an uninterrupted battle with criminals in the application of digital/computer technologies, and these days the advancement in the use of Smartphones and social media has exponentially increased this risk. Thus it requires the development of a sound methodology to investigate Smartphones in a well defined and secured way. Computer fraud and digital crimes are growing rapidly and only very few cases result in confidence. Nowadays Smartphones accounts for the major portion as a source of digital criminal evidence. This paper tries to enlighten the development of the digital forensics process model for Smartphones, compares digital forensic methodologies, and finally proposes a systematic Smartphone forensic investigation process model. This model adapt most of the previous methodologies with rectifying shortcomings and proposes few more steps which are necessary to be considered to move with the advancement in technology.

This paper present an overview of previous forensic strategies and the difficulties now being faced by the particular domain. The proposed model explores the different processes involved in the forensic investigation of a Smartphone in the form of an fourteen- stage model. The Smartphone forensic investigation process model (SPFIPM) has been developed with the aim of guiding the a effective way to investigate a Smartphone with more area of finding the potential evidence.

Keywords: Smartphone, Forensic, Digital Evidence.

1. INTRODUCTION

Due to advancements in technologies, mobile communication devices and Personal Digital Assistants (PDAs), such as the iPhone and Blackberry, are now not limited to making voice calls only, instead they are used for browsing the Internet and accessing emails in plethora, and as the technology is progressing, it is becoming cheaper, thereby easily available and accessible to more and more people. Although the amount of data stored in such devices is much less as compared to the amount of data stored in computers, but this small amount of data can be of

great use and is potent of revealing useful information, and thus forensic examinations of mobile communication devices can be extremely fruitful.

Digital evidence, which is defined as "*Digital evidence or electronic evidence is any probative information stored or transmitted digitally and a party to a judicial dispute in court can use the same during the trial*", can be found in memory modules and data storage areas of mobile telephones. These evidence can prove an important part of a criminal or civil prosecution. Deleted text messages can be recovered, which can reveal not only purposes and objectives but also suspect's plan of action. Billing records can put light over people close to suspect and his/her associates. Physical movement of a handset can be plotted to illustrate where a suspect may have moved to and from over a period of time, by cell site analysis.

As different mobile devices are built differently, specialized forensic techniques are required to ensure that mobile telephone forensics assessments conducted are done so in a forensically sound mode and that the information extracted will endure the inquiry of a court of law.

1.1 Why Smartphone forensics?

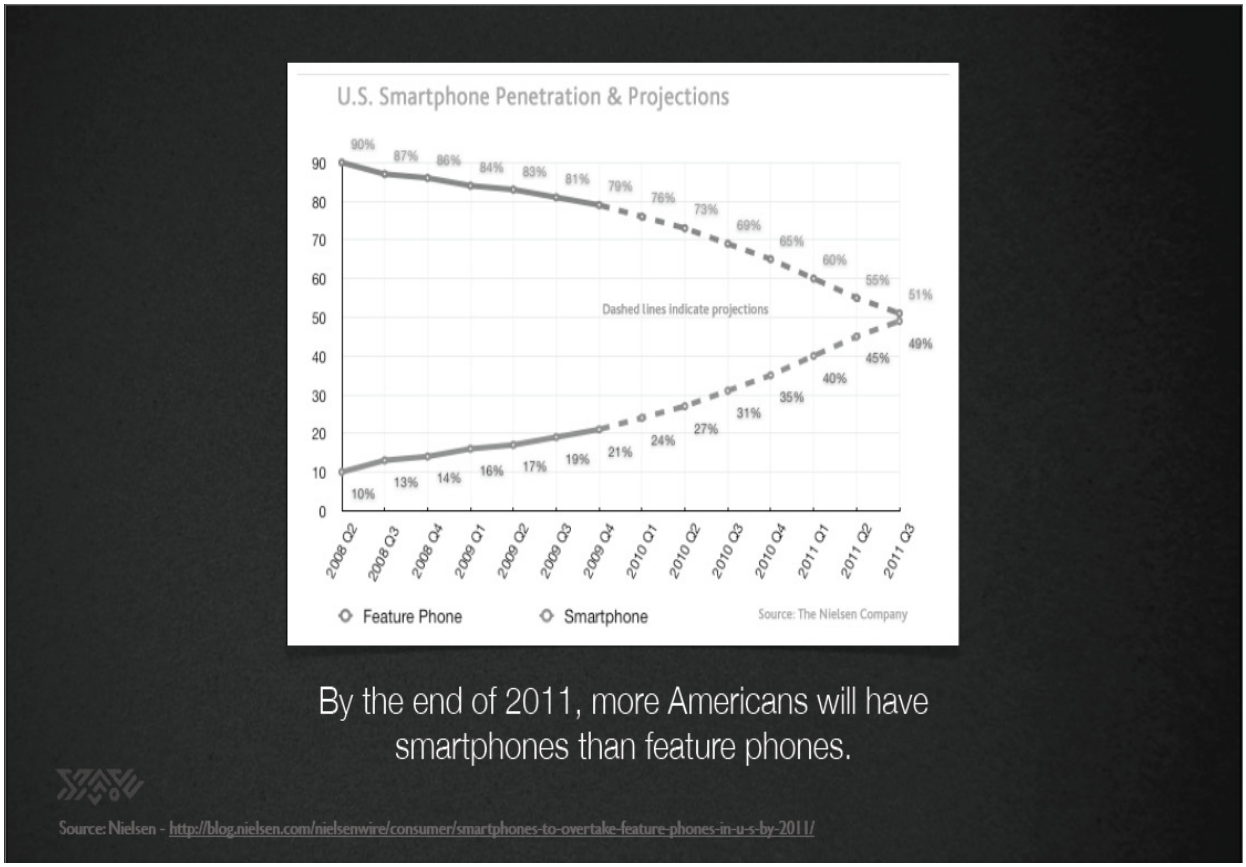

FIGURE 1: U.S. Smartphone Penetration and Projection

The following section of the paper will discuss the necessity of mobile device forensics by weighing the following:

- Use of mobile phones to amass and broadcast personal and community information
- Use of mobile phones in online transactions
- Law enforcement, criminals and mobile phone devices

1.2 Use of Smartphones to amass and broadcast personal and community information.

The evolution of mobile phone applications like Word processors, Spreadsheets, and database-based applications have transformed these devices into mobile offices with ability to store, view, edit and print electronic documents. The ability to send and receive Short Message Service (SMS) messages has transformed mobiles into a message centre. The average teenager sends 3,339 texts per month [1].The facility of Multimedia Messaging Service (MMS) in mobile phones has provided support for multimedia objects and seamless amalgamation with email gateways that enables users to send content rich emails using the MMS service. Moreover, further expediency and robust, reliable, user friendly and powerful communications capabilities are induced in mobile devices with development of technologies such as "push e-mail" and always-on connections. With Push e-mail mobile device users can access their emails at any instant, anywhere as soon as email notification arrives, using their mobile device as mail client and making it an email storage and transfer tool. Popularity of Smartphones has given this trend a whole new direction.com Score said that the July 2011 US Smartphone audience reached 82.2 million people. Morgan Stanley Research estimates that sales of Smartphones will exceed those of PCs in 2012.The Coda Research Consultancy predicts global Smartphone sales of some 2.5 billion over the 2010-2015 period, and also suggests that mobile Internet use via Smartphones will increase 50 fold by the end of that period.MS Research expects 420 million Smartphones to sell in 2011 or 28% of the mobile handset market. They predict this figure will rise to over 1 billion in 2016 (half the market).

15-25 year olds spend more than 3 hours per day on their Smartphones and 60% of this is on entertainment & browsing[2]. Data usage for 3G users is almost 44% more than 2G users.IDC (December 2009) estimates there were more than 450 million mobile Internet users worldwide in 2009; this will pass the 1 billion mark by 2013.

Nielsen Informate Mobile Insights, as the alliance is called, revealed in its most recent study that the average Smartphone user spends 2 and a half hours a day using their phones with 72% of their time spent on activities such as gaming, entertainment, apps and internet related content. Only 28% of their time is now used for voice calls and text messaging.

	15-24 years	**31 years**
Total Time spent on the Smartphone	**3 hrs**	**2 hrs**
Total Time spent on Browsing & Entertainment	**2 hrs**	**1 hr**
Total Time spent on Chat & SMS	**31mns**	**15mns**

Table 1: Time spent and activities on Smartphones
Source: Nielsen Informate Mobile Insight

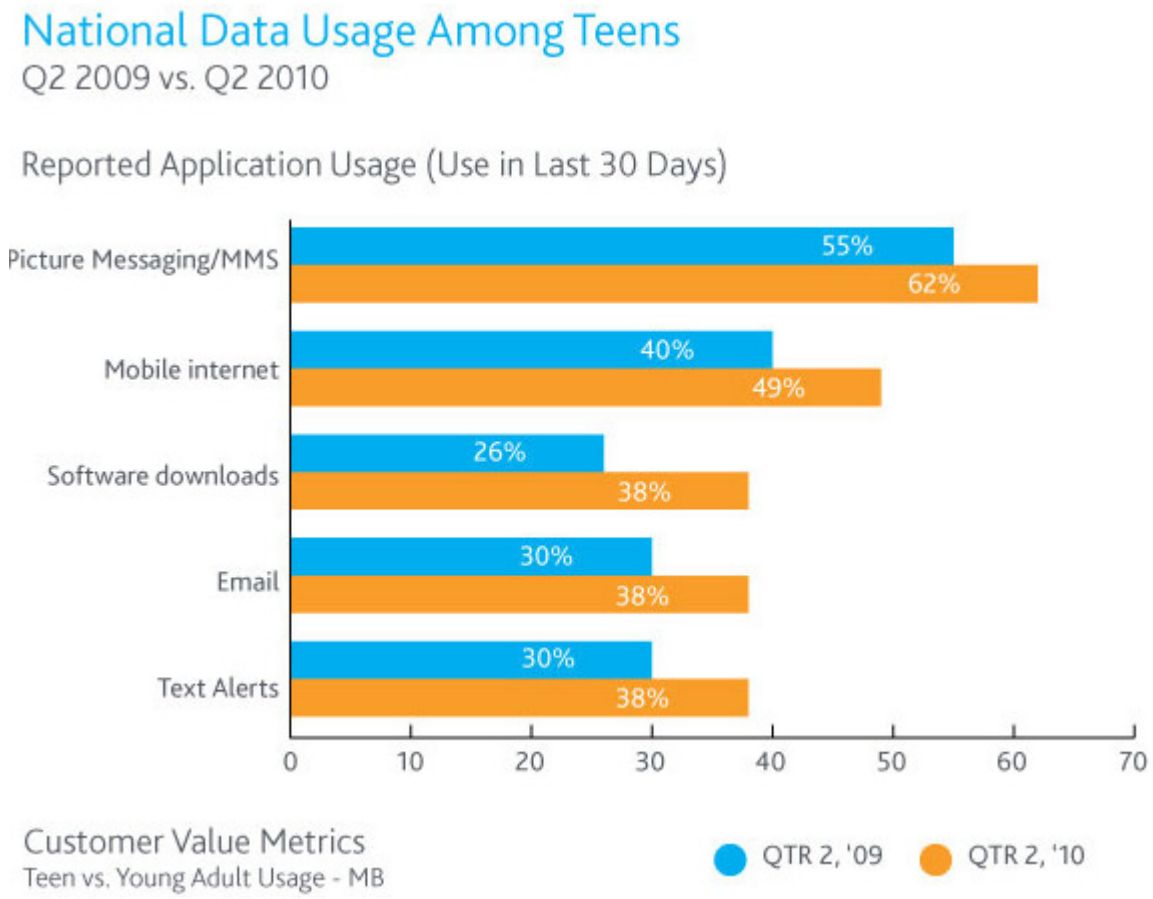

FIGURE 2: Data usage among teens
Source: Nielsen informate Mobile Insights

1.3 Use of mobile phones in online transactions

With help of Wireless Application Protocol (WAP) and technologies like digital wallets (E-Wallet) mobile phones can be used in online transactions conveniently. With enhancements in connectivity and security of mobile devices and networks enabled mobile phones to be used securely to conduct transactions such as stock trading, online shopping, mobile banking, hotel reservations and check-in [3] and flight reservations and confirmations [4]. Jeff Bezos, founder and CEO of Amazon.com(July 2010) stated that "In the last twelve months, customers around the world have ordered more than US$1 billion of products from Amazon using a mobile device" .Global Industry Analysts(GIA) (February 2010) predicts the global customer base for m-banking will reach 1.1 billion by the year 2015. Yankee Group (June 2011) predicts that there will be 500 million m-banking users globally by 2015. Currently, 27 percent of all survey respondents use mobile banking--far more than use m-commerce (13 percent), mobile coupons (11 percent) and mobile payments (9 percent).

1.4 Law enforcement, criminals and mobile phone devices

The focus on utilization of mobile phone technologies for controlling organized crimes involving usage of such technologies in one way or the other and thereby enforcing laws is surprisingly low. Mobile phones are continually used by criminals as a means to assist everyday operations and

planning from a long time back. sardonically, while it took decades to convince lawful businesses that mobile connectivity can improve their operations, just about every person involved at any level of crime already knew in the early 1980s that mobile phones can offer a considerable return on investment [5].On the other hand, due to some of the following reasons [6], law enforcement and digital forensics still lag behind when it comes to dealing with digital evidence obtained from mobile devices:

- The mobility facet of the device requires dedicated interfaces, storage media and hardware.
- The file system residing in volatile memory versus stand alone hard disk drives.
- Hibernation behavior in which processes are suspended when the device powered off or idle but at the same time, remaining active.
- The diverse variety of embedded operating systems in use today.
- The short product cycles for new devices and their respective operating systems.
- The evolving use of cloud is enabling to hide the presence of data form mobiles into web.

These differences make it important to distinguish between mobile phone and computer forensics.

2. COMPUTER FORENSICS V/S MOBILE PHONE HANDSET FORENSICS

The following sections of the paper evaluate mobile and computer forensics in the following aspects:

- Reproducibility of proofs in the case of dead forensic analysis
- Dead and live forensic analysis and their dependencies on connectivity options
- File systems (FS) and Operating systems (OS)
- Hardware variations
- Forensic technologies and tool-kits available

2.1 Reproducibility of proofs in the case of dead forensic analysis

In dead forensic analysis, an image of the entire hard disk is made after powering off the target device. The entire data of the original hard disk and the forensically acquired image of the entire hard disk is then computed using a one-way-hash function. This hash function generates a value for content of both, the original hard disk and the image of hard disk. The acquired image represents a bit-wise copy of the entire hard disk if the two values match. Then, sound forensic techniques are applied to analyze the acquired image in a lab using a trusted OS. This process is referred to as offline forensic analysis or offline forensic inspection.

A major distinction between conventional computer forensics and mobile phone forensics is the reproducibility of proofs in the case of dead forensic analysis. This is because mobiles, unlike traditional computers, remain active constantly and their content is continuously updated. The ever changing device clock in smart phones alters content of its memory constantly. Thus the forensic hash produced from such devices generates a different value every time the function is run on the device's memory [6]. This makes it impossible to obtain a bit-wise copy of whole data of a smart phone's memory.

2.2 Dead and live forensic analysis and their dependencies on connectivity options

Online analysis(Live analysis) means that the system is not taken offline neither physically nor logically [7]. The ways in which a device is connected to the outside world refers to the connectivity options. The connection may be wired or wireless. Although, connectivity options on smart phones are much more than those on traditional computers and are further evolving at a great rate, nothing noteworthy in field of live analysis has been done when it comes to smart phone handset forensics.

2.3 File systems (FS) and Operating systems (OS)

Digital forensic investigator have sound knowledge of computer operating systems but their knowledge and abilities to analyze digital evidences present in mobile phones is are very limited due to lack of knowledge about and familiarity with operating systems and file systems of mobile devices. Earlier, one of the main issues facing mobile forensics was the availability of proprietary OS versions in the market. Some of the OS versions were developed by well known manufacturers such as Nokia and Samsung while some were developed by little known Chinese, Korean and other regional manufacturers. This made developing forensics tools and testing them an onus task. Nowadays, as sales of Smartphones is peaking, most of which are produced by well known manufacturers like Apple, Google and RIM. This eases the scenario a bit as Apple and RIM devices use a specific OS and other mobile device manufacturers also use OS like Android by Google.

Now, the problem in developing efficient and reliable forensic analysis techniques is because the OS developers and even forensic tool developers are reluctant to release information about the inner workings of their codes as they regard their source code as a trade secret.

Another issue with mobile OS and FS when compared to computers is the states of operation. While computers can be clearly switched on or off, the same cannot be said about some mobile phone devices. This is especially true for mobile phones stemming from a PDA heritage where the device remains active even when it is turned off. Therefore, back-to-back dead forensic acquisitions of the same device will generate different hash values each time it is acquired even though the device is turned off [8]

A key difference between computers and mobile phones is the data storage medium. Volatile memory is used to store user data in mobile phones while computers use non-volatile hard disk drives as a storage medium. In mobile phones, this means that if the mobile phone is disconnected from a power source and the internal battery is depleted, user data can be lost. On the contrary, with non-volatile drives, even if the power source is disconnected, user data is still saved on the hard disk surface and faces no risk of deletion due to the lack of a power source. From a forensics point of view, evidence on the mobile phone device can be lost if power is not maintained on it. This means that investigators must insure that the mobile device will have a power supply attached to it to make sure data on the device is maintained.

One of the drawbacks currently facing mobile OS and FS forensic development is the extremely short OS release cycles. Symbian, a well known developer of mobile phone operating systems is a prime example of the short life cycle of each of its OS releases. Symbian produces a major release every twelve months or less with minor releases coming in between those major releases. This short release cycle makes timely development, testing and release of forensic tools and updates that deal with the newer OS releases difficult to achieve.

2.4 Hardware variations

As Smart phones are portable devices and have a specific set of functionalities unlike the large general purpose computers, the hardware architecture of smart phones is significantly different from that of computers. Thus the common characteristics of a smart phone vary from those of a computer in the way it stores the OS, its processor functions and behaves and it handles its memory(both internal and external).

The typical hardware architecture of a smart phone typically consists of a microprocessor, main board, Read Only Memory (ROM), Random Access Memory (RAM), a radio module or antenna , a digital signal processor, a display unit, a microphone and speaker, an input interface device (i.e., keypad, keyboard, or touch screen) and a battery. The OS is generally stored in ROM, which may be re-flashed and updated by the user of the phone by downloading a file from web and executing it on a personal computer that is connected to the phone device. These ROM updates may be hardware specific or OS specific. Other user data and settings are stored in RAM.

Some smart phones might include supplementary devices and modules like a digital camera, Global Positioning device (GPS), and even a small hard disk. OS is highly customized by manufacturers to fulfill user demands and to suit their hardware devices [9]. Thus different hardware devices may have different OS versions and specifications even if the two devices are fairly similar to each other, the same is applicable for two different(in terms of hardware) devices manufactured by a same manufacturer. Various phone providers may customize some options in device's ROMs, this causes variations to occur between two identical phones purchased from different providers. Proprietary hardware is a further concern for smart phone forensics. Mobile forensic tools hardly provide any support for such devices. About 16% of mobile phones in the market today come from proprietary manufacturers and are not supported by forensic tools. Furthermore, several smart phones have an interface which can not accessible through a computer. IN such cases forensic analysis of the device becomes even harder.

Another major factor which hinders use of existing forensic tools and presents challenges for new forensic tools under development is the small product cycle. Smart phones get out dated and out of use so rapidly that continually building forensic tools fit for newer device type is a uphill task.

2.5 Forensic technologies and tool-kits available

Earlier as mobile phones had very limited functionalities and a very limited information storage capacity, the focus was more on phone records from the telecommunications companies rather than the analysis of the device itself. But, today smart phones have large memories, loads of functionalities and applications, and many connectivity options. Mobile phone forensic tools and toolkits are not as advanced as required to match up the growth in mobile phone devices. These forensic tools are developed by third party companies and are rarely tested and verified using any sound methodology. The developers of the toolkits admit to using both, manufacturer supplied and self developed commands and access methods to gain data access to memory on mobile devices [10]. One such tool supports a very limited number of devices. Also, the extent of information a tool can extract varies and is generally quiet limited. Moreover, while some toolkits provide acquisition capabilities, they do not provide examination or reporting facilities [8]. Furthermore, direct access to information on the smart phone is not always attainable. Phone software and/or hardware must be used to obtain data from the smart phone's memory as shown in Figure:

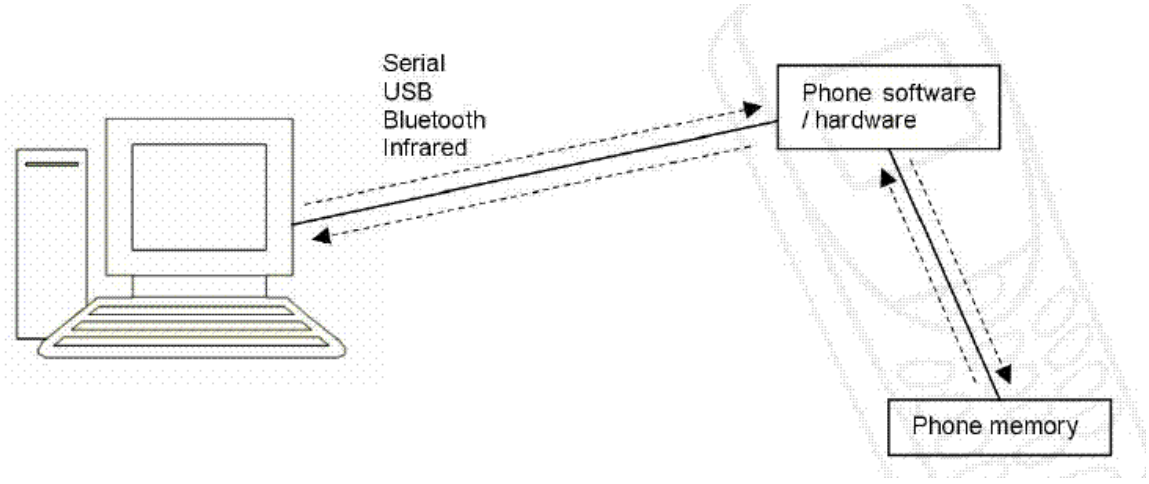

Figure 3: Indirect Access to Data in Mobile Phone Memory via Software and Hardware Commands and Methods [10].

To make this data trustable, evaluation of mobile forensic tools becomes a fundamental component of their development process Today, only a single tools evaluation document is available for mobile phone forensics and it is published by the National Institute of Standards and

Technology (NIST) in the United States [6]. Eight mobile phone forensic toolkits are evaluated in the document. A variety of devices from basic to smart phones are covered in the document. The document agreed on the state that no toolkit is available for successful forensic analysis of all mobile phone devices. But the document restricted its scope to a set of scenarios with a specific set of given activities that were used to estimate the capabilities of each of the eight toolkits under evaluation. Also, the document tested the toolkits in one set of conditions which was a virtual machine installed on a windows machine. This insured toolkit segregation and ruled out the possibility of conflicts amongst the tools [8].

3. MOBILE PHONE DATA AS EVIDENCE

This section of the paper will highlight some forensic definitions, principles and best practice guidelines and how they address mobile phone forensics issues. In this section, some of the forensic guides that address mobile phone forensics are discussed and their shortcomings or flaws are mentioned.

3.1 Definition of Digital Evidence

According to the Scientific Working Group on Digital Evidence (SWGDE), Digital Evidence is "information of probative value that is stored or transmitted in binary form". Thus any useful information stored or transferred in digital mode is an evidence regardless of the devices or interfaces used to store or transfer it. Therefore, smart phones are a promising site for collecting such evidence.

The Australian Standards HB171 document titled "Guidelines for the Management of IT Evidence" refers to IT Evidence as: "any information, whether subject to human intervention or otherwise, that has been extracted from a computer. IT evidence must be in a human readable form or able to be interpreted by persons who are skilled in the representation of such information with the assistance of a computer program". It is a flawed definition as it overlooks all possible sources for collecting digital evidence other than computers. Even the Information Technology Act 2000 (No. 21 of 2000) is not modernized to comprise information about mobile phone evidence

3.2 Principles of Electronic Evidence

United Kingdom's Association of Chief Police Officers (ACPO) Good Practice Guide for Computer based Electronic Evidence, proposed four principles to be followed while dealing with Computer-Based Electronic Evidences [11]:

- Principle 1: No action taken by law enforcement agencies or their agents should change data held on a computer or storage media which may subsequently be relied upon in court.
- Principle 2: In exceptional circumstances, where a person finds it necessary to access original data held on a computer or on storage media, that person must be competent to do so and be able to give evidence explaining the relevance and the implications of their actions.
- Principle 3: An audit trail or other record of all processes applied to computer based electronic evidence should be created and preserved. An independent third party should be able to examine those processes and achieve the same result.
- Principle 4: The person in charge of the investigation (the case officer) has overall responsibility for ensuring that the law and these principles are adhered to.

ACPO's guide regards computer based electronic evidence as no different from documentary evidence and as such is subject to the same rules and laws that apply to documentary evidence [11]. The ACPO guide also recognized that not all electronic evidence can fall into the scope of its guide and gave an example of smart phone evidence as evidence that might not follow the guide. It is also mentioned in ACPO's guide that an evidence collected without following the guide can be considered as a viable evidence.

However, Principle 1 of the ACPO guide can not be complied with when it comes to smart phone forensics. This is because smart phone storage is continually changing and that may happen automatically without interference from the mobile user. Thus, the goal with mobile phone acquisition should be to affect the contents of the storage of the mobile as less as possible and adhere to the second and third principles that focus more on the competence of the specialist and the generation of a detailed audit trail [8]. According to Principle 2, the specialist must be skilled enough to understand the internals of both hardware and software of the specific smart phone device they are dealing with and be proficient with the tools used to attain evidence from the device.

More than one tool is recommended to be used when acquiring evidence from mobile phone as some tools do not return error messages when they fail in a particular task [8]. Coming to Principle 3, When it comes to the recovery of digital Evidence, "The Guidelines for Best Practice in the Forensic Examination of Digital Technology" publication by the International Organization on Computer Evidence (IOCE) considers the following as the General Principles Applying to the Recovery of Digital Evidence [12]:

- The general rules of evidence should be applied to all digital evidence.
- Upon seizing digital evidence, actions taken should not change that evidence.
- When it is necessary for a person to access original digital evidence that person should be suitably trained for the purpose.

All activity concerning to the seizure, access, storage or transfer of digital evidence must be fully documented, conserved and accessible for evaluation. An individual is responsible for all actions taken with respect to digital evidence whilst the digital evidence is in their possession.

As with the ACPO principles, principle 2 cannot be strictly applied to evidence recovered from Smartphone devices because of their dynamic nature. Furthermore, mobile phone acquisition tools that claim to be forensically sound do not directly access the phone's memory but rather use commands provided by the phone's software and/or hardware interfaces for memory access and thus rely on the forensic soundness of such software or hardware access methods [10]. Hence, when using such tools for extracting information, the phone's memory may get modified unknowingly.

3.3 Mobile Phone Evidence Guides

There are a number of guides available, that concisely state potential evidence on a smart phone device. In this section, some of these guides are highlighted and their pitfalls are described.

The National Institute of Justice (NIJ), which is under the United States Department of Justice lists mobile phones under the heading of "Telephones" in their "Electronic Crime Scene Investigation: A guide for First Responders" publication [13].The details provided in these guides are not sufficient in describing an effective forensic approach for evaluating smart phones. These guides are not up to date and demand some serious modifications and extensions. Both guides though mention that mobile phones might have some potential evidence on them. The degree of the coverage is little and does not deal with smart phone storage capabilities and applications on them.

The USSS document also lists a set of rules on whether to turn on or off the device [12]:

- If the device is "ON", do NOT turn it "OFF".
- Turning it "OFF" could activate lockout feature.
- Write down all information on display (photograph if possible).
- Power down prior to transport (take any power supply cords present).
- If the device is "OFF", leave it "OFF".
- Turning it on could alter evidence on device (same as computers).

- Upon seizure get it to an expert as soon as possible or contact local service provider.
- If an expert is unavailable, USE A DIFFERENT TELEPHONE and contact 1-800-LAWBUST (a 24 x 7 service provided by the cellular telephone industry).
- Make every effort to locate any instruction manuals pertaining to the device.

On the other hand, the NIJ guide for first responders lists the following as potential evidence [13]: Appointment calendars/information., password, caller identification information, phone book, electronic serial number, text messages, e-mail, voice mail, memos, and web browsers.

The guide overlooked the possibilities that external storage device may be attached to a smart phone.

Both the guides fail to point out that smart phones may have electronic documents, handwriting information, or location information on them. The guides do not any significance of phone based applications such as Symbian, Mobile Linux and Windows Mobile applications. Both, Symbian and Windows Mobile based phones were found to execute malicious code such as Trojans and viruses especially ones transferred via Bluetooth technology. Non malicious applications on smart phones might be used to carry out criminal actions or can have log files or useful data and thus they could also be considered as evidence or source of evidence. Thus, every phone application and content associated to it should be regarded as a probable evidence including the Bluetooth logs, Infrared (IrDA) logs , Wi-Max and Wi-Fi communications logs and Internet related data such as instant messaging data and browser history data. Java applications should also be considered as evidence as many mobile phone operating systems support a version of Java [10].

The United Kingdom's Association of Chief Police Officers (ACPO) Good Practice Guide for Computer based Electronic Evidence lists the following instructions (CCIPS, 2002) to be followed while handling mobile phones for evaluation processes:

- Handling of mobile phones: Any interaction with the handset on a mobile phone could result in loss of evidence and it is important not to interrogate the handset or SIM.
- Before handling, decide if any other evidence is required from the phone (such as DNA/fingerprints/drugs/accelerants). If evidence in addition to electronic data is required, follow
- the general handling procedures for that evidence type laid out in the Scenes of Crime Handbook
- or contact the scenes of crime officer.
- General advice is to switch the handset OFF due to the potential for loss of data if the battery fails or new network traffic overwrites call logs or recoverable deleted areas (e.g. SMS); there is also potential for sabotage. However, investigating officers (OIC) may require the phone to remain on for monitoring purposes while live enquiries continue. If this is the case, ensure the unit is kept charged and not tampered with. In all events, power down the unit prior to transport.

The on/off rules here initially conflict with the USSS guide. Here again the guide is not up to date for it considers only SMSs, voicemail and address book/call history details as potential source of evidence from a smart phone device. N flow chart is provided for seizure process of a smart phone.

4. PROPOSED SMARTPHONE FORENSIC MODEL: Smartphone Forensic Investigation Process Model

Many digital forensic models have already been proposed by now. However the most appropriate one has not been figured out yet. The varying frameworks developed are such that they work well with one particular type of investigation. But none of them emphasize on the specific information flow associated with the forensic investigation of Windows mobile devices.

The Windows mobile device forensic process model has been developed in an attempt to overcome the major limitations of the existing digital forensic models. It helps forensic practitioners and law enforcement officials in the investigation of crimes emphasising a systematic and methodical approach for digital forensic investigation keeping in mind that the standard practices and techniques in the physical and digital investigation world are incorporated, wherever appropriate.

The proposed model consists of twelve stages, which are explained in the subsequent sections.

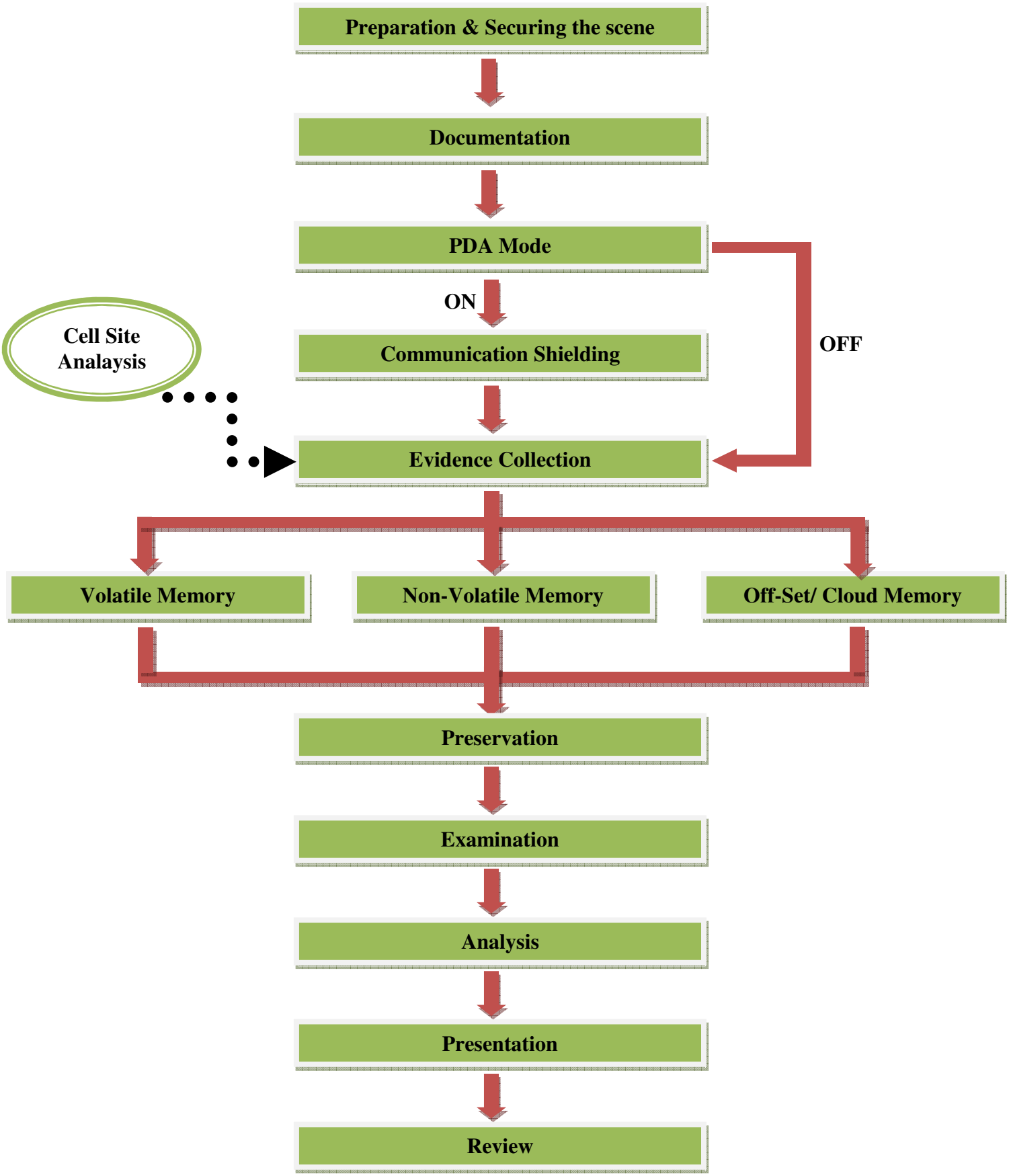

Figure 4: SFIPM

4.1 Phase One - Preparation

In order to enhance the quality of evidence and minimizes the risks associated with an investigation the preparation phase is planned out. This phase is associated with getting an initial understanding of the nature of the crime and activities. Being conducted prior to the occurrence of actual investigation, this phase involves preparation of the tools required for standard portable electronic device investigations, accumulating materials for packing evidence sources, building an appropriate team assigning roles to each personnel which may include case supervisor, crime scene sketch preparer, evidence recorder and so on, etc.

A critical assessment of the circumstances relating to the crime is carried out taking in consideration the knowledge of various mobile devices, accessories, features, specific issues etc. One more issue concerned with investigations involving Windows mobile devices is that the power runs out before evidence collection is over. So a toolkit consisting of standard power supplies, cables and cradles must be maintained properly.

A systematic strategy for investigation should be undertaken, keeping in mind the incident's nature and other technical, legal and business factors. While investigation the various legal constraints and jurisdictional as well as organizational restrictions should be ensured. Search warrants, support from the management, privacy rights of suspects, required authorizations and several other issues should not be overlooked during the process. A notification to all the concerned parties indicating the forensic investigation is also issued. Training, knowledge and experience of personnel are undoubtedly the prime contributors here.

4.2 Phase Two - Securing the Scene

Preventing the contamination and corruption of evidences and security of the crime scene from unauthorized access are the prime concerns of this stage. This is done protecting the integrity of all evidences and by maintenance of a formal protocol for ensuring systematic and secure custody at the crime spot. The evidences may get destroyed or destructed when the number of people at the crime scene increases. So Investigators are responsible for the control of the scene by defining the boundaries of the crime and controlling the gathered crowd over there. At the same time, safety of all the people at the scene must also be ensured.

It should be avoided to determine the contents in the devices and external storage devices at this stage. The devices must be left in their existing state until a proper assessment is made. If the device is on, it is better to leave it on. Similarly, if the device is off, never turn it on. No electronic device should be allowed to touch or tampered with.

4.3 Phase Three - Documenting the Scene

In order to maintain a proper chain of custody and circumstances surrounding the incident, documentation being a continuous activity is required in all stages. Things like the existing state on mobile phone when just spotted after the crime should be documented. A record of all visible data must which would help in recreating the crime scene any time during the investigation or say during a testimony in the court must be maintained. Photographs, sketches and crime-scene mapping all are merged together into a single documentation. The photographs may include device components such as power adaptors, cables, cradles and other accessories as already discussed earlier. It is necessary to keep a log of those who were present on the scene, those who reported afterwards, and those who left etc., along with the summary of their activities while they were at the scene. Classification of people into separate groups like victims, suspects, bystanders, witnesses and other assisting personnel etc. is carried out. Their location at the time of entry is recorded and documented.

4.4 Phase Four - PDA Mode

It is always advised that never to change the state of device it is working in. This phase decides the first course of action when device in hand depending upon the working of the device.

i) Active Mode: When the device is running/working, it is in Active mode or On mode. We would first need to shield it from external network and further communication without changing its mode so that the potential vulnerable volatile evidences remain intact. For this purpose device is first moved to Communication Shielding phase before working further.

ii) Inactive Mode: When the device is switched off, it is in Inactive or Off mode. Since we want to keep the evidences intact, it is not advised to turn the device on because this may lead to overriding of old data with new data. Thus we can continue with phase six and can skip communication shielding.

4.5 Phase Five – Communication Shielding

Occurring prior to the phase of evidence collection, Communication Shielding emphasizes to block the further communication options on the devices. This is done to ensure that no overwriting of the existing information on the devices is done. Even if the device appears to be in off state, some communication features like wireless or Bluetooth may be enabled.
The possibilities of overwriting and hence corruption of evidences may persist which should be avoided. Similarly, when the device is in the cradle connected to a computer and synchronization mechanisms using ActiveSync are enabled, remove any USB or serial cable, which connects it to the computer. The best option after seizing a device is to isolate it by disabling all its communication capabilities.

4.6 Phase Six – Volatile Evidence Collection

Since majority of the evidences involving mobile devices are volatile in nature, their timely collection and management is required. Volatile evidences are again prone to destruction as the device state and memory contents may change.

Depending upon the nature of evidences and the particular situation, evidences are either collected on the spot at the crime scene or they may be analyzed at the forensic laboratory afterwards. This decision may also depend upon the current power state. There may be a case of information loss if the device is running out of battery power. Hence, adequate power needs to be maintained if possible by using the power adaptor or replacing batteries. The device can also be switched off to preserve battery life and the contents of the memory. Alternatively, the contents of the memory can be imaged using appropriate commercial foreign tools like Paraben PDA Seizure which is used for memory acquisition. Several other open source forensic tools are also available which may be combined together to obtain better results.

4.7 Phase Seven – Non-volatile Evidence Collection

At this stage evidences are extracted from external storage devices like MMC cards, compact flash (CF) cards, memory sticks, secure digital (SD) cards, USB memory sticks etc. Along with this, evidences from computers and systems which are synchronized with these devices are also collected. Evidences of non-electric nature like written passwords, hardware and software manuals and related documents, computer printouts etc. are also looked for. Hashing and write protection of evidences is done to ensure their integrity and authenticity. Again forensic tools must be used in order to ensure the admissibility of evidences in the court of law. If the device has integrated phone features, the acquisition of sim card information takes place at this stage.

4.8 Phase Eight -- Off-Set

Until now there has been no bifurcation for the offset storage of data but the latest advancement in the field of cloud computing and other offset storage technologies has led to serious consideration of this phase for the search of potential evidence.

Smartphones are now equipped with cloud computing advantage to store their personal data online to cross mobile storage limits and access that data from anywhere anytime from any device. This could rise a possibility to hide the criminal evidence online which is not easy to track from device easily. Special consideration needs to be given to see what online data transactions have been made to have a track of activities done.

4.9 Phase Nine -- Cell Site Analysis

Cell site analysis is the science of being able to pinpoint a specific position, or positions where a mobile phone was or is. If a call is made from a mobile phone or a call is received from another phone to the mobile phone in question, or if an SMS is either sent or received then there will be records of this particular event.

Cell Site Analysis is associated with the science of locating the geographical area of the phone whenever calls are made, SMS or downloads are made or received, either in real time or historically. Such services are generally used by law enforcement agencies with the purpose of ensuring that a suspect was indeed present at on the spot and during the time when the crime was being held. The information provided is generally used for evidential purposes and is supported in courts by the expert witnesses. Using data from the networks and the skills of the network engineers, mobile phone signal strength readings are taken from various locations around the site in question to narrow down exactly where a mobile phone is being used.

4.10 Phase Ten – Preservation

To ensure the safety of evidences gathered their packaging, transportation and storage is carried out in this phase. Identification and their labeling are done before packaging. Plastic bags cause static electricity and hence may damage the evidences. Therefore, anti-static packaging envelopes are used for sealing the evidences like devices and other accessories.

Shocks, excessive pressures, humidity, temperature etc. may damage them during their transportation to the forensic workshop from the crime scene. Hence adequate precautions are necessary. Afterwards the device can be moved to a secure location where a proper chain of custody can be maintained and examination and processing of evidence can be started. Even after a safe transportation to the final destination, the packaged evidences may be prone to electromagnetic radiations, dust, heat and moisture. Unauthorized people should not have access to the storage area. National Institute of Standards and Technology guideline highlights the need of proper transportation and storage procedures, for maintaining a proper chain of custody. Proper documentation is done to avoid their altering and destruction.

4.11 Phase Eleven – Examination

To resolve and sort out the case, critical examination of the evidences collected and their analysis is carried out by the forensic specialists. Data filtering, validation, pattern matching and searching for particular keywords with regard to the nature of the crime or suspicious incident, recovering relevant ASCII as well as non- ASCII data etc. are some of the major steps performed during this phase. Personal organizer information data like address book, appointments, calendar, scheduler etc, text messages, voice messages, documents and emails are some of the common sources of evidence, which are to be examined in detail. Finding evidence for system tampering, data hiding or deleting utilities, unauthorized system modifications etc. should also be performed. Detecting and recovering hidden or obscured information is a major tedious task involved.

Significance of evidences is analyzed keeping in mind their originality is maintained. Appropriate number of evidence back-ups must be created before proceeding to examination. Huge volumes of data collected during the volatile and non-volatile collection phases are filtered and split into manageable chunks and form for future analysis. Data filtering, validation, pattern matching and searching for particular keywords with regard to the nature of the crime or suspicious incident, recovering relevant ASCII as well as non- ASCII data etc. are some of the major steps performed during this phase. A critical search and examination for decoding passwords and finding unusual hidden files or directories, file extension and signature mismatches etc. is carried out. The expertise of the investigator and capabilities of forensic tools used by the examiner also plays a major contribution for the efficient examination of evidences. When the evidence is checked-out for examination and checked-in, the date, time, name of investigator and other details must be documented. It is required to prove that the evidence has not been altered after being possessed

by the forensic specialist and hence hashing techniques like md5 must be used for mathematical authentication of data.

4.12 Phase Twelve – Analysis

Identifying relationships between fragments of data, analyzing hidden data, determining the significance of the information obtained from the examination phase, reconstructing the event data, based on the extracted data and arriving at proper conclusions etc. are some of the activities to be performed at this stage. This stage constitutes the technical review of the investigators on the basis of the results of the previous examination stage of the evidence. The analysis of whole situation at the crime scene should be such that the chain of evidences and timeline of events is consistent. Additional steps in the extraction and analysis process are analyzed and properly documented. Using a combination of tools for analysis will yield better results. The National Institute of Justice (2004) guidelines recommend timeframe analysis, hidden data analysis, application analysis and file analysis of the extracted data.

4.13 Phase Thirteen - Presentation

After the whole analysis of the results presentation of results to the wide variety of audience including law enforcement officials, technical experts, legal experts, corporate management etc. is done. This actually depends on the nature of the crime. The findings must be presented in a court of law, if it is a police investigation or before appropriate corporate management, if it is an internal company investigation. Allegations regarding the crime are discarded or confirmed during this stage. The results of examination and analysis are reviewed in their entirety to get a complete picture. This is because the individual results of each of the previous phases may not be sufficient to arrive at a proper conclusion about the crime.

A report consisting of a detailed summary of the various events that took place during the crime and the complete description of the steps in the process of investigation and the conclusions reached is documented and provided. Along with the report, supporting materials like copies of digital evidence, devices spotted at the crime scene, a chain of custody documents, printouts and photographs of various items of evidence etc. should also be submitted. The complex terms involved in various stages of investigation process and the expertise and knowledge of the forensic examiner, the methodology adopted, tools and techniques used etc. are all likely to be challenged before a jury and needs to be explained in layman's terminology.

4.14 Phase Fourteen - Review

A complete review of all the steps during the investigation and identification of the areas of improvement are included in this final Review stage of the Windows Mobile Forensic Process Model. Results and their interpretations may be used in future for further refining the gathering, examination and analysis of evidence in future investigations. In many cases, much iteration of examination and analysis phases are required to get the total picture of an incident or crime. Better policies and procedures are established in place in future by means of this information.

Smartphone Forensic Investigation Process Model	**NIJ Law Enforcement Model**	**DFRWS Model**	**Abstract Digital Forensic Model**	**IDIP Model**	**Systematic Digital Forensic Investigation Model**
Preparation			✓	✓	✓
Securing the scene		✓		✓	✓
Survey and Recognition		✓	✓	✓	✓
Documenting the scene				✓	✓
Mode Selection/ Shielding					

Volatile Evidence Collection					✓
Non-volatile Evidence Collection	✓	✓	✓	✓	✓
Off-Set/ Online Storage					
Cell Site Analysis					
Preservation		✓	✓	✓	✓
Examination	✓	✓	✓	✓	✓
Analysis	✓	✓	✓		✓
Presentation	✓	✓	✓	✓	✓
Review				✓	✓

Table 2: Comparison of major forensic models with Smartphone Forensic Investigation process model

As it is clear from the above comparison table that not only our model is accommodating all the necessary steps but it is also including needed processes to be added to walk with the advancement in technology and look for the more efficient evidence sources. It facilitates mode selection/shielding, off-set/online storage and cell site analysis which were otherwise not supported in the rest of the models making it more effective and versatile for evidence management.

5. FUTURE CHALLENGES IN MOBILE FORENSICS

The mobile industry is moving with such a fast pace, it's often hard to keep up with it. There is a large number of future trends going to be seen with Smartphones around us. With every major mobile phone release, users are treated to an ever-expanding list of advanced features. Some are more useful than others, but they represent an industry that is always on the move.

We tried to roundup some of the best. All of these developments may have an impact on mobile device forensics.

5.1 Processor optimization

Mobile phones today are easily available with a processor speeds ranging from 300 Mhz to 600 Mhz, and even to latest Smartphones providing upto 1 to 1.3 Ghz.

Ed Hansberry stated in his article ' The Value Of Multi-Core Processors On Phones ' that Smartphones can take the benefits of symmetric multiprocessing (SMP) as well. The reason Apple has given for not allowing third party multitasking is power consumption.

Qualcomm has announced (2011) their multi-core Snapdragon line of processors, but they aren't the only one in the mobile SMP game. Most Nokia Smartphones, the Palm Pre, Motorola Droid and hundreds of other phones have an OMAP chip inside, likely an OMAP 3. Their new OMAP 4 line is based on the ARM Cortex A9 architecture.

These dual core systems do more than make things go faster, as they even overcome the challenges for mobile processors:

- Power consumption
- Digital Signal Processing
- Peripherals Integration
- Multimedia Acceleration
- Code Density

Such change in processor architecture will make an undesirable impact on Smartphone forensics.

5.2 Battery life

Power source/batter life is a major concerned these days as it was few years back before the popularity of Smartphones. Mobile phones Mobile phones typically use NiMH (nickel metal hydride), Li-ion (lithium-ion),or Li-polymer batteries. At a stage in mobile phone development these batteries were very good at their performance – look at best selling java phones like the Nokia 1100,3315,6600, Sony Ericsson T-600 etc had a battery life of almost 140 hours of standby time – nearly two weeks. But they are no so optimized enough to give this much support ot these days Smartphones simply because they require high amount of computation and continuous numerous activities running on them like GPS, wi-fi etc.

Peter Bruce, a professor of chemistry at the University of St Andrews is taking up that challenge with his "Air-Fuelled" rechargeable lithium battery. Put very simply, the Stair cell (St Andrews air cell) uses nothing more complicated than air as a reagent in a battery instead of costly chemicals. By freeing up space and exploiting one of the few elements that is free, [14] can squeeze more power into a smaller space at a reduced cost. "By using air in the cell we can get much higher energy storage up to a factor of 10.

As volatile data can be lost if the device gets turned off thus Battery life makes a huge impact on a mobile forensic investigation.

5.3 Storage memory

Smartphone's OS and applications are installed in RAM,ROM or flash memories because of the smaller OS and application as those of computers. These days latest Smartphones are available with up to 1GB of RAM to store application code and up to 64GB of internal (flash) memory for system code and user data.

Nearly every mobile phone these days also support external storage like micro SD cards varying from small storage capacity to up to large capacity of 64gb for high end Smartphones. Devices today even allow swapping in and out of external storage devices without turning off the device.
The storage medium used and the file system used by the OS to store data on it stands as a major evidence for Smartphone forensics.

5.4 Advance imaging

Smartphones are not just smart in business, they are even leading in the race of entertainment. Every Smartphone leading brand has now believed that people don't want to carry and an extra camera or camcorder to take pictures and videos thus every Smartphone releasing today is equipped with better camera technology, high pixel sensor and quality optics for advanced high quality pictures and high definition (1080p) videos.

Imaging is just not kept till photography rather advanced imagine capabilities with new mobile applications also allow to take 360 degree view of a place and make a whole map, guide, blue print of the place which can even be used for criminal offense. Thus advance imaging also stands as a source of evidence in mobile forensics and require high end image steganalysis.

5.5 Cloud computing

cloud computing is just not limited to computers in fact Smartphone leaders are trying hard to incorporate cloud with Smartphones which is going to revolutionaries the flexibility and mobile computing abilities of Smartphones.

Cloud computing rips off all the barriers which were there on the computational power of Smartphones by flexibility of the devices getting their work done remotely without having the suite installed on the device itself. It will also remove all the brand-based constraints which would be a

firm benefit with cloud implementation to any device and any application. Its de-centralized storing of documents, photographs and other data also enable users to work seamlessly with colleagues and even share devices without loss of data.

But we can deny that cloud computing in Smartphones also increases the risk of criminal activities being carried out in a more planned and larger scale because of seamless sharing and group working of people which could lead to large terrorist activities too.

5.6 4G and beyond

After 3G, the arrival of 4G is threatening to occur immediately and with it comes a new array of functionality along with higher specification hardware and improved network infrastructure which is going to enhance the speed of our mobile lives. It also going to provide fast and stable data connection.

This rapid change in technology and its extension is a big hurdle in Smartphone forensics. With new technology comes the requirement of newer way of Smartphone forensics.

6. CONCLUSION

Motivated by the rapid increase in mobile frauds and cyber crimes, this research work took tried to put forward the need and way of Smartphone forensics. This paper starts with the discussion on the increasing need of smartphone forensic then how is it different from computer or other digital forensics and then moving on to potential evidences and strategies defined earlier.
The proposed Smartphone Forensic Investigation Process Model (SPFIPM) benefits as follows:
• Serve as benchmark and reference points for investigating Smartphones for criminal cases.

• Provide a generalized solution to the rapidly changing and highly vulnerable digital technological scenario.

7. REFERENCES

1. Nielsen. “Mobile Texting Status.” Internet: http://mashable.com/2010/10/14/nielsen-texting-stats/, Oct. 14, 2010 [Jan. 13, 2012]

2. D. Paul. “the year of mobile customers.” Internet: http://www.themda.org/documents/PressReleases/General/_MDA_future_of_mobile_press_release_Nov07.pdf, Nov. 07, 2011 [Jan. 15, 2012]

3. FoneKey. Internet: www.FoneKey.net, 2008 [Dec. 12, 2011]

4. Ducell. Internet: www.DuCell.org, 2008 [Dec.15, 2011]

5. D, Mock. “Wireless Advances the Criminal Enterprise.” Internet: http://www.thefeaturearchives.com/topic/Technology/Wireless_Advances_the_Criminal_Enterprise.html, Jun. 18, 2008 [Jan. 17, 2012]

6. R. Ayers, W. Jansen, N. Cilleros & R. Daniellou. “Cell Phone Forensic tools: An Overview and Analysis.” Internet: http://csrc.nist.gov/publications/nistir/nistir-7250.pdf, 2007 [Jan. 21, 2012]

7. B. D. Carrier. “Risks of Live Digital Forensic Analysis.” Communications of the ACM, 49(2), 56-61.

8. W. Jansen & R. Ayers “Guidelines on Cell Phone Forensics” Internet: http://csrc.nist.gov/publications/nistpubs/800-101/SP800-101.pdf, 2006 [Feb. 03, 2012]

9. P. Zheng & L. M. Ni. “The Rise of the Smart Phone.” IEEE Distributed Systems Online, 7(3), art. no. 0603-o3003.

10. P. McCarthy. “Forensic Analysis of Mobile Phones.” Unpublished Bachelor of Computer and Information Science (Honours) Degree, University of South Australia, Adelaide.

11. ACPO. “Good Practice Guide for Computer based Electronic Evidence.” Internet: http://www.acpo.police.uk/asp/policies/Data/gpg_computer_based_evidence_v3.pdf, 2003 [Dec. 22, 2012]

12. IOCE. “Best Practice Guidelines for Examination of Digital Evidence.” Internet: http://www.ioce.org/2002/Guidelines%20for%20Best%20Practices%20in%20Examination%20of%20Digital%20Evid.pdf, 2007 [Feb. 09, 2012]

13. NIJ. “Electronic Crime Scene Investigation: A Guide for First Responders.” Internet: http://www.ncjrs.gov/pdffiles1/nij/187736.pdf, 2003 [Feb. 11, 2012]

14. U. Simon. “Batteries: The power behind the phone” Internet: http://www.independent.co.uk/life-style/gadgets-and-tech/features/batteries-the-power-behind-the-phone-1872933.html, Jan. 2010 [Feb. 18,2012]

15. A. Ankit, G. Megha, G. Saurabh & C. Gupta. “Systematic digital forensic investigation model” Vol. 5 Internet: http://www.cscjournals.org/csc/manuscript/Journals/IJCSS/volume5/Issue1/IJCSS-438.pdf, 2011 [Jan. 15, 2012]

Analysis of N Category Privacy Models

Marn-Ling Shing *sshing@tmue.edu.tw*
Early Child Education Department and Institute of Child Development
Taipei Municipal University of Education
Taipei, Taiwan, R.O.C y

Chen-Chi Shing *cshing@radford.edu*
Information Technology Department
Radford University
Radford, VA 24142, U.S.A.

Lee-Pin Shing *shingle@vt.edu*
Biology Department
Virginia Tech
Blacksburg, VA 24061, U.S.A.

Lee-Hur Shing *leehurshing@yahoo.com*
Engineering Department
Virginia Tech
Blacksburg, VA 24061, U.S.A.

Abstract

File sharing becomes popular in social networking and the disclosure of private information without user's consent can be found easily. Password management becomes increasingly necessary for maintaining privacy policy. Monitoring of violations of a privacy policy is needed to support the confidentiality of information security. This paper extends the analysis of two category confidentiality model to N categories, and illustrates how to use it to monitor the security state transitions in the information security privacy modeling.

Keywords: Privacy Model, Confidentiality Model, Information Security Model, and Markov Chain Model.

1. INTRODUCTION

Information assurance includes "measures that protect information and information systems by ensuring their availability, integrity, authentication, confidentiality, and non-repudiation. These measures include providing for restoration of information systems by incorporating protection, detection, and reaction capabilities" [1]. In business the job of keeping the company's infrastructure and network safe is growing increasingly complex as the perimeter expands, threats become more sophisticated, and systems become more complex and embedded. Companies are becoming increasingly more dependent on technology for the liability of security and privacy as required by the legislations in state and federal laws such as HIPAA, Sarbanes-Oxley, and California's Security Breach Information Act 1386 [2]. Information security is more than just using a good Intrusion Detection System or firewall; it involves keeping users educated, creating and maintaining good policies, getting the right budget, and sometimes monitoring user activities. In academia maintaining a strict privacy and security of student and employee records are required. There are state and federal laws that protect records containing information directly identifying, or revealing private information for students and employees. For example, they are the Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA), and the Gramm-Leach-Bliley Act (GLBA),. There are also a variety of security technologies and procedures used to help protect students' and employees' information from unauthorized access, use, or disclosure. When highly private information (such as a credit

card number or password) are transmitted over the Internet, usually they are protected through the use of encryption, such as the Secure Socket Layer (SSL) protocol. Password management on different servers to maintain privacy policies are necessary. Constant monitoring is needed to prevent the damage as a result of any violation of privacy policy.

Since security becomes an important component in the various services, new standards are emerging for these services. For example, the new auditing standards No. 99 provides a general guideline for the responsibilities and anti-fraud activities of a manager [3]. ISO 17799:2005 provides a general organization security structure [4]. These security standards may be influenced by existing security models. For example, ISO 17799:2005 provides a general organization security structure but many basic security principles may have been discussed and defined in various security models. While supply chain information is similar to any other information systems, it has unique features on confidential and integrity. In order to monitor the security in a supply chain network, it is necessary to model the security state transition in the Bell-LaPadula model [5]. A two-category model has been proposed [6]. This paper intends to generalize from the two-category model to N-category and provides details on analyzing the model and illustrates how to use it to monitor the security state transitions in privacy model.

2. LITERATURE REVIEW

2.1 Information Security Models

There are various models which provide policies from different aspects of security. The Bell-LaPadula model is one of the security models for information confidentiality and has been adopted by the military for a long time. For example, the Bibba model provides security policy in data integrity [7, 8]. On the other hand, Bell-LaPadula model provides security policy to guard against unauthorized disclosure [9]. Bell LaPadula model has been used in the military and is primarily designed for modeling confidentiality [10-11]. It classifies the access levels for the subject into a set of security clearances, such as: top security (TS), security (S), confidential (C), and unclassified (UC). In the mean time the objects have also been classified as corresponding security levels. It does not allow a subject to read the objects at security levels higher than the subject's current level. Every subject must belong to one and only one of the security clearance levels. In addition, every object must also belong to one and only one of the classification levels. For example, a colonel, who is in the TS security clearance, can read the Personnel files. Whereas, a soldier, who is in the UC security clearance, can read the telephone lists. The colonel can also read the telephone lists; however, the soldier cannot read the personnel files.

2.2 The Bell-LaPadula Model for Supply Chain Networks

In a supply chain network, prices offered by suppliers are often confidential due to competition and also are not public information in the buyer's company. The confidentiality of the supplier information is essential in nowadays competitive business world. Chen et al. [12, 13] proposed to use the Bell-LaPadula model for the supply chain network. In order to investigate the security in the supply chain model, it is necessary to be able to model the security state transition in the Bell-LaPadula model. Shing et al. proposed using Markov chain model [6, 14, 15] in a two-category. According to the Bell-LaPadula model, we can classify the employees (or subjects) in the purchasing company into several security clearance levels and different information (or objects) into different security classification levels. For simplicity, assuming that there are two security clearance levels for all employees in a purchasing company (see Figure 1). They are the top officer and other employees. The top officer can access or read two documents: both supplier evaluations and purchasing decision. On the other hand, other employees can only access (read) two documents: the public bidding notices and the public purchasing price list. The top officer can also access the documents which a general employee can access. Other employees cannot access documents for both supplier evaluations and purchasing decision. Table 1 shows security classifications and clearance levels for a purchasing company and its suppliers.

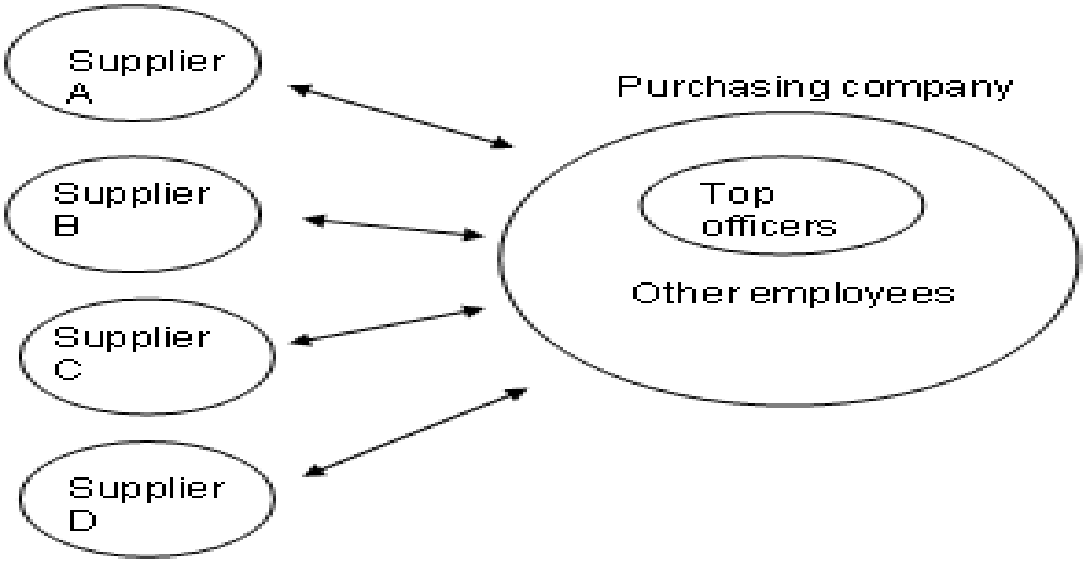

FIGURE 1: Purchasing Company and Their Suppliers.

Security Classification	Purchasing Co. and Suppliers	Documents/ Information
Top Secret (TS)	Managers	Supplier evaluations
Secret (S)	Other employees	Public bidding notices

TABLE 1: Security Classifications in a Supply Chain Network.

The abstract model of the table 1 can be represented as

Subject Security Clearance	Object Classification Level
S_1	O_{11}, O_{12}
S_2	O_{21}, O_{22}

TABLE 2: Security Abstract Classifications in a Supply Chain Network.

3. SEMI-MARKOV CHAIN MODEL

A Markov process is a stochastic process which states that the probability of a system at a state depends only on the previous state, not on the previous history of getting to the previous state [16]. If the states and their transitions at discrete points in time are discrete, it is called a Markov Chain [17-19]. Suppose p(0) represents the vector of the probability that the system is in one of those n states at time 0,

$p(0)=\begin{bmatrix} p_1(0) \\ p_2(0) \\ \dots \\ p_n(0) \end{bmatrix}$, $\sum_{i=1}^{n} p_i(0) = 1$, where $p_i(0)$ represents the probability of the system is in state i at time 0. Then the probability that the system is in one of those n states at time 1 is represented by p(1),

$p(1)=\begin{bmatrix} p_1(1) \\ p_2(1) \\ \dots \\ p_n(1) \end{bmatrix}$, $\sum_{i=1}^{n} p_i(1) = 1$, where $p_i(1)$ represents the probability of the system is in state i at time 1.

And P(1)=T p(0), where T is the transition probability matrix,

$$T=\begin{bmatrix} p_{11} & p_{21} & \dots & p_{n1} \\ p_{12} & p_{22} & \dots & p_{n2} \\ \dots & \dots & \dots & \dots \\ p_{1n} & p_{2n} & \dots & p_{nn} \end{bmatrix}, \quad \sum_{j=1}^{n} p_{ij} = 1, \text{ for } i=1,2,\dots,n \qquad \text{(Eq 3.1)}$$

and p_{ij} is the probability of the system in the state j, given it was in the state i. Suppose the probability that the system is in one of those n states at time s is represented by p(s),

$p(s)=\begin{bmatrix} p_1(s) \\ p_2(s) \\ \dots \\ p_n(s) \end{bmatrix}$

where $p_i(s)$ represents the probability of the system is in state i at time s. Then $P(s)=T(T(\dots(Tp(0))))=T^s p(0)$,where T is the transition probability matrix. A Markov Chain is a special case of a random walk process [20, 21] , which is defined as “a random variable X_n that has values in set of integers Z”, with

$P(X_n = X_{n-1}+1)=p$ and $P(X_n = X_{n-1}-1)=1-p$, where $p \in (0,1)$.

In general, a random walk process is a semi-Markov Chain. Every entry of the transition probability matrix in a semi-Markov Chain can be arbitrary [19], as in the Definition 3.1 below:

Definition 3.1
A semi-Markov process is a stochastic process that has an arbitrary distribution between state changes and any new state is possible given it is in the current state.

The Markov Chain model will be extended to the semi-Markov Chain using two category confidential model in Section 4.

4. ANALYSIS OF ABSTRACT SEMI-MARKOV CHAIN MODEL

The results in this section were for two category model and proved in [22]. They are listed here for completeness.

Definition. 4.1 A state is recurrent if a state will return back to itself with probability one after state transitions. If the state is not recurrent, then it is a transient. If a recurrent state is called recurrent nonnull if the mean time to return to itself is finite. A recurrent state is a recurrent null if the mean time return to itself is infinite. A recurrent state is aperiodic if for some number k, there is a way to return back in k, k+1, k+2, … transitions. A recurrent state is called periodic if it is not aperiodic.

Definition 4.2 A semi-Markov chain is irreducible if all states are reachable from all other states. It is recurrent nonnull if all its states are recurrent nonnull. It is aperiodic if all its states are aperiodic.

Definition 4.3 If a semi-Markov chain is irreducible, recurrent nonnull and aperiodic, it is called ergodic.

The eight states in the semi-Markov chain model is presented in Table 3 for abstract model in Table 2.

State	Object Classification
1	$(S_1, O_{11}), (S_2, O_{21})$
2	$(S_1, O_{12}), (S_2, O_{21})$
3	$(S_1, O_{21}), (S_2, O_{21})$
4	$(S_1, O_{22}), (S_2, O_{21})$
5	$(S_1, O_{11}), (S_2, O_{22})$
6	$(S_1, O_{12}), (S_2, O_{22})$
7	$(S_1, O_{21}), (S_2, O_{22})$
8	$(S_1, O_{22}), (S_2, O_{22})$

TABLE 3: Semi-Markov Chain States for Table 2.

Properties 4.1
P(System is at the state (S_1, O_{ij}) at time t | System is at the state (S_2, O_{2m}) at time t-1)=P(System is at the state (S_1, O_{ij}) at time t) , where $i, j=1, 2, 3, 4, m=1,2$.

Properties 4.2
P(System is at the state (S_1, O_{ij}) (S_2, O_{2m}) at time t)=P(System is at the state (S_1, O_{ij}) at time t) * P(System is at the state (S_2, O_{2m}) at time t), where $i, j, m=1,2$.

Properties 4.3
P(System is at the state (S_2, O_{2j}) at time t | System is at the state (S_1, O_{im}) at time t-1)=P(System is at the state (S_2, O_{2j}) at time t) , where $i, j=1, 2, m=1, 2, 3, 4$.

Using the following notation:
P(System is at the state (S_1, O_{11}) at time 0) $=a_{11}$
P (System is at the state (S_1, O_{12}) at time 0)$=a_{12}$
P (System is at the state (S_1, O_{21}) at time 0) $=a_{13}$
P (System is at the state (S_1, O_{22}) at time 0) $=a_{14}$
P (System is at the state (S_2, O_{21}) at time 0) $=b_{11}$
P (System is at the state (S_2, O_{22}) at time 0)$=b_{12}$, where $\sum_{j=1}^{4} a_{ij} = 1$ and $\sum_{j=1}^{2} b_{ij} = 1$, the initial state of the system is given by

Property 4.4
P(System is at the state m at time 0) = $a_{1i}\, b_{1j,}$
where $i=1, 2, 3, 4, j=1, 2,$ and $m=i+4(j-1)$.

In calculating the state transition probability, the following trivial properties are needed:
Property 4.5
P(System is at state (S_1, O_{1j}) (S_2, O_{2m}) at time t | System was at state (S_1, O_{1n}) (S_2, O_{2m}) at time t-1) = P(System is at state (S_1, O_{1j}) at time t | System was at state (S_1, O_{1n}) at time t-1) * P(System is at state (S_2, O_{2m}) at time t | System was at state (S_2, O_{12m}) at time t-1),
where *j, n =1, 2, 3, 4,* and *m=1, 2*.

Property 4.6
P(System is at state (S_1, O_{ij}) (S_2, O_{2m}) at time t | System was at state (S_1, O_{ij}) (S_2, O_{2n}) at time t-1) = P(System is at state (S_1, O_{ij}) at time t | System was at state (S_1, O_{ij}) at time t-1) * P(System is at state (S_2, O_{2m}) at time t | System was at state (S_2, O_{2n}) at time t-1),
where *i=1, 2, j =1, 2, 3, 4,* and *m, n=1, 2*.

Property 4.7
The transition probability
$p_{ij} = q_{ij} * r_{11}$
$p_{i\,(j+4)} = q_{ij} * r_{12}$
$p_{(i+4)\,j} = q_{ij} * r_{21}$
$p_{(i+4)\,(j+4)} = q_{ij} * r_{22}$
where *i=1, 2, j =1, 2, 3, 4*.

Property 4.8
If a semi-Markov chain is ergodic, then there exists a unique steady-state or equilibrium probability state.

Depending on the structure of the transition probability matrix, it may not have any steady state exists. For example, a symmetric random walk process which has p=0.5, is periodic. [9].

Property 4.9
For any semi-Markov chain if all the entries of its transition probability matrix are non-zero, then it is recurrent nonnull and aperiodic.

Corollary 4.9
For any semi-Markov chain, if every entry of its transition probability matrix has non-zero in all the entries, then it has an equilibrium state.

Corollary 4.10
If T keeps no change in Corollary 4.9, then the equilibrium state described in Corollary 4.9 is the eigenvector of T', the transpose of the transition probability matrix, of eigenvalues 1.

The next section will generalize the model to n categories and describe some of the properties

5. GENERAL MODEL

The most general abstract model of the Table 1 can be represented as in Table 4 below.

Subject Security Clearance	Object Classification Level
S_1	$O_{11}, O_{12}, \ldots, O_1 v_1$
S_2	$O_{21}, O_{22}, \ldots, O_2 v_2$
…	…
S_n	$O_{n1}, O_{n2}, \ldots, O_n v_n$

TABLE 4: Security General Abstract Classifications.

Definition 5.1.
The abstract model given in Table 4 is called an n category confidential model.

Since the subject S_i can access the objects O_{ij} where $j \geq i$, the states in semi-Markov Chain n category confidential model for Table 4 are listed in Table 5 below.

State	Object Classification
1	$(S_1, O_{11}), (S_2, O_{21}), (S_n, O_{n1})$
2	$(S_1, O_{12}), (S_2, O_{21}), (S_n, O_{n1})$
…	…
$\sum_{i=1}^{n} v_i$	$(S_1, O_n\ v_n), (S_2, O_{21}), (S_n, O_{n1})$
…	…
$\prod_{i=1}^{n} \sum_{j=i}^{n} v_j$	$(S_1, O_n v_n), (S_2, O_n v_n), (S_n, O_n v_n)$

TABLE 5: Semi-Markov Chain States for Table 4.

The total number of states in semi-Markov Chain grows exponentially fast as the total number of subjects n as in the following property:

Property 5.1.
Assume that the total number of subjects is n. The total number of states in the semi-Markov Chain model is $O(n!v^n)$, where $v=max(v_1, \ldots, v_n)$.
Proof:

The total number of states in semi-Markov Chain model is $\prod_{i=1}^{n} \sum_{j=i}^{n} v_j$, where $\sum_{j=i}^{n} v_j$ = n. Hence,

$\prod_{i=1}^{n} \sum_{j=i}^{n} v_j \leq nv(n-1)v \ldots v = n!\, v^n$.

For example, if there are three categories and at most two objects in each category (i.e., n=3 and v=2), by Property 5.1, the total number of states in the model is no more than 6x8=48 states. Similar properties as Properties 4.1-4.7 are still valid for three category model.

In order to calculate the transition probability matrix for the n category confidential model, we need to use some notations and define a binary operator $\triangleright$ on matrices.

Definition 5.2.

Let the matrix $Q_k = (q_{(k),ij})$ of size $\sum_{i=k}^{n} v_i$ x $\sum_{i=k}^{n} v_i$, where k=1,..,n and $q_{(k),ij}$ is the probability from state *i* at time 0 to state *j* at time 1 for subject S_k. And let n_k = number f rows/columns of Q_k. Define

$$Q_r \triangleright Q_s = \begin{bmatrix} q_{(s),11}Qr & q_{(s),12}Qr & . & . & . & . & q_{(s),1,\sum_{i=s}^{n} v_i}Qr \\ q_{(s),21}Qr & q_{(s),22}Qr & . & . & . & . & q_{(s),2,\sum_{i=s}^{n} v_i}Qr \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ q_{(s),\sum_{i=s}^{n} v_i,1}Qr & q_{(s),\sum_{i=s}^{n} v_i,2}Qr & . & . & . & . & q_{(s),\sum_{i=s}^{n} v_i,\sum_{i=s}^{n} v_i}Qr \end{bmatrix}$$

where every entry of the matrix
$q_{(s),ij}Qr$ is obtained by multiplying the entry of the matrix Q_r by $q_{(s),ij}$.
Note that the sum of each row of the matrix Q_i, where i=1,…, n is equal to one. The transition probability matrix T can be calculated using the operator in the following:

Property 5.2.
The transition probability matrix of the n category confidential model is given by
$T = (\ldots(Q_1 \triangleright Q_2) \triangleright Q_3)\ldots \triangleright Q_n)$.

Proof: The proof is similar to the one given by Property 4.7.

Property 5.3.
Every sum of each row in the transition probability matrix T is equal to 1.

Proof: The proof is similar to the one given by Property 4.8.

A result similar to Property 4.4 for n category confidential model can be obtained in Property 5.4.

Property 5.4.
The initial state in the n category confidential model can be calculated by multiplying all initial probability entries of every category in the following:

$$P(0) = (p^{(0)}_i) \prod_{m=1}^{n} n_{mx1},$$

where

$$p^{(0)}_i = \prod_{k=1}^{n} q_k^{(0)},$$

$q_k^{(0)}$ is the probability matrix at state (S_k, O_{k1}), (S_k, O_{k2}),…, $(S_k, O_k v_k)$, …, (S_k, O_{n1}), (S_k, O_{n2}),…, $(S_k, O_k v_n)$ at time 0.

Proof: Similar to that of Property 4.4.

For a two category confidential model described in Section 4, n=2, v_1=2, and v_2=2. Total number of states = $(v_1 + v_2)v_2$=8.

$$Q_1 = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{bmatrix}, \qquad \sum_{j=1}^{4} q_{ij} = 1, \text{ for } i=1,2,3,4$$

$$Q_2 = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

$$T = \begin{bmatrix} q_{11}*r_{11} & q_{12}*r_{11} & q_{13}*r_{11} & q_{14}*r_{11} & q_{11}*r_{12} & q_{12}*r_{12} & q_{13}*r_{12} & q_{14}*r_{12} \\ q_{21}*r_{11} & q_{22}*r_{11} & q_{23}*r_{11} & q_{24}*r_{11} & q_{21}*r_{12} & q_{22}*r_{12} & q_{23}*r_{12} & q_{24}*r_{12} \\ q_{31}*r_{11} & q_{32}*r_{11} & q_{33}*r_{11} & q_{34}*r_{11} & q_{31}*r_{12} & q_{32}*r_{12} & q_{33}*r_{12} & q_{34}*r_{12} \\ q_{41}*r_{11} & q_{42}*r_{11} & q_{43}*r_{11} & q_{44}*r_{11} & q_{41}*r_{12} & q_{42}*r_{12} & q_{43}*r_{12} & q_{44}*r_{12} \\ q_{11}*r_{21} & q_{12}*r_{21} & q_{13}*r_{21} & q_{14}*r_{21} & q_{11}*r_{22} & q_{12}*r_{22} & q_{13}*r_{22} & q_{14}*r_{22} \\ q_{21}*r_{21} & q_{22}*r_{21} & q_{23}*r_{21} & q_{24}*r_{21} & q_{21}*r_{22} & q_{22}*r_{22} & q_{23}*r_{22} & q_{24}*r_{22} \\ q_{31}*r_{21} & q_{32}*r_{21} & q_{33}*r_{21} & q_{34}*r_{21} & q_{31}*r_{22} & q_{32}*r_{22} & q_{33}*r_{22} & q_{34}*r_{22} \\ q_{41}*r_{21} & q_{42}*r_{21} & q_{43}*r_{21} & q_{44}*r_{21} & q_{41}*r_{22} & q_{42}*r_{22} & q_{43}*r_{22} & q_{44}*r_{22} \end{bmatrix}$$

This is confirmed by Property 4.7 and 4.8. The example of calculating transition probability for states in Table 1 can be found in [6].
The transition probability matrix can be obtained from Property 5.2 in the following:

Corollary 5.1.
For a three category confidential model of two states each, n=3, ν_1 =2, ν_2 =2 and ν_3 =2. In addition, n_1 =6, n_2 =4 and n_3 =2. Total number of states = $(\nu_1+\nu_2+\nu_3)(\nu_2+\nu_3)\nu_3$ =48. Assume $Q_1 = (q_{ij})_{6x6}$, $Q_2 = (r_{ij})_{4x4}$ and $Q_3 = (w_{ij})_{2x2}$. Then $T=(p_{ij})_{48x48}$, where m=1,2, t=1,2, k=1,2,3,4, s=1,2,3,4, i=1,..,6, j=1,…,6 and
$p_{i+6s+24t,j+6k+24m} = q_{ij} * r_{s,k} * w_{t,m}$

Proof: A block of 6 rows/columns for changing indices of $r_{s,k}$ and a block of 24 rows/columns for changing indices of $w_{t,m}$.

A similar result to express the entry of the transition probability matrix explicitly for n category confidential model is in the following:

Corollary 5.2.
For the n category confidential model, assume $Q_k = (q_{(k),ij})$, where i, j = 1, .., n_k , k=1,…, n.
Then $T=(p_{ij})$, with

$$p_{i+\sum_{s=1}^{n-1} i_{s+1} \prod_{m=1}^{s} n_{m,\,j+\sum_{s=1}^{n-1} j_{s+1} \prod_{m=1}^{s} n_m} = \prod_{k=1}^{n} q^{(k)}}$$

where i, j = 1, …, n_1,
$q^{(k)} = q_{(k),\, i_k\,,\, j_k}$
and
$i_k\,,\ j_k = 1, \ldots, n_k$
k = 1, 2, …, n.

Proof:
The proof is similar to the one in Corollary 5.1.

For example, if n=3, ν_1 =7, ν_2 =3 and ν_3 =2. In addition, n_1 =12, n_2 =5 and n_3 =2. Total number of states = $(\nu_1+\nu_2+\nu_3)(\nu_2+\nu_3)\nu_3$ =120. Assume $Q_1 = (q_{(1),ij})_{12x12}$, $Q_2 = (q_{(2),ij})_{5x5}$ and $Q_3 = (q_{(3),ij})_{2x2}$. Then $T=(p_{ij})_{120x120}$, where i, j=1, 2, …, 12
$p_{i+n_1 * i_2 + n_1 * n_2 * i_{3\,,\,j+n_1 * j_2 + n_1 * n_2 * j_3} = q^{(1)} * q^{(2)} * q^{(3)}$, with

$q^{(1)} = q_{(1),ij}$
$q^{(2)} = q_{(2),i_2,\ j_2}$
$q^{(3)}) = q_{(3),i_3,\ j_3}$
and
$i_2,\ j_2$ = 1, …, 4, 5.and $i_3,\ j_3$ = 1, …, 4, 5.

Although, by Property 5.1, the number of states grows extremely fast, in some cases it can be reduced significantly and the transition probability matrix can become a block matrix. For example, if for a two category confidential model n = 2 and ν_1 =2, and ν_2 =2 if

$Q_1 = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$, where A is a 2x2 matrix and

$Q_2 = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$. Then

$$T = \begin{bmatrix} Ar_{11} & 0 & Ar_{12} & 0 \\ 0 & 0 & 0 & 0 \\ Ar_{21} & 0 & Ar_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

And there are only four rows and four columns are non-zero rows or columns. Therefore there are only four states that are non-zero states, instead of eight states.
Likewise, for a three category confidential model of two states each, n=3, ν_1 =2, ν_2 =2 and ν_3 =2.,, instead of using 48 states, there are some possible cases which can reduce the number of states significantly, as shown in the following assuming parent company can access all states in the child company.:
Case 1: If a parent company (Q_1) invests its own Supplier1 company(Q_2) and has Supplier2 company(Q_3) for competition. There are only 4x2x2=16 states.
Case 2: If a company (Q_1) has Supplier1 company(Parent Q_2) and also has Supplier2 company(Child Q_3) for competition. There are only 2x4x2=16 states.
Case 3: If a company (Q_1) has 2 unrelated Supplier1 company(Q_2) and Supplier2 company(Q_3) for competition. There are only 2x2x2=8 states.

In the next section a two category (e.g. manager and employee categories) and a three category confidential models will be simulated under a variety of distributions.

6. SIMULATION OF STATE TRANSITIONS

6.1 Simulation of Two Category Model (when ν_1, ν_2 =2)

The simulation methodology can be found in [23] and the results can be found in [22].

6.2 Simulation of Three Category Model (when ν_1, ν_2, ν_3 =2)

A similar simulation results as a two category model for a three category one are shown in Figure 2 and 3 below. The initial state for all different distributions used in the simulation is p(0)=(p1(0), p2(0), …,p48(0))
=(
0.012280573930,0.033676035784,0.041838055971,0.114729155008,0.002450696980,0.006720350341,0.000716168443,0.001963891447,0.041838055971,0.114729155008,0.142535921970,0.390864859910,0.008349153550,0.022895216079,0.002439877449,0.006690680806,0.002450696980,0.006720350341,0.008349153550,0.022895216079,0.000489058225,0.001341105259,0.000142917737,0.000391911882,0.000436835332,0.001197898595,0.001488231833,0.0040810

59139,0.000087174348,0.000239051244,0.000025475005,0.000069858068,0.000237081035,0.000650128364,0.000807699191,0.002214888899,0.000047311614,0.000129738856,0.000013825897,0.000037913653,0.000042252076,0.000115864488,0.000143946425,0.000394732769,0.000008431775,0.000023121782,0.000002464022,0.000006756890) and =1.

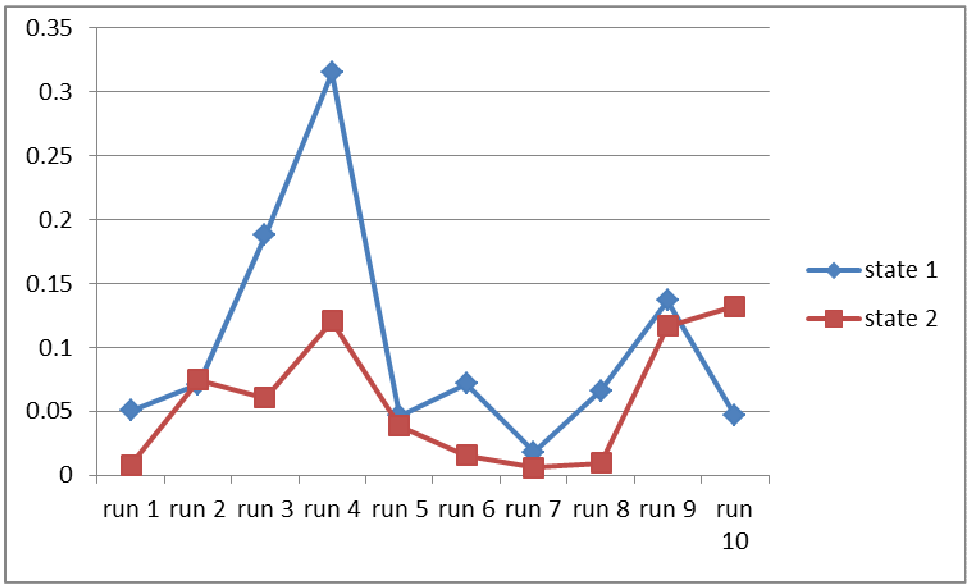

FIGURE 2: The steady states of the ten simulation runs when both category 1, 2 and 3 distributions are uniform (0,1) distribution.

The average steady states of all ten simulation runs for different distributions are shown in Figure 3.

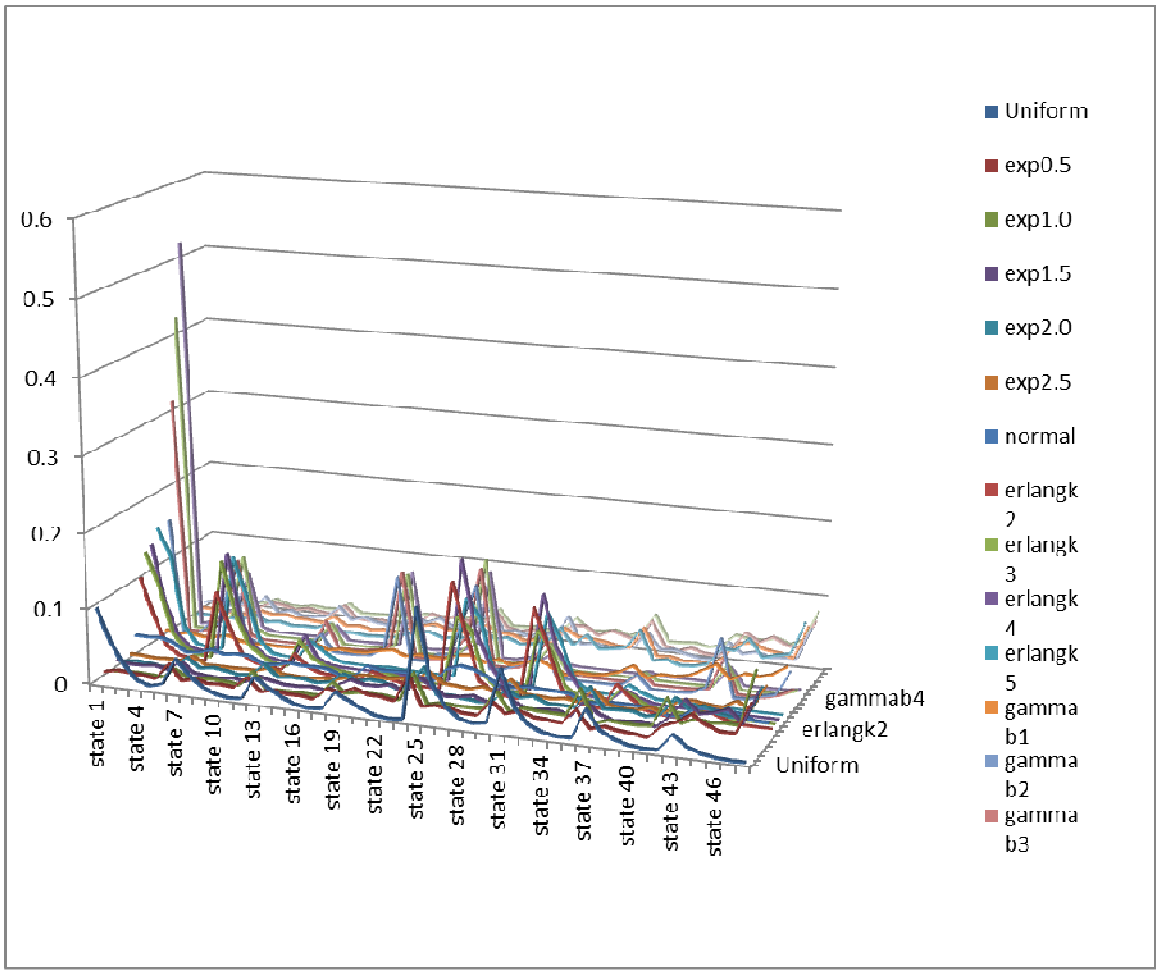

FIGURE 3: Comparison of Steady States for Different Distributions.

In Section 6 it describes how to validate an n category confidential model. An example of a two category model used in previous section is used to show how to use those properties.

7. VALIDATION FOR SEMI-MARKOV CHAIN MODEL

According to Rencher [24], the following two properties show how to test the hypothesis for the mean based on the average of the observations

Property 7.1.
In the n category confidential model if each steady state of an m observations $y_i \sim N_p(\mu_0, \Sigma)$, i=1,2,…,m are independently identically distributed normal random variables of p parameters each and if Σ is unknown, then the average $\bar{y} = \sum_{i=1}^{m} y_i /m \sim N_p(\mu_0, \Sigma/m)$. To test the hypothesis $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$. We reject H_0 at α level if $n(\bar{y} - \mu_0)' S^{-1} (\bar{y} - \mu_0) > T^2_{\alpha,p,m-1}$, where S is the sample variance-covariance pxp matrix $\sum_{i=1}^{m} (y_i - \bar{y})(y_i - \bar{y}) \Big/ (m-1)$ and $p = \prod_{i=1}^{n} n_i$ is the total number of states and T^2 is the Hotelling's T^2 test.

Property 7.2.

In the n category confidential model if each steady state of an m observations $y_i \sim N_p(\mu_0, \Sigma)$, i=1,2,…,m are independently identically distributed normal random variables of p parameters each and if Σ is known, then the average $\bar{y} = \sum_{i=1}^{m} y_i /m \sim N_p(\mu_0, \Sigma/m)$. To test the hypothesis H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$. We reject H_0 at α level if $m(\bar{y} - \mu_0)' \Sigma^{-1} (\bar{y} - \mu_0) > \chi^2_{\alpha,p}$, where Σ is the variance-covariance matrix, $p = \prod_{i=1}^{n} n_i$ is the total number of states and, χ^2 is the Chi-Square distribution.

Therefore for the two category confidential model described in Section 4 where p=8 and m=10 we can test the hypothesis H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$. We reject H_0 at 0.05 level if $10(\bar{y} - \mu_0)' \Sigma^{-1} (\bar{y} - \mu_0) > \chi^2_{0.05,8} = 15.51$ if Σ is known (Σ can be known from analyzing data from a long history or have a substantial evidence to support it). If Σ is unknown, we reject H_0 if $10(\bar{y} - \mu_0)' S^{-1} (\bar{y} - \mu_0) > T^2_{\alpha,8,9}$, where $\bar{y} = \sum_{i=1}^{10} y_i /10$ is the average of 10 final states of all runs. For the two category model described in Section 5, the mean μ_0 of final states of all ten runs of a randomly generated standard normal distributions N(0, 1) for both manager and employee transition matrices are recorded and the sample variance-covariance matrix Σ is created for each case (See Figure 4).

	state 1	state 2	state 3	state 4	state 5	state 6	state 7	state 8
run 1	0.16311	0.14008	0.15553	0.10319	0.12716	0.10920	0.12124	0.08045
run 2	0.03894	0.17662	0.08872	0.06968	0.06518	0.29565	0.14852	0.11665
run 3	0.27862	0.07009	0.07362	0.12027	0.23486	0.05908	0.06206	0.10138
run 4	0.23942	0.17365	0.20483	0.26506	0.03173	0.02301	0.02714	0.03513
run 5	0.19214	0.13896	0.17713	0.18278	0.08591	0.06213	0.07919	0.08172
run 6	0.02380	0.00512	0.00151	0.00139	0.72377	0.15591	0.04615	0.04231
run 7	0.25390	0.10694	0.07547	0.29020	0.09556	0.04025	0.02840	0.10923
run 8	0.24306	0.46896	0.05676	0.10998	0.03352	0.06469	0.00783	0.01517
run 9	0.25840	0.07277	0.04623	0.07815	0.30881	0.08697	0.05525	0.09339
run 10	0.12906	0.25043	0.14481	0.07030	0.08798	0.17073	0.09872	0.04793
Avg	0.18205	0.16036	0.10246	0.12910	0.17945	0.10676	0.06745	0.07233

Covariance	0.00969	0.00471	0.00625	0.00756	0	0	0	0.00115
	0.00471	0.01349	0.00397	0.00260	0	0	0	0
	0.00625	0.00397 75	0.00683 62	0.00581 77	0	0	0.00285 18	0
	0.00756 28	0.00260 6	0.00581 77	0.00957 94	0	0	0	0.00210 2
	0	0	0	0	0.02219 86	0.00524 13	0	0
	0	0	0	0	0.00524 13	0.00861 07	0.00553 92	0.00276 65
	0	0	0.00285 18	0	0	0.00553 92	0.00470 32	0.00295 42

0.00115			0.00210		0.00276	0.00295	0.00367
75	0	0	2	0	65	42	08

FIGURE 4: Average μ_0 and the sample variance-covariance matrix of Σ of 10 runs of randomly generated standard normal distributions for both manager and employees.

For example, assuming both manager and employee distributions are standard normal, there are ten observations (y_i , i=1,2, …,10) of states in the long run obtained by a manager. They are listed below

(state 1	state 2	state 3	state 4	state 5	state 6	state 7	state 8):
(0.671259	0.122945	0.075425	0.075867	0.038695	0.007087	0.004348	0.004373),
(0.347069	0.270521	0.137034	0.060936	0.07849	0.061179	0.03099	0.013781),
(0.524644	0.067031	0.039512	0.057259	0.237426	0.030335	0.017881	0.025912)
(0.179443	0.277235	0.076472	0.076476	0.114907	0.177528	0.048969	0.048972)
(0.27958	0.099922	0.054449	0.072657	0.272286	0.097315	0.053028	0.070762)
(0.54851	0.187313	0.115711	0.028219	0.074972	0.025603	0.015816	0.003857)
(0.364298	0.008646	0.001995	0.041343	0.510827	0.012123	0.002797	0.057972)
(0.35618	0.043088	0.049767	0.043699	0.366685	0.044359	0.051235	0.044988)
(0.11346	0.075408	0.044019	0.040133	0.302115	0.200791	0.117212	0.106863)
(0.256683	0.064596	0.027857	0.107778	0.305092	0.076779	0.03311	0.128104)

To test the hypothesis H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$, where μ_0 =

(0.182050	0.160366	0.102465	0.129105	0.17945	0.106764	0.067454	0.072339
1	8	4	8	3	9	9	)

The average of ten observations is $\bar{y}$ =

(0.364113 0.121671 0.062224 0.060437 0.23015 0.07331 0.037539 0.050558)

If Σ is unknown , from Property 7.1, $10(\bar{y} - \mu_0)' S^{-1} (\bar{y} - \mu_0) = 1979 > T^2_{0.05,8,9} = 697.356$. The null hypothesis is rejected. Therefore the security may have been breached.

The next section shows how the model helps manager to manage the dynamics of the security states.

8. MANAGERIAL IMPLICATIONS

The proposed model is general enough in practice. For example, in a supply chain network there are three groups involved: the purchasing group and two supplier groups. The model uses three security classifications which contains top secret (the purchasing group), secret (the first supplier group) and confidential (the other supplier group). If the first supplier group is not allowed to access the information in the second supplier group, we only need to set the appropriate part of the transition probability matrix to zero. The characteristics of the model are completely determined by the transition matrix. Based on the transition matrix we can determine whether the process will reach to the equilibrium state after a long period of time. Suppose we have a **scenario A** that a manager group first randomly evaluates suppliers before sending out a bid notice and request the bidding price. Then it repeats the whole process. In the mean time, an employee group randomly performs either providing bidding price or reading bidding notices. The transition matrix was given and the semi-Markov chain is periodic with period 4. Therefore, it is recurrent non-null. Since all states can be reachable from all other states, it is irreducible [25]. If at time 0 it has 42% chance that manager evaluates supplier, 40% chance that manager makes buying decision, 2% chance that manager reads biding notice and 16% chance that manager reads retail price and if it has 35% chance that employee reads biding notice and 65% chance that employee reads retail price, then at time 1000000 it has 14% that manager evaluates

supplier evaluation and employee reads bidding notice [15]. However, if we follow the **scenario B** that the purchasing group randomly performs those four actions and the supplier group performs those two actions randomly, then each state is a recurrent non-null and aperiodic. That is, the semi-Markov chain is ergodic [25] and the system has a steady state p(s) when time s is large. The simulation result is in the Figure 5 below:

state	1	2	3	4	5	6	7	8
Purchasing initial state	0.284308	0.216797	0.442041	0.056855				
Supplier initial state	0.964208	0.035792						
Time=0 state	0.274132	0.209038	0.426219	0.054820	0.010176	0.007760	0.015821	0.002035
State Transition probability matrix T	0.040019	0.028175	0.646285	0.514467	0.003465	0.002440	0.055963	0.044549
	0.335822	0.428346	0.019121	0.117963	0.029080	0.037091	0.001656	0.010215
	0.007283	0.041699	0.032780	0.014115	0.000631	0.003611	0.002839	0.001222
	0.334275	0.219179	0.019213	0.070854	0.028946	0.018979	0.001664	0.006135
	0.015764	0.011099	0.254588	0.202661	0.052318	0.036834	0.844909	0.672579
	0.132289	0.168736	0.007532	0.046468	0.439031	0.559990	0.024997	0.154216
	0.002869	0.016426	0.012913	0.005560	0.009521	0.054515	0.042855	0.018454
	0.131679	0.086340	0.007568	0.027911	0.437009	0.286540	0.025117	0.092630
Time=999997 state	0.047418	0.076108	0.006181	0.050500	0.215712	0.346228	0.028120	0.229734
Time=999998 state	0.047418	0.076108	0.006181	0.050500	0.215712	0.346228	0.028120	0.229734
Time=999999 state	0.047418	0.076108	0.006181	0.050500	0.215712	0.346228	0.028120	0.229734
Time=1000000 state	0.047418	0.076108	0.006181	0.050500	0.215712	0.346228	0.028120	0.229734

FIGURE 5: A semi-Markov chain Simulation Run using the scenario B.

Figure 5 shows the simulation run using protocol B after one million state transitions. We can see that the system has a steady state p(s)=(0.047418, 0.076108, 0.006181, 0.050500, 0.215712, 0.346228, 0.028120, 0.229734), where s=1000000. And it satisfies Tp(s)=p(s).
Any state which does not belong to one of the possible eight states is violates the security requirement. If an employee evaluates supplier, the system will warn the security manager to take actions. A large manufacturer may have more than hundreds of suppliers for various parts acquisition in different time periods. The semi-Markov chain model can help the managers to understand the confidential status of each supplier and then implement necessary security strategy for the organizations.

9. RESEARCH RESULTS AND CONCLUSION

By combining the subjects and objects possible security levels, all possible states can be listed in the semi-Markov chain model. In conclusion, since the confidentiality policy for the supply chain networks can be modeled by Bell-LaPadula model, semi-Markov chain model can be used successfully to simulate the state transitions dynamically for the Supply Chain networks. As we mentioned early, security standards today are emerging but many basic security principles in the standards can be traced back to existing security models. These standards and models are further impacting on the business strategy for the managers in an enterprise [21, 26]. ISO/IEC 17799:2005 provides “guidelines and general principles for initiating, implementing, maintaining, and improving information security management in an organization. The objectives outlined provide general guidance on the commonly accepted goals of information security management.” [4]. The semi-Markov chain model discussed in this paper shows the process of the secured state

during the time period in the supply chain network. Any state which does not belong to one of the possible state is considered as impeaching the security. For example, in the previous section only those eight states are allowed. If a general employee is conducting supplier evaluation, which is not in one of those eight states, the system will not allow the process to proceed to the next possible state and managers will be warned on security impeachment. In reality, a supply chain network is fairly complex. A large manufacturer may have more than 500 suppliers for various parts acquisition in different time periods. The semi-Markov chain model can help the managers to understand the status of each supplier and then implement necessary security strategy for the organizations. Although the model is useful for managers, however, because of Property 5.1, the number of states grows exponentially fast when the number of categories grows. It is suggested to be used when both the number of categories and the number of objects are small. This model can be also applied to password management in order to prevent threats from using the same password on other web sites.

10. ACKNOWLEDGEMENTS

This paper is in memory of my mother Sharn-Yun Chen for her life time encouragement, support and insatiable hunger for knowledge. Without her this paper would have never been completed.

11. REFERENCES

[1] CNSS (The Committee on National Security Systems) 4009, 2003.

[2] S. Lipner, Non-discretionary control for commercial applications, Proceedings of the 1982 Symposium on Privacy and Security, 2-10, 1982.

[3] K. Hsu and Z. Zhu, “SAS 99 – Consideration of fraud in a financial statement audit: a new auditing standard”, International Journal of Services and Standards, Vol. 1, No.4, 2005, pp. 414 – 425.

[4] M. Lee and T. Chang, “Applying ISO 17799:2005 in information security management”, International Journal of Services and Standards, Vol. 3, No.3, 2007, pp. 352 – 373.

[5] M. Bishop, Computer Security, Addison-Wesley, 2003.

[6] M. Shing, C. Shing, K. Chen and H. Lee. (2006). “Security Modeling on the Supply Chain Networks”, Journal of Systemics, Cybernetics and Informatics, 2008 , Vol. 5, No. 5, pp. 53-58.

[7] K. Biba, “Integrity considerations for secure computer systems”, Technical Report MTR-3153, 1, Bedford, MA: MITRE Corporation, 1977.

[8] D. Brewer and M. Nash, “ The Chinese wall security policy”, Proceedings of the 1989 IEEE Symposium on Security and Privacy, 1989, pp.206-214.

[9] D. Clark and D. Wilson, “A comparison of commercial and military security policies”, Proceedings of the 1987 IEEE Symposium on Security and Privacy, 1987, pp. 184-194.

[10] D. Bell and L. LaPadula,, “Secure computer systems: Mathematical foundations”, Technical Report MTR-2574, I, Bedford, MA: MITRE Corporation, 1973.

[11] D. Bell and L. LaPadula, “Secure computer system: Unified exposition and multics Interpretation”, Technical Report MTR-2997, Rev. 1, Bedford, MA: MITRE Corporation, 1975.

[12] K Chen, H. Lee and J. Yang, ‘Security Considerations on the Design of Supply Chain Networks’, the Proceedings of Southwest Decision Sciences Institute, Vol. 14, No. 1/2/3, 2006.

[13] K. Chen, M. Shing, C. Shing and H. Lee, "Modeling in Confidentiality and Integrity for a Supply Chain Network," Communications of the IIMA, 2007.

[14] M. Shing, C. Shing, K. Chen and H. Lee. (2006). "Security Modeling on the Supply Chain Networks", Proceedings of EIST 2006, Orlando, FL.

[15] M. Shing, C. Shing, K. Chen and H. Lee. "A Simulation Study of Confidentiality Modeling in a Secured Supply Chain Network", Proceedings of International Symposium on Intelligent. Information Technology Application conference, Dec. 22-23,2008, Shanghai, China.

[16] P. Bremaud, Markov Chains. New York: Springer, 1999.

[17] M. Aburdene, Computer Simulation of Dynamic Systems, Wm. C. Brown Publishing, 1988.

[18] Bhat, N. (1972). Elements of Applied Stochastic Processes, John Wiley & Sons.

[19] M. Molloy, Fundamentals of Performance Modeling. New York: Macmillan Publishing., 1989.

[20] G. McDaniel, IBM Dictionary of Computing. New York, NY: McGraw-Hill, Inc., 1994.

[21] A. Smith, "Strategic aspects of electronic document encryption", International Journal of Services and Standards, Vol. 3, No.2, 2007, pp. 203 – 221.

[22] M. Shing, C. Shing, L. Shing. (2012). "Analysis of a Two Category Confidentiality Model In Information Security", Journal of Communication and Computer, USA, 3(1), 2012.

[23] J. Banks, J. Carson, B. Nelson, Discrete Event System Simulation, New Jersey, Prentice Hall, 1996.

[24] A. Rencher, Methods of Multivariate Analysis. New York: John Wiley & Sons, 1995.
[25] E. Parzen, Stochastic Processes. San Francisco: Holden-Day., 1967.

[26] A. Smith, "Supply chain management using electronic reverse auctions: a multi-firm case study", International Journal of Services and Standards, Vol. 2, No.2, 2006, pp. 176 – 189.

Interactive Projector Screen with Hand Detection Using LED Lights

Padmavati Khandnor *padma_khandnor@yahoo.co.in*
Assistant Professor/Computer Science/Project Mentor
PEC University of Technology
Chandigarh, 160012,Iindia

Aditi Aggarwal *adi.23.pec@gmail.com*
Student/Computer Science/Final Year
PEC University of Technology
Chandigarh, 160012,Iindia

Ankita Aggarwal *ankita.aggarwal.pec@gmail.com*
Student/Computer Science/Final Year
PEC University of Technology
Chandigarh, 160012,Iindia

Swati Sharma *sharma.swati1990@gmail.com*
Student/Computer Science/Final Year
PEC University of Technology
Chandigarh, 160012,Iindia

Abstract

There are many different ways to manipulate OS, we can use a keyboard, mouse or touch screen. While doing a presentation, it is inconvenient to control the OS and explain the presentation at the same time. Our system, interactive wall, allows you to use hand to control OS on the projection screen which will act as touch screen. You can now experience a novel approach to control your cursor. Our system can be applied to the existing equipment, so you don't need to purchase any expensive touch screen. The system requires one web camera to detect the hand and then you can start to experience our system. Also, three LED lights are required to put on the fingers to help the system to detect the location of the hand and the gesture performing.

1. INTRODUCTION

1.1 Overview

The goal of this project is to build an interactive wall display for presentation or classroom use. The physical equipment required is not restricted. Any size of projection screen can be set up for the system or you can use a wall or a table to project the screen. User can be more flexible to enjoy our interactive wall at any place. Human Computer Interaction in the field of input and output techniques has developed a lot of new techniques over the last few years. With the recently released full multi-touch tablets and notebooks the way how people interact with the computer is coming to a new dimension. As humans are used to handle things with their hands the technology of multi-touch displays or touchpad's has brought much more convenience for use in daily life. The usage of human speech recognition will play an important part in the future of human computer interaction. This project introduces techniques and devices using the humans

hand gestures for the use with multi-touch tablets and video recognition and techniques for interaction. Thereby the gesture recognition take an important role as these are the main communication methods between humans and how they could disrupt the keyboard or mouse as we know it today.[10]

1.2 Motivation

- The size of touch-screens is usually quite small and limited.
- Expenses are directly proportional to size.
- Furthermore, some control methods are not very user-friendly.

➢ SOLUTION: using cameras to allow the same kind of user input as touch-screens do, but with a lower price and a larger screen.

➢ Hence AIM: to design a web camera-based system to perform like an interactive screen but with a larger sized screen and using our own hands to control it. [2]

2. METHODOLOGY

2.1 INPUT

Input includes valid region detection, hand shape recognition and gesture detection. Our system first implements valid region detection to achieve the display screen resolution and it identifies the x-y coordinates in the captured image on the screen. Then it recognizes the hand's shape and location and passes the x-y coordinates to the interface/output part of the system. Since the user may use different hand gestures to control the virtual mouse, the system detects and interprets specific hand gestures and passes the respective command(s) to the interface.

2.2 INTERFACE/OUTPUT

The interface is the device that can communicate between the recognition part and the OS. After the input part passes the data to the interface, the interface performs the action immediately and produces associated output. The system design follows the general architecture's five main components which are required to form the interactive Wall. The Prototyping Software Development Process design was used to test and refine the hand recognition algorithm until it achieved optimal performance for building a well-structured and robust system.

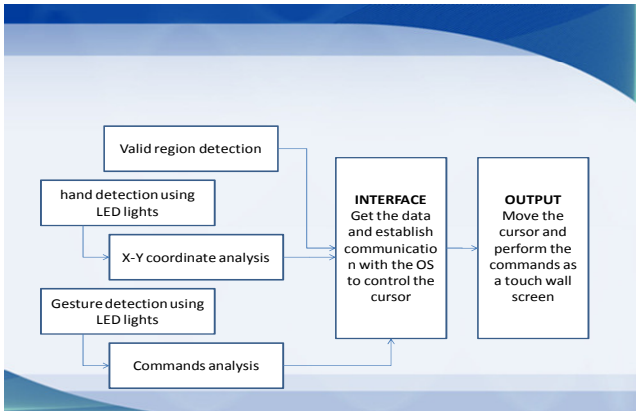

FIGURE 1: General architecture design

2.3 IMPLEMENTATION

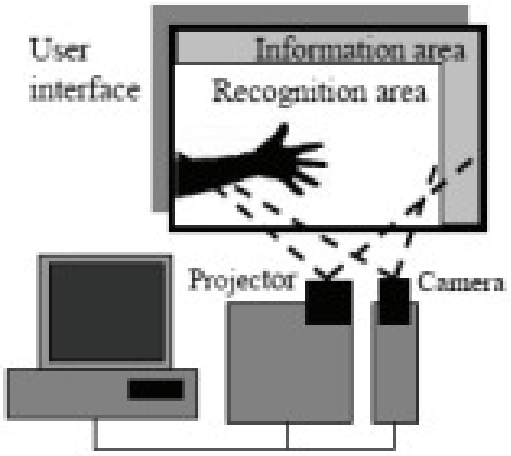

FIGURE 2: System Overview

2.3.1 VALID REGION DETECTION

For the system to be interactive, it requires real time processing and also synchronization with human hand gestures. In order to do this, we carefully identify the projected area for accurate positioning of the cursor in the design phase. If the hand is out of the projected area, then all the gestures or commands are invalid. The system only executes the commands when the hand is inside the projected area. Since the projected area must be a rectangle, we decided to use this characteristic to identify the area. By using the contour finding function (i.e. cvFindContours) from OpenCV, it was easy to find out four end points (i.e. the 4 corners of the projected area) of a contour. But this method failed as it may be possible that laptop, hence camera is at some angle to the screen, which caused incorrect mapping of x-y coordinates to the mouse cursor. So we decided to implement a function which allows the user himself to crop the image shown in the camera input to show up only the valid region(projector screen).[7]

2.3.2 HAND DETECTION

[1]LED lights provide a stable and efficient way to locate the position of the hand. The system only requires three LED lights because this number of lights provides sufficient commands and no interference affects the control.

Noise Filtering

1. Set the level of threshold to filter all the light of less intensity.
2. Cover the camera (either integrated or external) lens with a black film negative (photo reel film used in old cameras). This was the best method which let the system see only the bright white LED light.[9]

The current implementation tracks the brightest point in the current frame and show it encircled by a green colored pen appended with the x, y coordinates. This point was considered as centre (F). This gesture sends the MOUSE_MOVE command. If a second light is detected to the left of it (L) highlighted with a blue colored pen then a left click is generated and if a second LED light shows up on the right side (R) of (F),(highlighted with red color) then a right click is generated.

Both L and R are considered valid gestures only if they appear within region of interest (currently set by us as +-20 as radius).

2.3.3 GESTURE DETECTION

[3]After the hand detection, a maximum of three light sources can be detected: (i) the light that is closest to the track point is the first light F; (ii) the second light at the left hand side of F is L; (iii) the third light at the right hand side of F is R. We can imagine that F is the center of a circle with a certain radius. L and R will appear inside circle and be regarded as valid gesture commands.

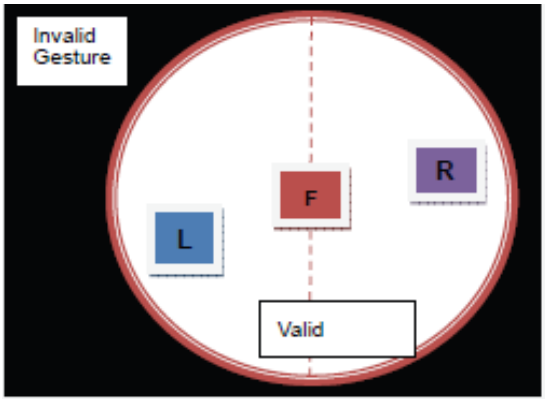

FIGURE 3: Gesture detection

The system relies on F, L and R to determine the actions. When only light F appears, the cursor synchronizes the position of the light on the projection screen. When light L appears on the left hand side of light F, then left click is executed. When light R appears on the right hand side of light F, then right click is executed. If three lights appear for a second, then the function of scrolling is executed. Since L and R can appear on any left part and right part in the circle respectively, it is difficult to keep tracks of L and R. Only F can be kept track of and used as the input method. The system analyzes the direction of movement of the track point and compares it with the pre-define pattern. If the input and the pattern match, then the corresponding command will be activated.

3. IMPLEMENTATION RESULTS

1. **Cursor movement** (with single LED detection): LED light on Index finger is detected by the camera as the centre point F. It takes this as the X-Y coordinate on the screen and maps the mouse cursor at this point.

Figure 4: Mouse Cursor tracking with single LED

2. **Right Click Implementation** (with LED detection to the right side of the index finger LED previously detected): LED light on Index finger is detected by the camera as the centre point F. Then another LED on the right side of F is detected as R. It takes F as the X-Y coordinate on the screen and implements right click at this point.

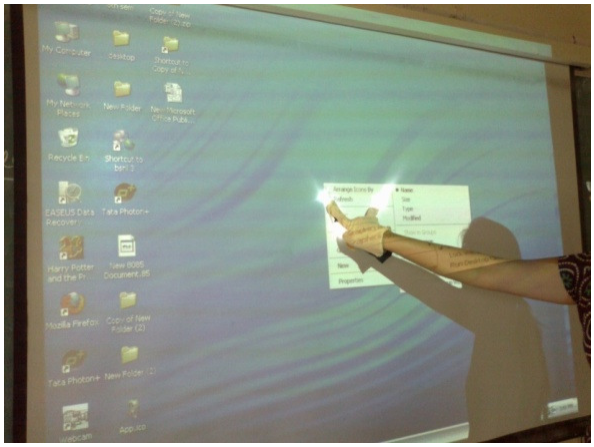

Figure 5: Right Click Implementation

3. **Left Click Implementation** (with an LED detection to the left side of the index finger LED previously detected): LED light on Index finger is detected by the camera as the centre point F. Then another LED on the left side of F is detected as L. It takes F as the X-Y coordinate on the screen and implements left click at this point.

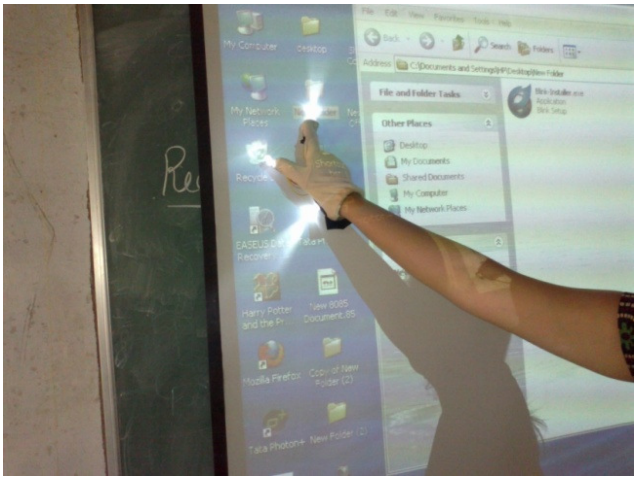

Figure 6: Left Click Implementation

4. CONCLUSION

1. Our system, Interactive Projector Screen, allows user to use LED hand glove to control OS on the projection screen but not touch screen.
2. The system requires one web camera to detect the LED light and then you can start to experience our system.
3. Also, three LED lights are required and sufficient to put on the fingers to help the system to detect the location of the hand and to perform mouse functions.
4. Any size of projection screen can be set up for the system or you can use a wall or a table to project the screen. User can be more flexible to enjoy our interactive wall in any places.

End Deliverable would be

1. A webcam with a particular resolution and other specifications with its lens covered with a black film negative to filter all the noise and light entering into the classroom to enable very smooth and stable functionality.
2. A Hand Glove with LED system integrated. This covers three 1Volt LED lights connected in parallel with a 9 volt battery. The battery lifetime is around 20 days if kept switched on continuously. The system also consists of 1pos On-Off switch which I used for simultaneous control of all 3 LED lights.

5. FUTURE SCOPE

1. Use with higher efficiency without preliminary training of gestures.
2. Interactive training of gestures to avoid retraining if an untrained user would like to use the system
3. To increase the number of gestures.
4. **Voice integration**
    a. Speech Detection is always mentioned as the most common straight-forwarded way, after the gestural motion, of how people interact between each other. This fact of course impacts also the design of human computer interfaces. Within the section of speech detection the main point is of course the software. For speech recognition itself you only need a normal microphone. The only thing you then have to consider is noise which will be also recorded with your actual planed voice. The major thing thereby is to create a good algorithm not only to select the noise from the actual voice but rather detecting what humans are actually saying.
5. **Gaming**
    a. The further implementation can be towards implementing features of gaming apart from the class room teaching and presentation applications , wherein gesture controls are able to manipulate the game operating system.[4]
6. **Key board implementation**
    a. Many softwares exist which allow to open up a virtual keyboard. By installing any such software and projecting it as well on the screen can allow us implement all keyboard functionality same as provided by Sixth Sense Device of Pranav Mistry.[2]

6. REFERENCES

[1]Attila Licsár1, Tamás Szirányi1, ”Hand Gesture Recognition in Camera-Projector System “,2009.

[2]Pranav Mistry, ”Sixth Sense Device”, November, 2009.

[3]Cristina Manresa, Javier Varona, Ramon Mas and Francisco J. Perales., “Hand Tracking and Gesture Recognition for Human Computer Interaction.”, Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain, 18 May 2005.

[4]Thomas Hahn,”Future Human Computer Interaction with special focus on input and output techniques”, at University of Reykjavik, March 26, 2010.

[5]Laura Boccanfuso1 and Jason M. O’Kanel, ”Adaptive Robot Design with Hand and Face Tracking for Use in Autism Therapy”, presented at Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29208, USA.

[6]Jun Park1, Yeo-Lip Yoon2, ”LED-Glove Based Interactions in Multi-Modal Displays for Teleconferencing“, presented at ik University, 2 Software Development Department, Homecast Company

[7]Eric Wong, “Use your webcam as a mouse”. Internet: www.youtube.com/watch?v=yUPaEnsKJYM,June 26, 2008.

[8]“Laser guided pointer”. Internet: http://www.youtube.com/watch?v=K_gnM9Ax-Kc,Jul.13, 2011.

[9] Audio-Visual Project lab, Elect. Eng. Div, Ngee Ann Polytechnic, Singapore, ‘‘DIY Penlight Mouse Controller - Interactive Webcam Inferface”. Internet:wwwww.youtube.com/watch?v=L476V10Ozi0,Feb.19, 2008.

[10] Audio-Visual Project lab, Ngee Ann Polytechnic, Singapore, “Interactive Projector Screen-Low cost Webcam Implementation”. Internet: http://www.youtube.com/watch?v=OTCWhraw2Xw,Feb.20,2008.

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Computer Science and Security (IJCSS)* is a refereed online journal which is a forum for publication of current research in computer science and computer security technologies. It considers any material dealing primarily with the technological aspects of computer science and computer security. The journal is targeted to be read by academics, scholars, advanced students, practitioners, and those seeking an update on current experience and future prospects in relation to all aspects computer science in general but specific to computer security themes. Subjects covered include: access control, computer security, cryptography, communications and data security, databases, electronic commerce, multimedia, bioinformatics, signal processing and image processing etc.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 7, 2013, IJCSS will be appearing in more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCSS LIST OF TOPICS

The realm of International Journal of Computer Science and Security (IJCSS) extends, but not limited, to the following:

- Authentication and authorization models
- Computer Engineering
- Computer Networks
- Cryptography
- Databases
- Image processing
- Operating systems
- Programming languages
- Signal processing
- Theory
- Communications and data security
- Bioinformatics
- Computer graphics
- Computer security
- Data mining
- Electronic commerce
- Object Orientation
- Parallel and distributed processing
- Robotics
- Software engineering

CALL FOR PAPERS

Volume: 7 - **Issue:** 1

i. Submission Deadline : January 31, 2013 **ii. Author Notification:** February 15, 2013

iii. Issue Publication: April 2013

CONTACT INFORMATION

Computer Science Journals Sdn BhD
B-5-8 Plaza Mont Kiara, Mont Kiara
50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2012
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org