Computer Forensics on Financial Crimes

Tsetini Maria
SID: 3301120006

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of
Master of Science (MSc) in Information and Communication Systems

OCTOBER 2013
THESSALONIKI – GREECE
Computer Forensics on Financial Crimes

Tsetini Maria

SID: 3301120006

Supervisor: Prof. Vasilios Katos
Supervising Committee Members: Assoc. Prof. Name Surname
                                Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of
Master of Science (MSc) in Information and Communication Systems

OCTOBER 2013
THESSALONIKI – GREECE
Abstract

This dissertation was written as a part of the MSc in Information and Communication Technology (ICT) Systems at the International Hellenic University. The topic of this dissertation is "Computer Forensics on Financial Crimes". The financial crime examined here is the Enron scandal, which is considered to be one of the most complex financial crimes in the history of corporate fraud. In order to examine this white-collar crime, the publicly available Enron corpus was used. Enron corpus includes emails exchanged by Enron's employees and it was used as a vehicle so that a computer forensics methodology would be created. Specifically, Enron email dataset was processed in order to provide useful statistical analysis results. Furthermore, Enron email dataset was analyzed so that social network visualization of it would be generated. Hence, by analyzing Enron corpus, this dissertation aims to act as a computer forensics toolkit for further analysis that can be conducted on this subject in the future and for financial crimes in general.

I would like to thank my supervisor, Prof. Vasilios Katos, for his valuable guidance and support throughout my thesis process. It was a privilege to have the opportunity to work with such a devoted scientist and inspiring teacher, who always provided me with significantly helpful advice.

Additionally, I would like to express my gratitude to my wonderful family for their emotional and financial support throughout my studies and in my life in general. Finally, I would like to thank my amazing friends for their constant encouragement and belief in me.

Tsetini Maria,
06/11/2013
Contents

ABSTRACT.........................................................................................................................III

CONTENTS ......................................................................................................................IV

LIST OF FIGURES ...........................................................................................................VI

LIST OF TABLES ................................................................................................................VIII

1 INTRODUCTION ..............................................................................................................1

2 LITERATURE REVIEW ....................................................................................................5

   2.1 COMPUTER FORENSICS ..........................................................................................5
       2.1.1 What is computer and digital forensics? .........................................................5
       2.1.2 Goals of forensic analysis.................................................................................7
       2.1.3 Computer forensics process ...........................................................................8
       2.1.4 Elements of a good process ...........................................................................12
       2.1.5 Rules of computer forensics .........................................................................14
       2.1.6 Computer forensics methods .......................................................................15
       2.1.7 The importance of Computer Forensics .......................................................17

   2.2 DIGITAL CRIMES ......................................................................................................18
       2.2.1 Digital crime definition ...............................................................................19
       2.2.2 Classification of crimes ...............................................................................20
       2.2.3 Types and sources of digital data in digital crimes ....................................23
       2.2.4 Types and rules of digital evidence ..............................................................26

   2.3 FINANCIAL CRIMES ................................................................................................28
       2.3.1 Financial crime definition ...........................................................................28
       2.3.2 Financial crime categories .........................................................................29

   2.4 THE CASE OF ENRON .............................................................................................35
       2.4.1 The outline of the scandal ...........................................................................35
       2.4.2 The rise of Enron .........................................................................................36
       2.4.3 Keys to Enron's success ..............................................................................38
       2.4.4 Enron's Innovations ....................................................................................40
       2.4.5 The fall of Enron .........................................................................................43
2.4.6 Reasons of Enron's collapse ................................................................. 46
2.4.7 Federal Investigation ................................................................. 55
2.4.8 Convictions ........................................................................... 56
2.4.9 Compensations ..................................................................... 58

3 PROBLEM DEFINITION ........................................................................... 60
  3.1 The Enron Email Dataset ................................................................. 60
  3.2 Related Work on Email Analysis ....................................................... 61
  3.3 Dissertation's Problem Definition .................................................... 62

4 CONTRIBUTION ..................................................................................... 65
  4.1 Methodology and Technologies .......................................................... 65
  4.2 Statistical Analysis ........................................................................ 72
  4.3 Social Network Visualization ............................................................ 83

5 CONCLUSIONS ....................................................................................... 100
  5.1 Overview of the Dissertation ............................................................. 100
  5.2 Results Evaluation ........................................................................ 101
  5.3 Personal Reflection ....................................................................... 102
  5.4 Future Work ............................................................................... 104
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Distribution of information security attacks per country</td>
<td>5</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Types of digital forensics</td>
<td>6</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Digital evidence</td>
<td>8</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Computer forensics process</td>
<td>11</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Information security triad (CIA)</td>
<td>20</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Examples of digital data sources</td>
<td>25</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Digital evidence</td>
<td>28</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Financial crime categories and subcategories (Peter Gottschalk, 2010)</td>
<td>30</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Enron logo - &quot;Endless possibilities&quot;</td>
<td>42</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Enron's stock price chart (January 2001 - January 2002)</td>
<td>45</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Enron's demise</td>
<td>46</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Enron, &quot;A true story of false profits&quot;</td>
<td>50</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Enron's slogan &quot;ask why&quot;</td>
<td>54</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Federal investigation about Enron</td>
<td>55</td>
</tr>
<tr>
<td>Figure 15</td>
<td>A tweet from CNBC regarding Skilling's new sentence deal</td>
<td>57</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Ironic depiction of Enron's slogan &quot;ask why&quot;</td>
<td>59</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Screen shot of the code that split the dataset into the tables</td>
<td>65</td>
</tr>
<tr>
<td>Figure 18</td>
<td>What it is included in each table</td>
<td>67</td>
</tr>
<tr>
<td>Figure 19</td>
<td>The number of records included in each table</td>
<td>68</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Creation of table &quot;output&quot;</td>
<td>69</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Insert values into table &quot;output&quot;</td>
<td>70</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Insert values to table &quot;result&quot;</td>
<td>71</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Tables included in database &quot;enron&quot;</td>
<td>72</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Code for counting emails (1998)</td>
<td>73</td>
</tr>
<tr>
<td>Figure 25</td>
<td>Total number of emails in 1998 per month</td>
<td>73</td>
</tr>
<tr>
<td>Figure 26</td>
<td>Total number of emails in 1999 per month</td>
<td>74</td>
</tr>
<tr>
<td>Figure 27</td>
<td>Code for counting emails exchanged outside working hours range (2000)</td>
<td>75</td>
</tr>
<tr>
<td>Figure 28</td>
<td>Statistics (2000)</td>
<td>76</td>
</tr>
<tr>
<td>Figure 29</td>
<td>Percentages of total exchanged emails per month</td>
<td>76</td>
</tr>
</tbody>
</table>
Figure 30: Statistics (2001)........................................................................................................... 77
Figure 31: Total number of emails in 2002 per month......................................................... 79
Figure 32: Code that prints the counted emails per year per month only for emails exchanged in table "employeelist"......................................................................................... 80
Figure 33: Total emails - Employeelist emails per month per year ................................... 81
Figure 34: Total emails - Employeelist emails per year ......................................................... 81
Figure 35: Jeffrey Skilling and Kenneth Lay statistics................................................................. 82
Figure 36: Code to produce 1997-1998 networks................................................................ 85
Figure 37: 1997-1998 network (Kamada-Kawai layout).......................................................... 86
Figure 38: Code to produce networks of 1999.......................................................................... 88
Figure 39: The network of the first semester of 1999 (Kamada-Kawai layout).................... 88
Figure 40: Network of December 1999 (Kamada-Kawai and Fruchterman-Reingold layout)................................................................................................................................. 89
Figure 41: Code to generate networks of working hours in 2000........................................ 90
Figure 42: Network for November 2000 during working hours (Bin Pack algorithm). 91
Figure 43: Code for the creation of networks for each month of 2000 for emails exchanged between "00:00:00" and "08:59:59" at weekends............................................................... 92
Figure 44: Network for December 2000 during "00:00:00" and "08:59:59" time range at weekends (GEM and Bin Pack layout)........................................................................................................ 93
Figure 45: Code for filtering year 2001 for every 6 days and during working hours.... 94
Figure 46: Network August 11-16, 2001, working hours (Physics layout).................... 95
Figure 47: Network October 22-27, 2001, working hours (GEM layout and Bin Pack layout).......................................................................................................................... 96
Figure 48: Network November 28 - December 2, 2001, working hours (GEM layout, Bin Pack layout).................................................................................................................... 97
Figure 49: Network March 2002 (GEM, Bin Pack, Physics layout)................................. 99
List of Tables

Table 1: Advantages and disadvantages of in-house investigation ......................... 15
Table 2: Advantages and disadvantages of police investigation.............................. 16
Table 3: Advantages and disadvantages of forensic specialist investigation........... 17
Table 4: Bankruptcy facts .................................................................................. 46
Table 5: Order of digital evidence volatility ....................................................... 109
Table 6: Enron Segment and Stock Market Performance, 1993 to 2001 ............. 109
Table 7: Timeline of Critical Events for Enron in the Period August 2001 to December 2001 ................................................................................................................. 110
1 Introduction

Nowadays, computer technology is used at a remarkably high degree in every aspect of our lives. This is mainly because it has made our lives easier in several ways. Particularly, it has provided modern society with a major asset, which is the acceleration of processes through highly increased productivity and efficiency. Additionally, access to any kind of information is incredibly improved. However, could this enhanced access to information act as a doubled-edged sword in the usage of computer technology?

Unfortunately, computer technology can also be used in order to perform dishonest and in some cases even illegal activities as well. These activities constitute information security incidents. Information security is defined as protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction (Jason Andress, 2011). The aforementioned unethical activities are attacks that aim to violate information security triad (CIA), which is Confidentiality, Integrity and Availability of computer data and systems.

It is a sad fact of modern society that the number of information security attacks is increasing dramatically every year. Therefore, there is an intense need for information security experts that can constrain these attacks. Information security experts' scientific field is called Computer Forensics. Specifically, computer forensics is defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems, networks, wireless communications, and storage devices in a way that is admissible as evidence in a court of law (US-CERT, 2008).

Within this dissertation computer forensics regarding financial crimes is researched. Specifically, the case of Enron scandal is the financial crime that it is examined here. The Enron scandal (October 2001) has been characterized as the greatest failure in the history of American capitalism and its collapse caused investors to lose a large amount of money and employees to lose their jobs, their medical insurances and their retirement funds as well. Federal investigation about the scandal, which lasted almost five years, concluded that Enron's executives overstated its expecting profits from several businesses and also manipulation of its financial statements was
discovered. Besides the financial crime scope of the scandal, it can be considered as a computer assisted crime as well, due to the vast amount of inculpatory digital evidence that was discovered. Particularly, Enron related digital evidence was approximately 31 terabytes and it was processed by the Regional Computer Forensics Laboratory (RCFL) in Houston.

A few years after the scandal, a certain part of this digital evidence was made publicly available for scientific research. This part was the Enron email dataset, which was posted on the web by William Cohen, a professor at Carnegie Mellon University (CMU). It should be stated that the Enron email dataset is an extremely large dataset that contains the emails exchanged by Enron's employees during 1997-2002. Successively, after it was publicly available, several scientists examined this dataset in order to provide emails' classification, statistical analysis, social network analysis and visualization and text mining of the emails. In 2005, Jitesh Shetty and Jafar Adibi cleaned the dataset from duplicate messages and created a MySQL database for the dataset in order to statistically analyze it and derive a social network from it. This database is also publicly available and it is used in this dissertation in order to perform dataset analysis.

At this point it is important to mention the key objectives of this dissertation. So, in this dissertation, the aforementioned database is locally stored and processed in order to provide some meaningful results about the Enron email dataset. Therefore, concerning the analysis part of the dataset, one objective is to examine the dataset and provide some helpful statistical analysis results, so that the dataset will be deeply comprehended. Also, another objective is to display a social network visualization of the dataset in order to understand the distribution of the network. Successively, the successful accomplishment of these objectives will be used as the basis for achieving the main goal of this dissertation. Particularly, the main goal of this dissertation is to create a computer forensics methodology that can be regarded as the foundation when analyzing extremely large datasets, like the Enron email dataset. Hence, this dissertation aims to act as a computer forensics toolkit for further analysis that can be conducted on this subject in the future and for financial crimes in general.

In order to achieve the above objectives the following procedure is conducted. Initially, chapter Literature Review of the topic is presented. Literature review includes four sections, which are Computer Forensics, Digital Crimes, Financial Crimes and
The case of Enron. Computer forensics section involves computer forensics definition, goals, process, rules, methods and eventually its importance. Also, digital crimes section includes digital crime definition and classification and the types of digital data and digital evidence. Moreover, financial crimes section includes financial crime definition and its categories. Finally, the case of Enron section is comprised of the outline of the scandal, the rise of the company, its innovations and success keys, the collapse of the company and the reasons behind it, the federal investigation on the scandal, ex-employees convictions and compensations to victims.

In succession, chapter Problem Definition is displayed. Problem definition includes three sections, which are The Enron email dataset, Related work on email analysis and Dissertation’s problem definition. The first section describes the Enron email dataset and its importance. The second section presents some of the previous work on email analysis. Most of this work presented used the Enron email dataset in their research. Finally the third section lists the technologies used in this dissertation and describes the problems that are faced while analyzing the Enron email dataset.

The next chapter, is Contribution and it includes the implementation part of the dissertation. It consists of three sections, which are Methodology and technologies, Statistical analysis, and Social network visualization. The first section describes in depth the methodology and the technologies used before the analysis part of the dissertation. The second section describes the statistical analysis of the Enron email dataset and the technologies that are used. In this section some indicative figures are presented as well. The third section presents the social network visualization of the Enron email dataset. Also, the technologies and the terminology that are used in this part are fully described.

The final chapter is called Conclusions and it consists of four sections, which are Overview of the dissertation, Results evaluation, Personal reflection and Future work. In the first section of this chapter a summarization of the dissertation is presented. Furthermore, the second section displays an evaluation of the dissertation results. Additionally, the third section describes the impact of the dissertation to the writer. Moreover, the last section provides some suggestions about the future work that could be made on this subject.

Finally, at the last part of the dissertation Bibliography and Appendices are presented. Bibliography includes all the references that are used in the dissertation. It
should be mentioned that in this dissertation scientific papers, books, reports, a documentary film and web references are used. Appendices include three tables, which are considered as additional information and therefore they are not presented in the main text.
2 Literature Review

2.1 Computer forensics

In this section, the computer forensics definition, goals, process, rules, methods and eventually its importance will be described.

2.1.1 What is computer and digital forensics?

The incorporation of computer technology in modern life has increased the productivity and the efficiency in several aspects of it. However, computer technology is not only used as a helpful tool that enhances traditional methodologies. In unethical hands, it can be used as a criminal tool as well. Particularly, technical skilled criminals exploit its computing power and its accessibility to information, in order to perform, hide or aid unlawful or unethical activities (Reith, Carr, Gunsch, 2002).

Nowadays, the number of information security incidents is increasing worldwide. As you can see in the figure below, USA is the major target of criminals' attacks. Consequently, in order to prosecute criminals involved in computer related incidents, investigators must be able to apply precise and reliable forensics procedures.

Figure 1: Distribution of information security attacks per country.
Computer or digital forensics is a relatively new scientific area and there are several definitions written to specify it. It should be stated that initially there was a distinction between computer and digital forensics. The former referred to evidence acquired from a computer, while the latter included the forensics of all digital technology. This distinction is shown in the figure below. Throughout the years the definition of computer forensics expanded, but some people still consider it as a branch of digital forensics science. Nevertheless, computer and digital forensics follow the same principles and procedures regarding digital evidence and therefore others, including this dissertation, consider them as equivalent.

![Diagram of Types of Digital Forensics](image)

Figure 2: Types of digital forensics.

Firstly, it should be defined what forensics is. Generally, in dictionaries forensics is defined as the process of using scientific knowledge for collecting, analyzing, and presenting evidence to the courts. One of the first definitions about computer forensics was demonstrated by Rodney McKemmish in 1999. He suggested that *Forensic Computing is the process of identifying, preserving, analyzing and presenting digital evidence in a manner that is legally acceptable.* A more specific definition about computer forensics was proposed in 2004 by the Information Security and Forensics Society (ISFS). According to ISFS, *Computer Forensics is the science of obtaining, preserving and documenting evidence from digital electronic storage devices, such as computers, PDAs, digital cameras, mobile phones, and various memory storage devices.* Additionally, the United States Computer Emergency Readiness Team (US-CERT) proposed in 2008 a more accurate definition, which
suggests that computer forensics is defined as the discipline that combines elements of law and computer science to collect and analyze data from computer systems, networks, wireless communications, and storage devices in a way that is admissible as evidence in a court of law. Finally, the more precise definition was cited by the Digital Forensics Research Workshop (DFRWS) in 2001. According to DFRWS, digital forensics can be defined as the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.

Moreover, digital forensics can be characterized as an art and a science at the same time. As Cory Altheide and Harlan Carvey successfully cited in their book "Digital Forensics with Open Source Tools", while applying digital forensics there are two aspects of the practice that become apparent. Firstly, the science of forensics, which is conducted by applying scientific method and deductive reasoning to data, arises. Secondly, the art of investigation evolves, since the interpretation of data and the reconstruction of an event can be illustrated as art. Consequently, digital forensics can be portrayed as the science of forensics combined with the art of investigation.

2.1.2 Goals of forensic analysis
The main goal of any forensic investigation is to find facts that were left on the system and reconstruct the truth of an event based on these facts, which can be called remnants. So, the investigator divulges the truth of the event by uncovering these remnants, which can be characterized as artifacts or evidence.

As for the digital forensics, traces can be left on the computer system by any action that is performed. It is remarkable that even the most simple actions can create artifacts, such as registers changing in the processor. Obviously, more complicated actions probably will cause longer-lasting impressions to the system. Also, even if an actor attempts to clean up the artifacts that he or she created can lead to the creation of more artifacts.

As it is cited before, investigator's job is to reveal the truth of an event. Every investigation is based on a hypothesis. Investigator's job is not to prove this
hypothesis. On the contrary, investigator's job is to discover artifacts or evidence that prove whether this hypothesis is valid or not. In legal terms, a valid hypothesis leads to inculpatory evidence and a non-valid hypothesis leads to exculpatory evidence. The following figure is a digital evidence depiction.

![Digital evidence](image)

**Figure 3: Digital evidence.**

Furthermore, investigators face a considerably significant problem, which is called the "digital paradox". As it is mentioned before, digital artifacts are tremendously resilient since they cannot be completely erased. However, digital artifacts are extremely fragile as well. They can easily be manipulated or entirely fabricated. Consequently, it is the investigator's task to examine the consistency of the digital evidence and decide whether it is the original or a manipulated one. Sometimes this intricate task is the unique objective of the whole investigation.

Besides these authenticity and integrity issues, investigators deal with another important problem, which is the huge volume of digital data. The majority of the important and relevant data is in a digital form. Thereby, investigators scrutinize a large amount of digital data, which is an extremely time consuming procedure.

Consequently, examiners face some incredibly complex issues. Therefore, it is imperative that a framework of procedures to handle the evidence be applied. This framework is described in the next section as the "Computer forensics process".

### 2.1.3 Computer forensics process

Several approaches that describe the computer forensics process have been proposed. In the context of this dissertation we decided to adopt the proposal made by the
Electronic Discovery Reference Model. EDRM is an industry working group that aims to establish an industry standard process for the analysis and the production of electronic data. EDRM suggested that the digital forensics process should abide by the following stages:

- Identification.
- Collection and preservation.
- Analysis.
- Production and presentation.

Identification

The first phase of the process is the initial point at which the investigator needs to determine what actions should be taken later on. There are five fundamental steps that are included in this phase:

1. **Determine scope and quantity of the data.** At this point the investigator must be able to discern what the investigation will cover and the amount of data that will be examined.
2. **Identify repositories.** The investigator must be able to detect the location of the data that could be proved to be evidence. The location could be any digital device, such as personal computers, cell phones and personal digital assistants. Additionally, it should be decided whether the investigator possess the appropriate tools to accomplish the investigation.
3. **Strategize preservation.** The data that is found must be protected at all costs. This means that it must not be modified, because otherwise questions about the integrity of the findings may be raised. At this particular point, the investigator should be able to define the steps that will be implemented so that the data would be protected.
4. **Establish chain of custody.** Chain of custody is a significant legal requirement that must be established, so that the investigator could be able to prove that the data is not modified. This can be achieved by recording all the steps taken to capture the data, such as who accessed the evidence, when they accessed the evidence and what they did with it.
5. **Preview the data.** At this step, the investigator is able to preview the data in a forensically sound manner so that no modification occurred.

Collection and Preservation
This stage refers to the actual collection of the data that will be examined. Also, this is where preservation techniques are performed. It can be distinguished into four main steps:

1. *Identify the source media*. As it is mentioned before, data is stored in digital devices. The type of this data and the way of accessing it must be known to the investigator.

2. *Select acquisition parameters*. At this point parameters must be defined so that accurate imaging is achieved. Before determining these parameters, the investigator should consider the type of the case and the legal requirements.

3. *Create the image*. At this step the image is created and it must be guaranteed that the image is complete and no modification to the data is occurred. This can be validated by using metadata along with the image.

4. *Authenticate*. The objective of this step is to guarantee that the image created by the investigator is identical to the original obtained data. This can be achieved by using metadata cryptographic hashes. Initially, a hash of the original data is created and then a hash of the duplicate data is generated. After that, these two hashes are compared and must match in order to have identical original and duplicate data. Otherwise, the integrity of the investigation could be questioned. Additionally, it should be ensured that a secure hashing algorithm is applied, such as MD5 (Message Digest 5) and SHA-1 (Secure Hash Standard). These algorithms act like fingerprints and help investigators prove mathematically that the evidence remained the same since the day it was collected.

*Analysis*

Analysis is the core of the investigation and it refers to the process of data reviewing and examination. Also, it includes the reconstruction of fragments of data and conclusion findings based on the evidence. It could be file system analysis, log analysis, statistical analysis or any other kind of it. Completeness is extremely important in this phase. Investigators should search for any evidence that it could be considered as relevant. Hence, investigators should be characterized as creative with unconventional thinking and completeness in their actions in order to successfully accomplish the phase of analysis. In addition, they should be exceptionally skilled and well-trained, considering the fact that they interpret the analysis findings based on
their training, expertise, experimentation and experience. Finally, it should be stated that the analysis phase may take several iterations of examination considering the fact that analysis results can lead to additional collection of data and thus additional analysis.

Production and Presentation

After completing the investigation, the results found in the analysis phase are shared with the interested party or parties. Particularly, the investigator creates a report of the actions that were performed, the artifacts that were unearthed and the meaning of these artifacts. In other words, investigators summarize their conclusions and provide precise explanation of them. It should be stated that the presentation of the evidence must be performed in a legally acceptable and understandable manner. Consequently, the report should be simple so that even people with no computer experience would be able to understand it. Also, investigators should have strong communication skills in this phase because they may need to defend their findings under challenge.

The figure below displays the computer forensics process as it was described above.

![Computer forensics process](image-url)

Figure 4: Computer forensics process.
2.1.4 Elements of a good process

Undoubtedly, the job of a computer forensics investigator is an extremely difficult one. Since forensics is a legal process and the case will probably go to court, investigators must combine time-tested forensic techniques, legal framework, investigation skills and cutting-edge technology in order to successfully support their findings. Additionally, the combined use of concise deterministic and repeatable processes must be carefully developed. According to the book "Hacking Exposed Computer Forensics", written by Philipp, Cowen and Davis (2010), there are several elements that define a good forensics process. These are:

- Cross-validation of findings.
- Proper evidence handling.
- Completeness of investigation.
- Management of archives.
- Technical competency.
- Explicit definition and justification of the process.
- Legal compliance.
- Flexibility.

Cross-validation

Investigators should use more than one forensic tool in order to validate their findings. This is because every tool may have some holes and the opposing counsel could take advantage of them. Consequently, multiple toolsets should be used so that findings would be cross-validated. Alternatively, another technically skilled investigator who would not be involved in the case from the beginning could test the results of the investigation and validate them.

Proper evidence handling

Computer evidence is subject to the so-called "observer-effect", which means that even the mere act of viewing data on the system without performing proper forensic methods can lead to data modifications. Investigators must be able to prove that no modifications occurred and the evidence remained the same since the day it was collected. So, investigators must guarantee the chain of custody of evidence after the collection of it so that the authenticity and the integrity could not be questioned in the court.

Completeness of Investigation
It is important that the discovery of evidence be conducted in a complete manner. No missing of relative evidence should occur, because otherwise the case would have a terrible outcome in the court. In order to avoid this, the investigator should conduct a solid, tested process in his attempt to collect, analyze and report the evidence.

**Management of Archives**

Since forensics is a legal procedure, a judge decision does not necessarily mean that the case has closed. Maybe in the future the case will be revisited and the investigator should be able to rework on it. Therefore, the investigator should always conduct precise archiving and case management via document retention, data storage and backup policies. Obviously, authenticity and integrity of the evidence must be ensured during this kind of revision as well.

**Technical Competency**

Computer forensics investigators must be exceptionally technical skilled. Also, they should have a deep insight on everything they conducted during the investigation. This means that they will be able to justify any action they performed in case they have to do it in the court. Moreover, they are encouraged to challenge their tools assumptions, so that they would be aware of their tools’ weak and strong points.

**Explicit Definition and Justification of the Process**

Investigators are ought to conduct the investigation in a way that they would be able to retrace all the steps they took during it. By doing this, they would also be capable of explaining their methodologies and to judges and opposing lawyers. Otherwise, if their methods are not clear and discrete, the whole investigation will fall apart.

**Legal Compliance**

Computer forensics investigators must make sure that the whole investigation process that they conduct complies with the existing laws on the particular subject. Therefore, they should request for some legal advice, otherwise their whole investigation might be meaningless. It is very important to always take into account that this is a legal procedure and their role is a supporting one in a bigger case.

**Flexibility**

Each computer forensics investigation is different, as it uses its own assumptions and requirements and has its own traps as well. As investigations continue, their requirements and assumptions may change due to technological changes. Therefore,
investigators should be able to ensure that they developed a process that can effectively deal with these changes.

2.1.5 Rules of computer forensics

Computer forensics is a relatively new scientific field but there are several rules that must be applied when conducting it in an investigation. This is mainly because it is foremost a legal process and a vast array of legal concepts and precedents are applied. Therefore, since it is a legal procedure rules play a crucial role while performing computer forensics.

The first rule, which is considered to be the most significant one, is to handle the original data in the minimum extent. The investigator must create duplicate copies of the original data and examine the duplicate data. This must be identical to the original data so that the duplicate data will not be characterized as controversial. So, the first rule must applied in order to ensure the authentication and the integrity of the evidence.

The second rule is to record any change that occurred to the data. In some cases, changes to the evidence may be unavoidable. For example, shutting down a computer can cause modifications to the memory and the temporary files. Thus, when modifications occur, some aspects of them like the nature, the extent and the reason why they occurred must be recorded. Also, if any actor participated in the modifications must be documented.

The third rule is the fact that investigators must comply with the rules of evidence. Investigators obey to these rules so that the concluding evidence collected in the forensics process will be legally accepted. Rules of evidence can be listed as admissibility, authenticity, completeness, reliability and believability. However, they will be described in the "Digital crimes" section.

Finally, the fourth rule of computer forensics is the fact that investigators should not exceed their knowledge. This means that investigators should not perform forensics in cases that are beyond their knowledge and skills. In these cases, the aid of a more experienced specialist investigator could be proved to be highly valuable. Additionally, if there is enough time, training could be helpful too. Nevertheless, investigators should not proceed with cases that they are not their field of expertise because this can cause irreparable damage to the outcome of such cases.
2.1.6 Computer forensics methods

Undeniably, the number of computer related crimes is increasingly growing over the last decade. Therefore, a great deal of private and public companies have decided to invest on information security, so that this number reduces. According to Karen Rydar (2002), companies can conduct computer forensics in three ways:

- In-house investigation.
- Police.
- Private sector forensic specialist.

**In-house Investigation**

In this case, investigation is conducted by the existing IT staff of the company. IT personnel have a great understanding of the system and collecting data from internal logs is their field of expertise. Hence, in investigations that internal logs must be examined, they are probably the best choice.

Nevertheless, computer forensics is not just about internal logs investigations. There is a variety of composite investigations that require the knowledge of a forensic specialist, so that the evidence collected could be admitted to the court. Additionally, sometimes investigators have to attend the court as expert witnesses and the existing IT staff of the company may not have the skills and the experience required for it.

Consequently, it is the least expensive method, but its effectiveness may be questioned in some incidents. Companies could hire forensic specialists in order to provide training to the existing staff. Training should be a continuous process, so that they have a deep knowledge of current trends and tools, which makes it costly for the company when no incidents occur. The following table summarizes the advantages and the disadvantages of this method.

Table 1: Advantages and disadvantages of in-house investigation (Karen Ryder, 2002).

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least expensive option.</td>
<td>Time intensive.</td>
</tr>
<tr>
<td>Quick response time.</td>
<td>Requires multi-skilled investigators.</td>
</tr>
<tr>
<td>No outside intervention for potentially &quot;brand&quot; damaging incidents.</td>
<td>Does not ensure evidence integrity.</td>
</tr>
<tr>
<td>Potential to develop in-house forensic teams.</td>
<td>Requires technical diversity.</td>
</tr>
<tr>
<td>Security staff know the company's system.</td>
<td>Requires constant awareness of hacker tools and methods.</td>
</tr>
</tbody>
</table>
The Police

In this case, investigation is conducted by the police. It is a method that provides guarantees, such as integrity and chain of custody of the evidence, but not all the police departments have specialist electronic crime units that can deal with computer crimes. Also, the police needs some evidence before performing an official investigation and this cannot happen in many cases. In addition, companies are unwilling to perform investigation through the police because law enforcement may cause a significant damage to their reputation. The following table presents the advantages and the disadvantages of this method.

Table 2: Advantages and disadvantages of police investigation (Karen Ryder, 2002).

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preserves the chain of custody.</td>
<td>Time intensive.</td>
</tr>
<tr>
<td>Ensures evidence integrity.</td>
<td>Resources not always available-could cause slow response time.</td>
</tr>
<tr>
<td>Specialized crimes units in operation in most states.</td>
<td>Requires constant awareness of hacker tools and methods.</td>
</tr>
<tr>
<td>Specialist units provide technical diversity.</td>
<td>Requires technical diversity that may not be available through local law enforcement offices.</td>
</tr>
<tr>
<td>Provides multi-skilled investigators.</td>
<td>Requires constant awareness of current forensic tools.</td>
</tr>
<tr>
<td>Produces evidence in court that is professional and easy to understand.</td>
<td>Requires constant awareness to changes in relevant legislation.</td>
</tr>
<tr>
<td>Provides recognized international qualifications.</td>
<td>Potential &quot;brand&quot; loss if certain incidents reach the public arena.</td>
</tr>
<tr>
<td>Availability of software utilities developed for law enforcement only.</td>
<td>May require some evidence prior to launching an investigation.</td>
</tr>
<tr>
<td></td>
<td>Restricted to their jurisdiction.</td>
</tr>
</tbody>
</table>

The Private Sector Forensic Specialist
In this case, company hires a forensic specialist to conduct the investigation. Undoubtedly, their experience, knowledge and expertise on these issues can result successfully to the outcome of the case. Nevertheless, the cost of such a method can be extremely high, reaching even hundreds of thousands of dollars.

Table 3: Advantages and disadvantages of forensic specialist investigation (Karen Ryder, 2002).

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preserves the chain of custody.</td>
<td>Time intensive.</td>
</tr>
<tr>
<td>Ensures evidence integrity.</td>
<td>Most expensive option.</td>
</tr>
<tr>
<td>Quick response time.</td>
<td>Requires constant awareness of hacker tools and methods.</td>
</tr>
<tr>
<td>Resources available.</td>
<td>Potential &quot;brand&quot; loss if certain incidents reach the public arena.</td>
</tr>
<tr>
<td>Provides multi-skilled investigators.</td>
<td>Requires constant awareness of current forensic tools.</td>
</tr>
<tr>
<td>Produces evidence in court that is professional and easy to understand.</td>
<td>Requires constant awareness to changes in relevant legislation.</td>
</tr>
<tr>
<td>Provides recognized international qualifications.</td>
<td></td>
</tr>
<tr>
<td>Provides technical diversity.</td>
<td></td>
</tr>
<tr>
<td>Skilled staff often have law enforcement background.</td>
<td></td>
</tr>
</tbody>
</table>

Consequently, organization's management should opt for the best method based on the incident and organization's status. There is no solution that can fit in all cases. Companies should take into account that in some cases maybe all the methods will be used. However, under no circumstances should the method be chosen after the incident occurs. Organizations should be well-prepared by developing an incident response plan before it occurs, so that the damage would be minimized.

2.1.7 The importance of Computer Forensics

Nowadays, more and more people are using computers and devices with computing capability, which means that a vast amount of significant and fragile digital data is stored in such devices. This has resulted in an increasing number of computer crimes. Therefore, computer forensics must be applied so that the integrity and the survivability of this data will be ensured. Additionally, computer forensics should be considered as a "defense-in-dept" security approach, which means that it is designed
on the principle that multiple layers of different types of protection from different vendors provide substantially better protection (US-CERT, 2008). For example, valuable information can be gained through a deep understanding of both legal and technical aspects of computer forensics.

Although a computer forensics procedure may be time consuming and disruptive, the potential costs of not conducting a proper digital forensics examination may be substantial if not disastrous (ISFS, 2004). Firstly, the loss or the destruction of vital evidence can severely damage the viability of the legal procedure. This means that a criminal may not be sentenced because of this. Secondly, while computer forensics is conducted, some organizational vulnerabilities and security risks may be discovered. So, without it organizations remain vulnerable to security risks and even future security attacks. This can harm the organizations in many aspects, such as decreasing reputation, loss in customer confidence and generally loss of business. Thirdly, if the protection of sensitive information like customer files, private data or confidential information fails, then the organization might face legal actions. Particularly, in cases of companies that their main job is to protect intellectual properties and maintain confidentiality, this would cause an irreparable damage, especially if data is not timely recovered.

Moreover, computer forensics help organizations realize the significance of computer security and the fact that appropriate access controls must be applied on digital devices. In addition, the increasing amount of corporate and organizational data that is generated and kept in digital devices must be carefully protected, since it can be proved to be vital in an investigation.

Consequently, computer forensics is mainly important because it can potentially save money. This makes it a cost-benefit analysis for organizations. If they do not take it seriously, they might be completely destroyed. The Enron scandal, which is thoroughly described later on, is an example of such a demise and its email dataset has been used in the trials and in many scientific analyses since then.

2.2 Digital crimes

In this section digital crime definition and classification and the types of digital data and digital evidence are included.
2.2.1 Digital crime definition

Digital crime is a particularly new kind of crime. Therefore, the literature demonstrates that the definition of the criminal area of digital crimes is not explicitly defined. As Hulst, R.C. van der and Neve, R.J.M. pointed out in their paper "High-tech crime, different crime types and perpetrators" in 2008, there is no conventional norm on shared concepts and often different definitions are used interchangeably. Also, the fact that different kind of experts, such as researchers, policy makers and law enforcement workers, use different terms when defining a criminal phenomenon, makes things even worse. Consequently, digital crime is not a consistent field right now, which triggers confusion and harms the development of a useful approach that can lead to a successful cooperation between experts in the field of digital crime.

As it is cited before, there is a variety of different definitions in the criminal area of digital crimes that are used interchangeably. For example, digital crime is referred in the literature as computer crime, cyber crime and high-tech crime. Specifically, according to ACPO guidelines (Association of Chief Police Officers), computer crime is defined as a crime which a computer or computers play a significant part. Additionally, cyber crime is characterized as any crime where a computer is a tool, target or both (Grabosky, 2007). Moreover, high-tech crime is defined as a range of criminal activities which make use of Information and Communication Technology (ICT) and may targets people, properties and organizations (using ICT as a means), or electronic communication networks and information systems (Hulst, R.C. van der, Neve, R.J.M., 2008). Furthermore, computer crime is classified as illegal activities that make use of electronic systems as a means to affect the security of computer systems and data (Michael Kunz, Patrick Wilson, 2004). Finally, digital crime is characterized as any offences against computer data or systems, which includes unauthorized access, modification or impairment of a computer or digital system (Australian Institute of Criminology, 2011). These crimes are offences against the information security triad, which is the confidentiality, integrity and availability of computer data and systems (The Council of Europe, 2012). The following figure portrays the aforementioned information security triad (CIA).
2.2.2 Classification of crimes

As it is mentioned before, according to ACPO guidelines, computer crime is as crime which a computer or computers play a significant part in their completion. However, ACPO guidelines defined two distinctive ways that a computer can participate in a crime. The first one is called "computer focused crime", while the second one is called "computer assisted crime".

Computer focused crime

In this case, the computer or ICT in general is used as the explicit target of the crime, which means that the criminal activity would not be operated without ICT. For example, web defacement, denial of service (DoS) and malware are considered as computer focused crimes. Consequently, this kind of computer crime mainly involves breaking into, manipulating and altering computer systems. Specifically, Hulst et al. (2008) demonstrated that computer focused crime can be categorized in four clusters:

- Unauthorized access to ICT.
- ICT failure due to data traffic.
- ICT failure due to data and system manipulation.
- Service performers.

Unauthorized access to ICT

Unauthorized access is described as electronic intrusion, or gaining access to resources via a computer resource without permission (Michael Kunz, Patrick Wilson, 2004). Hackers are those who perform criminal activities in order to gain
unauthorized access to computer systems and usually they have strong financial incentives to perform such activities. Their technical knowledge is of remarkably high level and they are able to violate secure and insecure systems and develop tools to trigger failures in these systems. For example, one tool that hackers usually develop to achieve their criminal activities is botnet. Botnet is a collection of remote-control computers that communicate with each other in order to perform a variety of computer crime activities, such as spamming, phishing and DoS attacks.

ICT failure due to data traffic
Criminals are capable of using data traffic in order to disturb the operation of systems. This can be achieved in various ways and the two most significant of them are DoS attacks and spamming. A Denial of Service (DoS) attack aims to violate the availability of a system. Specifically, criminals intentionally send massive quantities of data to systems in order to overload them and make them unavailable. As for the spamming, it refers to massive quantities of emails that are sent primarily for advertising purposes, but also causes system failures as well. Additionally, massive quantities of phishing emails are sent so that sensitive personal information is obtained by criminals in order to accomplish their goal, which is to extort money from people.

ICT failure due to data and system manipulation
In this cluster, ICT failures are a result of data and system manipulation, which mainly involves performing actions in order to damage, delete, change or destroy them. One of the most widespread tool for such manipulations is malware. Malware, which is short for "malicious software" refers to software programs that run on a computer in order to damage or perform other unwanted actions on it. Obviously, no authorization from the system administrator is provided in such tools. However, they manage to remain unobserved. Malware can be used to achieve several purposes. For example, it can be used to gather sensitive personal information (Spyware), damage data and system (Virus), grant external access to computers (Trojan horses), block or alter websites (Web-defacement) and perform hacktivism.

Service performers
This cluster refers to cases that criminals hire experts with high level technical knowledge in order to accomplish or hide their criminal activities. This means that in this case ICT service performers are connected to organized crime. Common types of
service performance are the corruption ICT staff, infiltration by criminal ICT workers and hiring ICT experts (Hulst et al, 2008).

**Computer assisted crime**

In this case ICT is not used as the target of the crime, but as the tool to commit it. Some examples of computer assisted crimes are distribution of pedophile content, embezzlement, money laundering, harassment, credit card theft, intellectual property theft. In all these cases computer and digital technology in general is used as a tool to perform a great deal of offences. Additionally, according to Hulst et al (2008), computer assisted crime can be classified into four main categories:

- Legal communication and covert shielding.
- Illegal trade.
- Economic and financial crime.
- Illegal communication.

**Legal communication and covert shielding**

According to Hulst et al (2008), it includes three subjects: radicalization and extremism, terrorism and ideologically motivated crime, and innovative covert shielding using ICT. In this case, internet plays a crucial role because it is extensively used, especially by young people. Additionally, it is used to collect information and mobilize people. As a result, the development of knowledge is harmed and some significant trends and indications may not be taken seriously. In order to achieve their goals, criminals use some methods to protect their communication from others. For example, they use encryption and stenography and sometimes they hire experts to help them.

**Illegal trade**

Nowadays, the tremendous growth of internet has caused the creation of illegal trading of goods and services through ICT as a mean of communication. However, internet is also used as a distribution channel of illegal trading as it provides access to a large market and criminals believe that the possibility of being revealed is quite low. For example, internet is used extensively by criminals in order to distribute child pornography and therefore high level techniques are used by experts in order to constrain it.

**Economic and financial crime**

This category contains non-violent crimes that are linked to the fact of earning illegal profit through fraud, deception and embezzlement. Fraud can be achieved via several methods, such as market manipulation, blackmailing and money laundering. The Enron case was a fraud and involved such cases. It should be noted that internet fraud uses some techniques that were described in computer focused crime, such as phishing and spamming. Also, digital blackmailing can be associated to hacking and DoS attacks that were mentioned above.

*Illegal communication*

Nowadays, more and more people are using digital devices to communicate over the internet. Unfortunately, in many cases this communication can be characterized as illegal as it includes the exchange of messages that contain illegal content. For example, illegal communication may involves stalking, discrimination and grooming. Grooming can be described as the actions taken by adults in order to inappropriately approach children and sexually assault them. Additionally, there is another aspect of illegal communication through digital devices. This is the unauthorized and illegal interception of communication, which is performed by criminals using several different techniques, such as hacking, spyware, keyloggers and dishonest staff that is willing to perform illegal actions in order to make some profit.

### 2.2.3 Types and sources of digital data in digital crimes

Criminals who perform digital crimes always leave some traces after they finish with their illegal activities. These traces are digital data that are stored in criminals’ digital devices and they are used as evidence in order to prosecute them. Therefore, it is extremely significant to deeply understand the nature of such data and the source of it so that it can be appropriately used.

Up to date digital devices have extremely high computation power, which causes them to produce and store huge amounts of data. According, to Information Security and Forensics Society (ISFS, 2004), this data can be classified into active, residual and backup data.

Active data is the kind of data that it is generated by the users. For example, customer information and word processing documents can be characterized as active data. Moreover, active data can be further classified into metadata, operating system data, temporary files data and communications data.
Metadata refers to "data about data", which is basically important data that is kept within data files. For example, the date that a file is created and the user that created it is stored within the file and it is considered as metadata. Additionally, metadata provide a great deal of information about the data that they describe, such as when it was the last time that it was modified, printed or saved and the directory in which this was performed. As it is obvious, this kind of information is remarkably significant in criminal investigations.

Operating system data is data that is generated by a computer's operating system. This kind of data contain valuable information about a user's actions. Particularly, when a user surfs the internet, web browsers store all kind of information about the user's activities on the web and if cookies are enabled then even passwords are stored. Consequently, in digital investigations operating systems provide a vast amount of information which can be proved to be precious.

Temporary files data is data that is temporarily stored on the hard drive of a digital device when the user of the device executes a program. Even though documents or files that are processed by the user are not permanently saved on the device, temporary recovery files are created. So, in case of criminal activities, temporary files data present actions performed by the criminal without intentionally storing them. Therefore, this kind of data is very significant.

These days digital devices are extensively used for communication purposes and communication data is produced when users communicate with each other over these devices. For example, this data involves information about the sender and the receiver of the communication, the topic, the date and the time of their discussion and the files they were exchanged between them. Additionally, if the criminal attempts to erase all this kind of information, this would be conceived by the experts in case of investigation. Hence, communication data provide exceptionally helpful information when it is required.

It is a commonly believed that when someone delete files from a digital device, such as a computer, these files are entirely removed from it. However, this is not true because the operating system of these devices keep a directory of the name and the location of the each file. So, when a file is deleted it does not mean that the file data is removed. It means that the space is now available but the actual content of it is still kept on the device and it is stored in a location until it is overwritten. This
kind of data is called residual data and it provides vital information when it is known how to access it. Additionally, residual data contain other kinds of data, such as file fragments, unallocated data and slack space data, which also offer significant information.

Backup data can be produced manually by the user or automatically by the software features and contain information that is duplicated into portable digital devices, so that the user can retrieve this information even in a case of system failure. Hence, backup data can contribute to the investigation of criminal activities in the digital world.

As for the sources of digital data, there are numerous such sources. Obviously, the hard drive of a desktop or a laptop computer contain valuable digital data. Also, its memory and its cache memory include important digital data. Additionally, according to ACPO guidelines (Association of Chief Police Officers), its main unit, monitor, keyboard, mouse, modems and wireless network cards can contain significant digital data. Also, storage devices associated with computers can contain precious information. For example, external drives, shared drives, CDs, DVDs, floppy disks, Jaz/Zip cartridges, back-up tapes, PCMCIA cards, memory sticks, memory cards, and all kind of USB connected devices. Moreover, all electronic devices are susceptible to keep digital evidence. The most common devices in digital crimes are PDAs (Personal Digital Assistant), personal digital organizers, electronic organizers, palmtop computers, smart phones and generally mobile phones. Furthermore, power supply units, routers, digital cameras, printers and smart cards can contain digital data. In the following figure some of the most frequent digital data sources are shown.

Figure 6: Examples of digital data sources.
Besides the hardware examination, there are other sources of digital data that must be scrutinized. For example, countless sites require to initially sign up in order to have full access to it. Hence, the accounts of suspected criminals must be examined, because they may give us significantly useful information about the criminals. One account that is vital to be investigated is the email account of the suspect. Emails usually keep extremely important information about the suspect's activities that might be related to the crime. Such a case is the case that will be examined in this dissertation, which is the Enron email dataset.

All in all, since we live in a society that uses digital devices at a very high extent, digital data is everywhere. These devices are frequently used by criminals in order to perform illegal activities. In these cases, digital data contained in their devices can be used as digital evidence in order to recreate the events and prosecute them.

2.2.4 Types and rules of digital evidence
In cases that the aforementioned investigation of digital data is led to a legal proceeding, then this digital data must comply to certain rules in order to be accepted by courts as digital evidence. Also, it is important to be mentioned that evidence in general are categorized into three main types. These types are:

- **Real evidence.** Real evidence refers to unambiguous evidence that is based on demonstrable characteristics and it could stand on its own in courts without the need to depend on other factors. For example, in cases of computer crimes logs generated by audit functions could be considered as real evidence, supposing that logs are not modified at all.

- **Testimonial evidence.** Testimonial evidence is any kind of evidence (oral or written) that is provided by a witness, who in some cases is also the victim. The validity of this evidence category depends on the reliability of the witness. If the witness is considered to be significantly reliable, then this type of evidence is unquestionably helpful and it is considered to be as strong as the real evidence. However, if the witness is unreliable, then this kind of evidence is not useful.

- **Hearsay.** According to Oxford Dictionary, *hearsay evidence is information gathered by other people that one cannot adequately substantiate; rumor.* This
means that this kind of evidence is generated by witnesses that have no direct knowledge of what happened. Therefore, hearsay evidence lack reliability and integrity and it should not be presented in the court.

As for the rules of evidence, according to Matthew's Braid (2001) paper published for the Australian Computer Emergency Response Team (AusCERT), these rules are five and they shouldn't be overlooked, so that the digital evidence is useful and acceptable in court. The rules of evidence are:

- **Admissibility.** It is the most fundamental rule of evidence. Digital evidence must be in a form that it is generally accepted by courts. In case this rule is failed to be obeyed, then evidence is as it was never gathered. This means that the time and effort that were spent for gathering it was not worthwhile.

- **Authenticity.** Authenticity refers to the fact that digital evidence must be related with an incident in a relevant way because otherwise nothing can be proved.

- **Completeness.** Completeness of digital evidence is a very significant part of the whole process. Evidence that it is collected by investigators shouldn't just present one aspect of a criminal incident. Instead, all evidence should be included and evaluated for completeness, so that evidence unquestionably prove the case. For example, in a computer crime attack, it should be proved that the suspected attacker was logged in when the attack occurred, but also it should be explained why other users that were logged in at the same did not perform the attack. In legal terminology this is referred as "exculpatory evidence".

- **Reliability.** The authenticity and veracity of the whole process must not raise any questions or doubts about the way it was performed.

- **Believability.** Digital evidence that is displayed in a court ought to be well-defined, clear and in a plain form, so that it is understandable and believable by all people who perform their jury duties. However, the relationship between the simple form of the evidence and original scientific data should be presented, so that the evidence is not subjected to defraud.

Additionally, it should be stated that the chain of custody or the continuity of evidence is well preserved during the whole process. Specifically, investigators should be able to prove that a particular piece of evidence was at a particular place, at
a particular time in a particular condition (Karen Ryder, 2002). In computer crimes, this must be applied on the hardware collected from the criminal and on the data gathered from the hardware as well. Hence, the management of the digital evidence and the quality control of it is an extremely significant procedure. The following figure portrays an interesting image of digital evidence.

Figure 7: Digital evidence.

Finally, it is vital that it is considered that certain types of digital evidence will not remain in the memory for a long period of time. Such data is called volatile evidence and it should be examined at the very beginning. Some examples of volatile evidence in a computer crime and their order of volatility is presented in Appendix A.

2.3 Financial crimes

This section describes financial crimes and it classifies them into four main categories and several subcategories.

2.3.1 Financial crime definition

According to the UK Financial Services Authority (FSA) a financial crime includes any offence involving money laundering, fraud or dishonesty, or market abuse. However, this is relatively plain definition of financial crimes. Nowadays, there is an increasing growth of financial crimes, mainly because of the technological improvements that made it possible for criminals to extensively exploit them and perform financial crimes. Additionally, financial crimes in corporate organizations have increased considerably because of the size and complexity of accounting services and the inability of the statutory auditors in conjunction with further
constrains such as company laws and standards (Izedonmi Famous, Ibadin Peter Okoeguale, 2012). Therefore, it is imperative that a more accurate definition of financial crimes is presented. The Economic and Financial Crime Commission (EFCC), which is a Nigerian law enforcement agency that investigates financial crimes, published in 2004 a more detailed definition about financial crimes. So, according to EFCC financial crimes include violent, criminal and illicit activities committed with the objective of earning wealth illegally in a manner that violates existing legislation and these include any form of fraud, narcotic drug, trafficking, money laundering, embezzlement, bribery, looting and any form of corrupt malpractices and child labor, illegal oil bunkering and illegal mining, tax evasion, foreign exchange malpractice including counterfeiting, currency, theft of intellectual property and piracy, open market abuse, dumping of toxic waste and prohibited goods, etc.

Moreover, it should be mentioned that financial crimes are considered to be "white-collar" crimes. This means that they refer to crimes against property through a series of illegal activities so that the person who initiated these activities would be financially benefit at the expense of other people. People who commit this kind of crimes are considered to be highly educated, intelligent and opportunists with high self esteem who do not consider themselves as criminals.

2.3.2 Financial crime categories

There is a vast amount of financial crimes that has been performed by individuals and institutions up to now. Therefore, it would be helpful to classify them into categories in order to examine them more easily. In 2010, Peter Gottschalk in his book "Investigation and Prevention of Financial Crime" proposed a financial crime classification and the following picture displays this classification.
Fraud Crime

Fraud refers to the intentional distortion of truth in order to deceit someone, who is the victim, based on a series of unlawful actions. Based on the above picture, fraud crime includes:

- **Advance fee.** In this kind of fraud, lawbreakers get in touch with their victims via letters, emails, telephones and faxes, which are sensitive personal information and is obtained without any authorization. Particularly, in advance fee frauds, lawbreakers try to ensure a prepaid commission for a deal that will
never actually be accomplished. Nowadays, advance fee fraud is dramatically increased mainly because of the widespread use of internet, which enables communication all over the world.

- **Bank fraud.** Bank fraud refers to criminal unlawful actions, taken by offenders who intentionally perform a well-designed plan in their attempt to deceive a financial institution. In the case of Enron there was bank fraud among other crimes since it took financial support from financial institutions under false statements.

- **Cheque fraud.** Cheque fraud refers to the situation that a company's cheque is stolen, altered or forged by a lawbreaker who get access to its funds and is able of performing several actions concerning the account, such as closing it.

- **Click fraud.** Click fraud is a kind of fraud that takes place on the internet in pay per click online advertisements. In this kind of fraud, an individual user or a computer software deceitfully clicks on such online advertisements in order to increase the profit of it, without any willingness to gather more information about it.

- **Consumer fraud.** Consumer fraud occurs when fraudsters deceptively force consumers into paying for products that they never actually receive or they are of low quality or overpriced. In this case also the widespread use of internet has deteriorated the situation. However, consumer fraud may refers to incidents in which the consumer is fraudulent. For example, consumer insurance fraud is a case where the consumer is fraudulent and attempts to deceit the insurance company.

- **Credit card fraud.** Credit card fraud refers to the incident in which the offender uses the details of a stolen credit card in order to purchase products or services. Credit card may be stolen or credit card details might be illegally acquired from an insecure system. Also, credit card details might be illegally acquired by people who have legal access to this kind of information. Generally, credit card fraud is commonly characterized as a subcategory of identity theft.

- **Embezzlement.** Embezzlement is defined as the fraudulent conversion of another's property by a person who is in a position of trust, such as an agent or employee (http://legal-dictionary.thefreedictionary.com/embezzlement). In
the case of Enron, its executives embezzled investor's money while presenting fraudulent earnings.

- **Hedge fund fraud.** Hedge funds are companies in the private sector that pool investors' money and reinvest them into a great deal of complex financial instruments. In these cases the management of the company and its trading strategies are not available to the public in order to avoid having their competitors imitate them. However, since the practices of the company are hidden, frauds can easily occur.

- **Identity fraud.** Identity fraud refers to an identity theft incident in which the criminal steals sensitive personal information from the victim and pretend to be the person whose personal information was stolen. Then this information (e.g. social security number) is used to break into other aspects of the victim's privacy, such as credit card details.

- **Mortgage fraud.** Mortgage fraud refers to actions taken by people who presented false documents to financial institutions in order to secure a loan. Sometimes, realtors participated in the process in order to gain more money from their commissions.

- **Occupational fraud.** Occupational fraud is basically company fraud and includes the use of someone's occupation status as a tool for increasing personal gain through the intentionally misuse and manipulation of the company's resources or assets. Also, in some cases is referred as internal fraud and it can contain the presentation of false financial statements in order to increase executive bonuses. Clearly Enron was an occupational fraud.

- **Subsidy crime.** Subsidy crime relates to criminal actions performed by an individual person or a company in order to obtain government subsidies. These actions include presenting false information when applying for them or not using them according to the rules defined in the deal. An example of such a crime is the illegal price fixing cartels.

The Enron scandal is obviously an occupational fraud since its executives misused its resources for their own personal gain and manipulated financial statements for the same reason. However, it included other types of crimes as well such as bank fraud, embezzlement and money laundering.

*Theft Crime*
Theft crime is described as the illegal taking of someone's property without their permission or consent. In this case, the victim could be an individual person, a group of people or a business. Theft crime is classified into:

- **Art theft.** Art theft includes burglary, robbery and deception of art pieces. Sometimes it also involves money laundering mostly in the sense of tax evasion.
- **Theft of cash.** Theft of cash refers to the act of stealing cash money. Skimming, lapping and embezzlement are some examples of theft of cash.
- **Identity theft.** Identity theft is defined as the crime of obtaining sensitive private information of the victims without their approval and knowledge about it. The difference among identity fraud and identity theft is that the former relates to the actual mistreatment of such information and the employment of illegal activities in order to impersonate the victim, while the latter refers to the acquisition of such information by being a thief.
- **Intellectual property crime.** The notion of intellectual property was created to legally protect an organization's creative and intangible assets. Hence, intellectual property crime includes copyright piracy and counterfeiting of products. It should be cited that these days this kind of crime is closely linked to digital crime.
- **Inventory theft.** It is the reduction in the amount of the company's inventory as a result of illegal actions (e.g. employee theft, shoplifting, vendor fraud etc.). Usually this kind of crime is called "shrinkage".

**Manipulation Crime**

Manipulation is described as a tool of obtaining control over someone else's activities by using artful, unfair and unlawful means. Manipulation crime is divided into:

- **Bankruptcy crime.** Bankruptcy crime relates to illegal actions taken by criminals before filing for bankruptcy procedure. Examples of such unlawful actions are concealing assets, filing deceitful documents and bribing the trustee mandated by court.
- **Bid rigging.** Bid rigging refers to the unfair and dishonest privilege that is given to a business in an open tender competition in order to overpower its competitors based on inside information.
Competition crime. Competition crime relates to cartel collaboration in which market competitors cooperate in order to reduce competition and fix market prices. Obviously, prices imposed by a cartel cooperation are higher than those in a market with non-fixed prices.

Computer and cyber crime. Refers to crimes described above, as computer focused crimes and computer assisted crimes. According to surveys, the most economically damaged crimes of this category are the Denial of Service attacks. Also, it should be mentioned that Enron email dataset constitutes a field of examination in this kind of crime.

Counterfeit currency. Currency counterfeiting is usually performed by criminals in order to finance their activities or hide their profits.

Extortion. Extortion is defined as the coercible actions committed by offenders in order to illegally acquire an entity's property. An entity could be a person, a group of people or even an organization.

Ghost employees. Ghost employees are people who are registered in the payroll system of a company without actually working in the company. So, the payment is deflected to a bank account that was deceitfully created.

Inflated invoices. In this case a company or an employee of a company inflates the bills without any prior agreement with the entities that will pay these bills.

Money laundering. Money laundering involves actions taken by criminals in order to present money acquired by performing illegal acts as legally gained money. It is closely linked to other types of crimes, such as embezzlement and fraud. Enron's executives were accused of money laundering due to their practices in Enron's trading patterns, internet money movements and other activities.

Income tax crime. This type of crime refers to offenders' actions to disregard the income tax laws that are applied in each country in order to minimize their tax fees. In some cases, like the Enron case, this is conducted by discovering loopholes in the legislation.

Corruption Crime

Corruption crime can be described as the dishonest and illegal behavior and generally mistreatment of people in authority positions for their own personal gain. Corruption crime is classified into:
Bribery. Bribery is a type of corruption crime that it is performed by offenders in their attempt to ensure favorable treatment.

Kickbacks. Kickbacks involve the dishonest and unlawful actions of posing threats and blackmailing.

Organizational corruption. Organizational corruption has to do with actions performed by organization's employees in order to intentionally mislead its resources and assets or misuse its practices for their own personal benefit. However, these actions may intercept the organization from achieving its goals and gradually cause its demise. Enron is an example of organizational corruption.

Public corruption. Public corruption refers to the misemployment of the power of elected politicians in order to benefit themselves.

Consequently, there are several forms of financial crimes that occur in our modern society. At this point it should be stated that a lot of them occurred in the case of Enron that will be extensively described in the next section. According to the classification of financial crimes, which was illustrated above, the financial crimes conducted in the case of Enron are bank fraud, embezzlement, occupational fraud, computer and cyber crime (Enron email dataset), money laundering, tax evasion and organizational corruption.

2.4 The case of Enron

This section describes in depth the case of Enron scandal. Specifically, it includes the outline of the scandal, the rise of the company, its innovations and success keys, the collapse of the company and the reasons behind it, the federal investigation on the scandal, ex-employees convictions and compensations to victims.

2.4.1 The outline of the scandal

The Enron scandal (October 2001) has been characterized as the greatest failure in the history of American capitalism and its collapse had a major impact on financial markets as investors and creditors lost their trust in the company. It caused investors to lose a large amount of money and employees to lose their jobs, their medical insurances and their retirement funds as well. Additionally, it caused the dissolution
of Arthur Andersen LLP, which was the audit company that performed both the internal and external accounting for Enron Corp.

Enron is the most famous company in the world that collapsed so quickly and so entirely. As it is stated in the documentary film "Enron: The Smartest Guys in the Room" (2005), it had taken Enron approximately 16 years to increase from 10 billion to 65 billion of assets and it took them only 24 days to go bankrupt. This may sound unbelievable, but it is what actually happened. However, how is it possible for a well respected company that was ranked by the global business magazine "Fortune" as the most innovative company in the United States to fall down so fast? The main reasons behind this are the exploitation of the energy deregulation, the manipulation of financial statements and several high-risk deals regarding financial assets operations. In addition, individual and collective greed that was evolved inside the company played a significant role to Enron's crumbling. These reasons will be described later on in detail.

2.4.2 The rise of Enron

Enron was founded in 1985 by Kenneth Lay. It was born from the merger of Houston Natural Gas and InterNorth, which was a pipeline company located in Nebraska. It is important to be mentioned that Enron was founded after the federal deregulation of natural gas pipelines.

Initially, this led Enron to lose its exclusive rights to the pipelines and incur massive debt. So, Kenneth Lay had to come up with a new and innovative business strategy in order to endure the operation of the newly merged company. The goal was to generate profits and cash flow and thus he hired a young consultant with background in banking and asset and liability management to help him accomplish it. The name of this consultant was Jeff Skilling, who climbed the company ladder relatively fast and even became the CEO (Chief Executive Officer) of it in February 12, 2001. Skilling proposed a revolutionary solution to credit, cash and profit problems that Enron faced after the merger. He suggested the creation of a "gas bank" in which Enron would buy gas from a network of suppliers and sell it to a network of consumers, contractually guaranteeing both supply and the price, charging fees for the transactions and assuming associated risks (Thomas, 2002). Thereby, Skilling's innovative idea introduced a new product and a new paradigm for the industry, which
was the energy derivative. Consequently, Lay wanted to reward Skilling for his golden idea and therefore he created a new division in 1990 called Enron Finance Corp. and hired Skilling to run it. It quickly became the dominant player in the natural gas market and provided Enron with superior profits as Enron was able to predict future prices with greater accuracy.

Nevertheless, deregulation had a positive impact on Enron as well. Prior to the deregulation, most contracts between natural gas producers and pipelines were "take-or-pay" contracts. These contracts defined that pipelines had to purchase a predetermined quantity at a given price or be liable to pay the equivalent amount in case of failure (Healy and Palepu, 2003). Similar long-term contracts were agreed between pipelines and local gas distribution companies. Consequently, long-term stability in supply and prices of natural gas was guaranteed by these contracts. However, after the natural gas market deregulation, energy supply became a free-market commodity, where prices were deregulated and more flexible contracts were allowed. As a result, by 1990 Enron was benefited by the admission of spot prices, increased gas supply and flexibility because it owned the largest interstate network of pipelines.

In succession, Enron adopted a diversification strategy so that to increase its growth. This strategy began by transforming its traditional pipeline business to natural gas trading and was extended to a financial trader and market maker in electric power, coal, steel, paper and pulp, water and broadband fiber optic cable capacity (Healy and Palepu, 2003). Additionally, it embarked on risky international projects (e.g. Dabhol plant) regarding construction and management of energy facilities. So, since Enron's trading business was considered to be very innovative, the company believed that they could approach in a similar way other activities as well. The results were very promising and by the year 2000 Enron was the seventh largest company in the United States and "Fortune" magazine listed it as the most innovative company of the country (Curral and Epstein, 2003).

In order to achieve this diversification strategy, the company decided to change its corporate culture in a way that it would apply to the new transformed company image. Consequently, Jeff Skilling started recruiting the company with the best and the brightest traders and with the most talented MBA (Master of Business Administration) graduates in the country. One of them was Andrew Fastow, who was
hired in 1990 and soon became Skilling's and Lay's protégé. Fastow moved quickly the corporate ranks and became the CFO (Chief Financial Officer) in 1998. The newly hired graduates had to follow exhaustive schedules and in order to balance this out Enron provided them with several corporate perks and also merit-based bonuses. Skilling's philosophy behind this was to allow traders to "eat what they killed". In addition, his philosophy about the corporate world was shown by the PRC (Performance Review Committee) system that he introduced. With PRC employees were graded from 1 to 5, with 5s usually being fired within a six month period. Grades were based on the values of Enron, which were Respect, Integrity, Communication and Excellence (RICE). Unsurprisingly, PRC became known as the roughest employee-ranking system in the United States as a result of Skilling's firm belief that people are only motivated by money and fear. Thus, intense internal competition and secrecy regarding trading contracts evolved inside the company.

As Enron's growth was increasing, capital markets were impressed and only a few raised crucial questions about its operations and business strategy. Undoubtedly, Enron's gas trading idea was very successful mainly because of deregulation. However, similar approaches to other activities and risky international projects can be reported as failures. So, when investors became more suspicious about its questionable accounting methods ("mark-to-market accounting") and questioned the value of the company's stock was the beginning of Enron's collapse.

2.4.3 Keys to Enron's success
By the late 1990's, Enron accomplished repeated value innovations, which reduced the cost of gas and electricity to customers at approximately 40-50 percent. At the same time, Enron reduced its own cost structure as well and created the first national spot market for gas in which commodity swaps, future contracts and other complex derivatives effectively stripped the risk and volatility out of gas prices (Kim and Mauborgne, 1999).

The diversification strategy of Enron made it possible for Enron to participate in the whole supply chain regarding the energy field. By creating the natural "gas bank", Skilling wanted to make Enron to act as a financial banking institution intermediating between suppliers and buyers of natural gas. He stated that Enron did not acquire a competitive advantage by its "heavy" assets (e.g. pipelines) and the only
way to obtain it and dominate in the trading market was through information. Consequently, with the "gas bank" idea Enron was not only an energy supply company that bought pipelines and electrical power plants, but also a financial institution that acted as a major dealer in wholesale and derivatives transactions. This means that Enron was actually a two sided company; the energy supply and the dealer side. The former was provided useful information by energy producers regarding production costs and distribution problems. The latter was given valuable knowledge of order flow, since market participants contacted the company to inform it about their transactions (buy or sell). Naturally, this critical knowledge acquired by the dealer side was shared directly with its energy supply side. Thereby, Enron was provided a major competitive advantage (E. Moncarz, R. Moncarz, A. Cabello, B. Moncarz, 2006).

The energy supply side of Enron's business increased swiftly by investing a great deal of money in the United States and abroad. Investments caused a rapid accumulation of assets, which was mostly financed by debt, and a high debt-to-equity ratio was produced as well. However, this was partially hidden from the investors with the aid of partnerships, known as "Special Purpose Entities" (SPE). Additionally, the dealer side of Enron's business boosted swiftly by being benefit from the development of derivatives trading and from trading during the energy crisis in California in the summer of 2000. During this crisis, electricity power plants were intentionally shut down so that an artificial shortage of energy would be created. This unethical practice caused electricity prices to increase dramatically. Enron took advantage of the situation and made a huge amount of money (around 2 billion dollars) by betting that prices would increase, which they did. So, electricity deregulation transformed California's electricity market into a casino due to the lack of a well designed legislation. Generally, Enron acting as a financial institution and performing trading activities provided it with large profits because it was both a dealer and a key market participant, which also helped it obtain more market power. So, Enron earned money both from its trading volume and its advantageous position in the market.

Enron was initially well positioned to act as the intermediate between producers and utilities because it possessed a large network of pipelines and had expertise in managing the physical logistics of delivering gas through these pipelines.
Then, it quickly evolved and gained expertise in managing the business trading risks (e.g. spot market volatility, gas price fluctuations etc.). Thus, it can be stated that Enron's model was to acquire physical capacity in each market and then leverage that investment through the creation of more flexible pricing structures for market participants, using financial derivatives as a way of managing risks (Healy and Palepu, 2003). Enron supported that the expertise it had obtained in the gas trading systems it could be applied to new markets and continuous growth was assumed due to its diversification from a pure energy firm into a broad-based financial company. Therefore, the successes of Enron was rooted in its ability to manage risks in complex transactions. However, these very risks ultimately brought Enron down (Chatterjee, 2003). Furthermore, Enron's corporate culture was based on innovation and competitiveness, where employees were solely evaluated according to their quarterly results. Thereby, success and failure sprang from this culture.

2.4.4 Enron's Innovations
Enron introduced several innovations as its corporate culture instructed creative and hardworking people with smart ideas. In 1988, the company approached additionally to its regulated pipeline business, unregulated markets as well. In 1989, it initiated the Gas Bank, which allowed gas producers and wholesale buyers to obtain gas supplies and simultaneously guarantee the price risk.

In order to accomplish risk management regarding future gas prices, Enron used derivatives to lower risks. Throughout the years, trading derivatives became one of Enron's strongest point and most of it was done without any exchange regulating these deals. In 1999, Enron started using "mark-to-market" accounting, which was an accounting method that Enron believed it fairly presented the results of managing in its portfolio of trades and contracts (Chatterjee, 2003). This accounting type will be described in more detail later on.

Enron used hundreds of Special Purpose Entities in an attempt to seclude special risk from its core operations and help it pay less when borrowing money. Specifically, by using an SPE a company is able to increase leverage and ROA (Return On Assets) without having to report debt on its balance sheet. The company exchanges hard assets and related debt with an SPE in return for an interest. Then, the SPE can borrow money from financial institutions to buy company assets or
accomplish other operations without the debt or assets being presented on the company's financial statements. Since the SPE possesses some assets but not the company's existing debt, it is able to borrow money from financial institutions at lower rates. According to FASB (Financial Accounting Standard Board) guidelines, only a 3% of the SPE must be owned by an outside investor in order not to be considered as a subsidiary. In case of subsidiaries the company must include SPE's financial position and results of operations in its financial statements. As for Enron's complex operations with its SPEs, they will be discussed as a reason of its collapse.

Similar to the SPE, Enron used Volumetric Production Payment (VPP) with its transactions with oil and gas production companies. Particularly, Enron loaned money to oil and gas production companies and required to be repaid in oil and gas. This was because Enron could not afford to generously give million of dollars to these companies and then wait some years for the gas sales. So, financial vehicles were created in order to raise money for the VPPs, which were perfectly legal and laid the groundwork for the more complicated entities Enron created in the late 1990's, that eventually led to the company's demise (Curral and Epstein, 2003). Another financial vehicle was the Cactus Funds, which was a group of VPP contracts that were used in order to raise the cash Enron needed to accomplish the VPP loans. Enron presented them as creative solutions to problems of raising funds inexpensively. By mid 1993, Enron had used Cactus partnerships to raise around $900 million. Additionally, JEDI (Joint Energy Development Investors) was 50-50 partnership between Enron and the California Public Employees' Retirement System (CalPERS). JEDI was the first partnership initiated by a pension fund.

In 1994, Enron started electricity trading and in 1997 traded its first weather derivative. It continued trading in other markets as well (coal, pulp, paper, plastic, bandwidth). It should be mentioned that in 1997 Enron purchased Portland General Electric Corp., which was an electric utility company, for about $2 billion. In 1999, it created Enron Online (EOL) that it would perform wholesale commodity trading and the same year constructed its first bandwidth trade. By using EOL Enron became a counterparty to every transaction made on the platform. Also, EOL provided a safe transaction environment which was extremely important for credit risk management. EOL was a huge success and it dealt with $335 billion in online commodity trades in 2000. In addition, in 2000 it launched the EnronCredit.com. Its goal was to help
companies hedge their risk in trading transactions by purchasing and selling credit risk. Furthermore, in July of 2000 Enron and Blockbuster reach an agreement so that to provide video on demand to customers all over the world through high-speed internet lines. Enron invested hundreds of millions into broadband with very little in return. However, its strategy was evaluated very positive by the stock market and the next month Enron's stock hit an all-time high of $90.56. The following figure displays Enron's logo of that time about "Endless possibilities", which expressed their philosophy about innovations.

![Enron logo](image)

**Figure 9**: Enron logo - "Endless possibilities".

Moreover, Enron's innovations include expanding its business outside the U.S. borders. Enron International was a fully owned subsidiary of Enron and was created to construct and manage energy assets in markets outside the United States that were deregulated regarding the energy field. Its first major international project started in 1993 and included the construction of the Teeside electric power plant in the United Kingdom. The same year, Enron and Maharashtra decided to construct the massive Dabhol power plant in India. This was a very expensive (around $2 billion) and a very challenging project that finally opened up in 1999. In 1998, it purchased a British water utility, Wesses Water for $2.2 billion. Enron dealt with several challenges while entering these new markets. This was because in some of these projects Enron had the expertise in the field that was needed, but in others it did not. So, it was just assumed that it could expand its business to other types of energy assets. Additionally, international diversification, particularly in developing economies such as India and China, exposed Enron to political risks (Healy and Palepu, 2003). For instance, the
Dabhol project in India raised considerable political opposition and controversy and the fact that Enron had no previous experience in managing the construction and the operations of power plants made things ever worse. On the contrary, its competitors were experienced in such operations and therefore even if the project had been successful it wouldn't mean that Enron could create a sustainable advantage over them.

Consequently, Enron made plenty of costly deals that were also very risky. Some of them were successful, while others failed and contributed to its demise. For more details about Enron's deals and stock market performance from 1993 to 2001 see Appendix B. It should be mentioned that the figures reported there are as originally announced by the company.

2.4.5 The fall of Enron

In February 2001, Kenneth Lay announced his decision to retire and assigned to Jeffrey Skilling the positions of both president and CEO of Enron. At the same time, Enron's stock was valued around $80 and Skilling boasted in the company's annual conference with analysts that it should be traded approximately at $120.

In March 2001, Enron and Blockbuster video-on-demand deal was cancelled which led Enron's stock to fall to the mid-$60s. During that time, Skilling had started looking despondent and nervous in public. There is an example of such a behavior held in a conference call with analysts in April 2001. One of the analysts asked Skilling questions about Enron's balance sheet and the cash flow of earnings. Skilling furiously replied to him in a rude and insulting way. This was the first time that Skilling behaved like that in public. The next few months, Enron faced huge cash shortfall as a result of several risky underperforming deals. Meanwhile, senior management continued selling Enron stocks and earning hundreds of millions of dollars. On August 14, Skilling announced his resignation claiming "personal reasons" for doing this and simultaneously he sold a great deal of his corporate shares and earned approximately $60 million. After Skilling's decision, Enron's stock price was valued below $40 that week and continued its falling to below $30 per share. The only short recovery took place in early October right after Portland General was sold but it did not last for long. Since Skilling resigned, Enron's chairman, Kenneth Lay, realized he had to postpone his retirement and returned as the company's CEO.
A few days after Skilling's resignation, Sherron Watkins, who was an Enron vice president sent a memo to Kenneth Lay expressing her concerns about the company's accounting practices. Particularly, she was skeptical about the lack of disclosure of the substance in transactions that involved the SPEs run by Enron's CFO, Andrew Fastow. Lay informed the company's attorneys, Vinson & Elkins, and its auditing firm Arthur Andersen LLP in order to scrutinize the situation. Vinson & Elkins reported that it was a serious issue but no action was needed because the accounting was acceptable, which makes the ethics of the legal profession issuing an opinion on accounting issues questionable (Cunningham and Harris, 2006). A few days after this report, Kenneth Lay and his wife sold some of their Enron shares, which later on caused them to be charged of inside information trading. Undoubtedly, Enron had started its falling down and it was only a matter of time to entirely crumble.

On October 16, after more than four years, Enron reported its first quarterly loss, which was approximately $1 billion due to underperforming businesses. Then, it put an end to some SPEs operations, like the "Raptor", in order to limit its damage. However, it was another disclosure that alerted the Securities and Exchange Commission (SEC). SEC was alerted when Enron disclosed the reversal of the $1.2 billion entry to assets and equities it had made as a result of dealing with arrangements (Thomas, 2002).

On October 17, it was announced that the company altered its plan for its "employees' 401(k) pension plan". Specifically, it legally locked employees investments for 30 days so that they couldn't sell their Enron stocks. The timing of this decision can be characterized as suspicious, considering the fact that senior management had already sold most of their stocks a few months earlier.

On October 22, it was announced that SEC began examining transactions between Enron and SPEs run by Andrew Fastow, who was fired a couple of days later. On November 8, it was announced that Enron proceeded to the restatement of its financial statements from 1997 in order to include the SPEs that were not involved in these statements. Also, it accomplished some proposed adjustments made by its auditing firm, Arthur Andersen LLP, regarding the same years. It is quite peculiar that these very adjustments had previously been characterized as "immaterial". This restatement resulted in another $591 million in losses over the four years as well as an
additional $628 million in liabilities as of the end of 2000. The equity markets immediately reacted to the restatement, driving the stock price to less than $10 per share (Thomas, 2002). After that, Enron's credit was downgraded and traders and investors lost their trust in it, which demonstrated that the complete demise was inevitable. The following figure displays Enron's stock price during January 2001 and January 2002.

![Enron stock price chart (January 2001 - January 2002)](image)

**Figure 10:** Enron's stock price chart (January 2001 - January 2002).

On November 9, a merger deal with a smaller intrastate competitor, Dynegy, was announced. However, this agreement fell through on November 28. Major credit rating agencies downgraded Enron's debt to junk bond status, making the firm liable to retire $4 billion of its $13 billion debt. Dynegy pulled out of the proposed merger. (Healy and Palepu, 2003). On November 30, Enron's stock was estimated at a shocking 26 cents per share. Naturally, unable to do anything more, Enron filed for bankruptcy protection on December 2. For more details about the chain of events that contributed to Enron's collapse from August 2001 to December 2001 see Appendix C.

Enron's demise caused several results. The following table summarizes these results as interesting bankruptcy facts. It should be stated that the number of employees that lost their jobs includes not only Enron's but also employees from Enron's associated partnerships.
Table 4: Bankruptcy facts (Enron: The Smartest Guys in the Room, 2005).

<table>
<thead>
<tr>
<th>Bankruptcy facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>20,000 employees lost their jobs and medical insurances</td>
</tr>
<tr>
<td>Average severance pay: $4,500</td>
</tr>
<tr>
<td>Top executives were paid bonuses totaling $55 million</td>
</tr>
<tr>
<td>Employees lost $1.2 billion in retirement funds (2001)</td>
</tr>
<tr>
<td>Retirees lost $2 billion in pension funds (2001)</td>
</tr>
<tr>
<td>Enron's top executives cashed in $116 million in stock (2001)</td>
</tr>
</tbody>
</table>

2.4.6 Reasons of Enron's collapse

As it is stated before, after approximately 16 years of "success stories", Enron collapsed within a few months after Skilling's resignation. The main reasons of this event are the complex use of SPEs, the role of deregulation with its political aspects, the manipulation of Enron's financial statements via the use of "mark-to-market" accounting and its accounting regulation and the role of audit committees and external auditors. Additionally, other reasons such as top management compensation, fund managers and sell-side analysts played a crucial role to its demise. The figure below presents Enron's demise.

![Figure 11: Enron's demise.](image)

The use of Special Purpose Entities

Special Purpose Entities were used by Enron in its effort to access capital or hedge risk. As it is mentioned in a previous part, by using SPEs Enron was able to increase
leverage and ROA (Return On Assets) without having to report debt on its balance sheet. Enron exchanged hard assets and liabilities with its SPEs in return for an interest. Then, SPEs could borrow money from financial institutions to buy Enron's assets or deal with other operations without the debt or assets being presented on Enron's financial statements.

Several SPEs were run by Enron's CFO, Andrew Fastow, who inducted even more complex operations with them. These operations were on the limit between legality and illegality regarding their accounting practices. Besides hard assets and liabilities, SPEs were fully exploited by the use of highly composite derivative financial instruments. So, as Enron's business became more intricate, problematic assets that their value was decreased were transferred to SPEs. For example, decreasing assets from certain overseas energy facilities were left to SPEs, so that the losses would not be visible in Enron's books.

Enron used a considerably large number of SPEs by 2001. Initially, they were used in order to make deals with gas producers under long term fixed contracts. However, throughout the years SPEs were created to accomplish financial reporting goals which made them be characterized as controversial. Two of the most notorious SPEs were LJM Cayman LP and LJM2 Co-Investment LP. Andrew Fastow run them and he was paid more than $30 million in management fees in a two-year period (from 1999 to July 2001). Senior management and the board of directors supposedly authorized such a salary, that remarkably exceeded his Enron salary. Apparently, these SPEs and several others (e.g. Chewco, Raptor etc) violated general accounting principles because as it is mentioned before according to FASB guidelines only a 3% of the SPE must be owned by an outside investor in order not to be considered as a subsidiary. Consequently, as Enron revealed in October 2001, it violated this requirement and avoided consolidating them. As a result of not consolidating them, Enron's balance sheet understated its liabilities and overstated its equity and earnings (Healy and Palepu, 2003). After this revelation, the value of these SPEs' assets fell, which meant immense debt for Enron.

Additionally, Cactus Funds that was considered to be a legal and creative solution, it actually concealed risks and debt that was created by VPP contracts from Enron's books. Moreover, executives were able to secrete risks from investors and from the board of directors through derivative contracts.
Consequently, Enron mainly exploited the use of SPEs in order to hide financial statements from its balance sheet. By doing this, it was able to present overstated equity and earnings in cases where losses should be presented instead. However, some of these practices are considered to be legal, which makes us wonder whether ethics play any role in the corporate world.

*The role of deregulation*

Enron was born from deregulation, which initially was referred to natural gas deregulation. It initially caused reduced prices, increased supply and increased volatility in gas prices. Enron offered long-term fixed prices arrangements with producers so that exposure to fluctuation in spot prices would be declined. It also used financial derivatives (e.g. swaps, future contracts etc) and off-balance sheet financing tools. These tools were called Special Purpose Entities and were described previously.

Later on, deregulation expanded to electricity too, which played a significant role to Enron's rise and fall. Many scientists (e.g. Geisst, Rossi, Kuttner etc) strongly believe that deregulation cannot be applied efficiently in electricity market. Particularly, Kuttner has suggested that other fields of the economy can be run as free-market, but electricity cannot. Moreover, he declared that even a well-designed market system creates opportunities for market manipulation (Kuttner, 2002). California electricity crisis in the summer of 2000 is an example that confirmed Kuttner's belief. During this crisis, electricity power plants were intentionally shut down so that an artificial shortage of energy would be created. This unethical practice caused electricity prices to increase dramatically. Enron took advantage of the situation and earned around $2 billion by betting that prices would increase, which of course they did. Thus, this crisis proved that electricity supply and demand are inelastic variables, because electricity is a good that cannot be stored. In deregulated electricity markets, entrepreneurs' goal is to gain the appropriate market power that would provide them price manipulation by discarding spare capacity. As a result, utilities and consumers are prey for traders (E. Moncarz, R. Moncarz, A. Cabello, B. Moncarz, 2006).

In general, several economists, mainly neoclassical, firmly believe that an economy would be benefit without economic regulation. Specifically, they argue that this would bring innovation, increase efficiency and reliability and reduce costs and prices. However, as in most cases theory is different from what is actually applied.
For example, in case of California consumers expected 30 percent reduction in electricity price and increased reliability. However, they actually experienced the significant growth of wholesale markets and the debacle of reliability.

Furthermore, deregulation always has some political implications. In the case of Enron, political aspects arise from the friendship between Enron's chairman, Kenneth Lay, and the Bush family. Lay was a strong supporter of the Republican party and a close friend of George Herbert Walker Bush. Lay supported his friend by making large contributions to his campaign, when he run for president. Later on, Lay became friend with Bush's son, George Walker Bush, who also became the president of the United States of America and Lay contributed with fund raisers in his campaign too. In fact, their relationship was so good that in 2000, when G. W. Bush was already the president, Lay was considered to be a candidate for President's Treasury Secretary, but eventually he was not selected. Thus, considering their close relationship there may be a connection between the fact that SEC (Securities and Exchange Commission) decided to involve in the Enron scandal too late. Also, there may be a link with the fact that federal government did nothing timely in the energy crisis in California in the summer of 2000. Federal government decided to impose federal price controls only after the crisis, when California consumers had already lost a lot of money. Additionally, it imposed the recall of California's democrat governor, Gray Davis, who was ultimately replaced by the republican Arnold Schwarzenegger.

All in all, deregulation can theoretically create an efficient market, but in practice this is not always the case. Specifically in the Enron's case, it created a system so complex that energy traders like Enron Corp. could manipulate supply and price, evade scrutiny, and fleece consumers (Kuttner, 2002). Also, it helped Enron become a market-maker in gas and electricity market, which may not be an appropriate field for deregulation. Finally, deregulation and its political aspects affected Enron's rise and demise.

Manipulation of Enron's financial statements

Enron used a composite business model due to the variety of the products and the diversity of operations that it dealt with. In order to present a promising picture of its performance, it took full advantage of accounting limitations in managing its earnings and balance sheet (Healy and Palepu, 2003). However, there were two challenging problems in Enron's approach, which are listed below.
Firstly, the adoption of "mark-to-market accounting" in trading business was problematic. Once Enron started the trading business it changed the straightforward accounting it used in its natural gas business to mark-to-market accounting. This change was necessary because trading included complex long-term contracts and mark-to-market accounting made Enron able to make forecasts of energy prices and interest rates well into the future. This accounting method instructs companies to adjust their quarterly balance sheets to fair market value once they have outstanding energy-related or other derivative contracts. Thus, when a long-term contract was signed, the present value of future profit was presented as revenues and the present value of the expected costs was recorded. Companies were allowed to develop and use models to estimate these values based on their own assumptions and methods. Consequently, unrealized gains and losses were reported on the balance sheet. In Enron's case, income was estimated as the present value of net future cash flows. However, the viability and the costs of some contracts were not taken into account, which considerably overstated profit. For example, as it is mentioned before, in 2000 Enron signed a 20-year period contract with Blockbuster in order to provide video-on-demand over its broadband network. A few pilot projects were executed and Enron recognized estimated profits more than $110 million based on these projects. Technical viability and market demand aspects were not taken into account and the deal fell through in 2001. The following figure is a poster of a play about Enron that displays the manipulation of its financial statements.

Figure 12: Enron, "A true story of false profits".
Secondly, the fact that Enron built its transactions extensively based on the creation of Special Purpose Entities (SPEs). As it is stated before, Enron used a large number of SPEs in order to accomplish financial reporting goals. Additionally, it provided only minimal disclosure on its relations with the SPEs. Enron made investors believe that transactions with SPEs confined downside risk, but they did not know that SPEs used Enron's stock and financial guarantees to do this and therefore Enron was not protected from downside risk. Furthermore, the fact that SPEs were run by Enron's key employees who were paid remarkably more than they did in Enron created a possible conflict of interest regarding whether they had accomplished their fiduciary responsibility to Enron's shareholders.

Thus, manipulation of Enron's financial statements portrayed a decent picture of its performance. Yet this picture was only superficial. In fact, by using deceptive methods a divergence between economic realities and accounting numbers was created. As it is cited in the documentary film "Enron: The Smartest Guys in the Room" (2005), the numbers just didn't add up, which was the main reason of its collapse. Top executives deceived the market with false reporting and manipulating accounting rules in their transactions due to their obsession with meeting analysts' estimates. Deception led to disclosure restatements of these transactions and revelation of accounting irregularities and business failures in October 2001. This significantly condued to its demise since the market immediately responded to them by dramatically reducing its stock price and increasing its borrowing costs. After a while, investors, traders, banks and generally everybody lost their trust in Enron which was catastrophic for a company that was highly relied on external funding and caused severe lack of liquidity.

Nevertheless, it should be stated that many of these accounting actions were legal. Off-balance sheet accounting is legal, but Enron did more than that. It designed transactions that satisfied the letter of the law, but violated the intent so that its balance sheet did not reflect financial risks (Healy and Palepu, 2003). Once again the lack of ethics in the corporate world is evident.

*The role of audit committees and external auditors*

The fact that Enron's problem remained undetected for many years is quite peculiar. The first question that everybody thinks of is *where were the auditors?* It is their job to examine cases like Enron and detect irregularities. But what happens when auditors
perform consulting and auditing services at the same time? Severe conflict of interest arises. In Enron's case, audit committees failed to recognize irregularities and problems but it was the external auditors who played a crucial role in its demise.

Generally audit committees in the corporate world, are external directors who annually meet a few times and heavily rely on information from management and external auditors. So, if management presents fraudulent financial statements and the auditors fail, they would probably not be able to distinguish the problem timely.

Enron's audit committee was superior comparing to others, because it comprised of remarkably respectable experts in the field (e.g. Dr Robert Jaedicke, accounting professor in Stanford University, Paulo Ferreira, former president and CEO of the State Bank of Rio de Janeiro in Brazil etc). But the meetings were only a few, very short and dealt with a variety of issues which obviously couldn't be examined in depth. So, the committee couldn't predict the accounting irregularities regarding the SPEs and the deceptive management reports. Nevertheless, the audit committee did not challenge several important transactions that were primarily motivated by accounting goals, was not skeptical about potential conflicts in related party transactions and did not require full disclosure of these transactions (Powers, Troubh and Winokur, 2002).

In 2001, Arthur Andersen LLP was considered to be one of the "Big Five" accounting firms and provided auditing, tax and consulting services. Its headquarters was located in Chicago, but it was Houston's office that collaborated with Enron, which was its second biggest client and Andersen performed its external and internal audits. Additionally, it provided Enron consulting services, which indicates that a conflict of interest existed. It is remarkable that in 2000 Arthur Andersen was paid by Enron $25 million for its audit services and $27 million for its consulting services. Consequently, Andersen auditors had conflicted financial incentives whether to maintain and satisfy Enron as a consulting or as an auditing client. Evidently, when its audit team faced with accounting issues chose to ignore them, acquiesced in silence to unsound accounting, or embraced accounting schemes as an advocate for its client (Cunningham and Harris, 2006).

Furthermore, the fact that Andersen's business model shifted from utterly auditing services to consulting, weakened auditors as monitors of management. Employees were not equipped for this change and many young auditors performed
consulting, whilst they did not take into consideration the risk of audit failure. Thus, they lacked expertise to evaluate successfully complex financial statements. Also, they failed to recognize transactions between SPEs that were created for financial reporting manipulation.

Traditionally, auditing companies have strong internal controls. However, Andersen had serious internal control vulnerabilities due to its lax standards. For example, Andersen's Chicago office allowed Houston office to overrule significant reviews of Enron's accounting practices. Its internal control mechanisms should have been reviewed and strengthened because it is a fact that investors lose confidence in accounting firms with weak internal controls. Also, considering the fact that it was involved in another scandal during the same time period and it was operating under a consent order from SEC, it makes no sense why this review was neglected.

Moreover, when SEC started the investigation about Enron, Andersen shredded documents related to its auditing services to Enron in its effort to conceal improprieties. Hence, it was convicted of obstruction of justice on June 15 2002. Although this decision was overturned by the US Supreme Court in 2005, Andersen's credibility was lost and irreparably recovered. Losing credibility meant losing clients as well. Also, the State of Texas removed its license to practice accounting due to professional misconduct. With its reputation of honesty destroyed, Arthur Andersen LLP collapsed, causing 29,000 people to lose their jobs.

Obviously, Enron's debacle was highly associated to the fact that auditors failed to recognize irregularities on time. Even if they did, eventually they succumbed to pressure from Enron's senior management and did not act effectively. Additionally, the fact that Andersen performed consulting services on top of their auditing services caused severe conflict of interest, which its inexperienced employees could not handle it successfully. This led to America's oldest auditing firm collapsing due to this gigantic scandal. Andersen's failure triggered a series of events concerning how auditors were perceived by the public; clients perceived them as an obligation and not as a value added service, investors felt they were not reliable, young professionals found auditing unattractive, regulators and the general public perceived them as confined to their clients.

Other reasons
Enron was a massive failure because of its complex operations and because market controls failed to guarantee integrity in such operations. It was not just executives' fault. Everybody knew what was going on, but they acted on their own best short-term interest.

Enron's senior management was highly compensated in stock options so that their interest would be aligned with its shareholders. However, stock compensation programs that were applied at that time, encouraged executives to make decisions that increased short-term stock performance, but failed to increase medium- or long-term value.

Additionally, institutional investors owned 60% of Enron's stock until early 2001. Some of these prestigious financial firms are Janus Capital Corp., Barclay's Global Investors, Merril Lynch, Morgan Stanley Investment Management, Goldman Sachs Asset Management (involved in a financial scandal in 2008 that triggered global economy's depression) and many others. These institutions did not recognize Enron's problems because they were misled by its accounting methods or by sell-side analysts as its high stock price was a result of unrealistic future performance expectations (mark-to-market accounting). Also, fund managers lacked of strong incentives to demand and act on thorough, long-term company analysis. This lies on the philosophy that not a lot of questions are raised when something is greatly profitable. This is quite ironic considering Enron's slogan "Ask Why", which is presented in the figure below and it was supposed to suggest that Enron questioned conventional wisdom and thus provided innovation. However, its debacle was heavily relied on the fact that no one or very few people actually asked why.

Figure 13: Enron's slogan "ask why".
Moreover, sell-side analysts failed to provide timely warning about Enron's issues. This is mainly because of their financial incentives to suggest Enron to their clients so that their firms' investment deals would be supported. So, analysts faced a conflict of interest when making their forecasts, because they relied on investment bankers at their firms and on Enron's management for inside information. As a result their forecasts could be illustrated as biased. A remarkable fact is that investment banks made more than $125 million from Enron during 1998-2000 and financial analysts of these banks were rewarded with generous bonuses.

Consequently, it is not an exaggeration to state that in Enron's case everybody involved in its transactions is to be blamed. Management failed, auditors failed, analysts failed, creditors/banks failed, investors failed, regulators failed and lawyers failed (Cunningham and Harris, 2006). Massive greed, collusion and lack of ethics characterized all of them, which raise questions about the credibility and the viability of the current capital market.

2.4.7 Federal Investigation

Enron case became the largest and most complex white-collar investigation in FBI history until then and its task force involved prosecutors, agents, analysts and computer scientists. It took them five years of investigation to lead Enron's top executives to jury convictions. Due to the importance of the case, it was assigned to the highest levels of the FBI, the Department of Justice and the Securities and Exchange Commission (SEC). The figure below implies the federal investigation about Enron.

Figure 14: Federal investigation about Enron.
In January 2002, agents investigated Enron's building for nine days and ferreted more than 500 boxes of evidence, which included critical documents. Additionally, more than 100 interviews were conducted by agents. In February 2002, Enron's board of directors assigned to a special investigation committee to conduct its own internal investigation (Powers Report), which concluded that accounting principles were violated and top executives enriched themselves through this violation. This report significantly helped FBI during the investigation and agents continued interviewing people in the USA and overseas. Moreover, financial analysts scrutinized bank and brokerage accounts to discover fraudulent purchases, which consolidated the argument of insider trading charges. Furthermore, agents with expertise in digital forensics gathered over four terabytes of data. At the same time, the Regional Computer Forensics Laboratory (RCFL) in Houston constituted the "Enron Task Force" and processed 31 terabytes of data that included 2,300 pieces of evidence, 600 employee emails, 130 computers, 10 million pages of documents, more than 3,000 outlook email boxes and 4,500 lotus notes email boxes. It should be pointed out that a terabyte is equivalent to 250 million pages of text, which makes us realize how difficult RCLF's job was in the investigation.

After the investigation, FBI concluded that Enron's executives overstated its expecting profits from businesses like the broadband trading in order to boost its stock price. Also, international assets were overvalued and manipulation of its financial statements was discovered. Powers report was a very helpful tool to FBI agents unearthing these results.

### 2.4.8 Convictions

After a five-year investigation conducted by FBI, Enron's former top executives were led to jury convictions. The main actors in these convictions were Jeffrey Skilling, Kenneth Lay, Andrew Fastow, Rick Cousey and Ken Rice. Other than Enron's executives, employees from companies that were related to Enron and also convicted will not be described in this dissertation.

Jeffrey Skilling was indicted on 35 counts of several financial crimes such as bank fraud, making false statements to banks and auditors, securities fraud, wire fraud, money laundering, conspiracy and insider trading. He pleaded not guilty to all charges. The trial began on January 30, 2006 in Houston and on May 25, 2006 the
jury decided that Skilling was guilty on the following counts; on one count of conspiracy, on one count of insider trading, on five counts of making false statements to auditors and on twelve counts of securities fraud. It was announced that he was not guilty on nine counts of insider trading. Five months later, on October 23 he was sentenced to 24 years and four months in prison and he was enforced to pay $45 million as a fine. However, on June 21, 2013 a sentencing deal that reduced Skilling's sentence by 10 years was announced. Under the deal, more than $40 million of Skilling's fortune, which has been frozen since his conviction in 2006, will be distributed to victims of Enron's collapse. The figure below presents a tweet from CNBC that announces Skilling's new sentence deal.

![CNBC](image)

**Figure 15:** A tweet from CNBC regarding Skilling's new sentence deal.

Kenneth Lay was indicted on 11 counts of several financial crimes including securities fraud, wire fraud and making misleading statements. Lay pleaded not guilty to all charges and he cited that he was misguided by top executives suggesting that Andrew Fastow was the main responsible for Enron's dissolution. He was found guilty on ten counts of securities and wire fraud and he was subjected to a maximum total sentence of 45 years in prison and approximately $90 million compensation as civil fines. However, Lay died on July 5, 2006 due to heart attack and on October 17, 2006 his convictions were abated. Although, Lay's wife Linda sold 500,000 shares of Enron shares a few days before Enron filed for bankruptcy, she was never indicted on any charges.

Andrew Fastow was indicted on 98 counts of financial crimes including money laundering, insider trading and conspiracy. On January 14, 2004 Fastow pleaded guilty to two counts of conspiracy and he agreed to cooperate with federal authorities in the prosecutions of other former Enron executives (Kenneth Lay, Jeffrey Skilling, Rick Cousey) in order to receive a reduced sentence. Therefore, he was
sentenced to 10 years in prison and he was enforced to pay a fine of $23.8 million. Because of his great cooperation with the authorities, Fastow's sentence was reduced to six years and he was eventually released on December 16, 2011. Fastow's wife Lea, who was an assistant treasurer in Enron, was also indicted on misdemeanor tax charge. She pleaded guilty and she was sentenced to one year in prison and an additional year of supervised release.

Rick Cousey was Enron's Chief Accounting Officer. He was indicted on six felony counts for manipulating Enron's financial condition. Although he initially pleaded not guilty, he changed to guilty and he was convicted to seven years in prison.

Ken Rice was the CEO of Enron Broadband and was indicted and pleaded guilty to securities fraud in broadband case. He also agreed to cooperate with federal prosecutors and he testified in the trial of Skilling and Lay. Because of his help in Skilling's and Lay's cases, he received a 27-month sentence in June 2007 and he was ordered to pay an additional $50,000 fine. It should be mentioned that he had already forfeited $13.8 in cash and property.

2.4.9 Compensations
It is estimated that Enron's shareholders lost nearly $74 billion in the four years prior to its bankruptcy. Out of this huge amount of money, around $40 to $45 billion loss was due to the fraud. Additionally, Enron was in debt of approximately $67 billion because of liabilities to its creditors, employees and its shareholders. In order to pay a part of its debt, Enron conducted auctions where it sold its assets such as logo, art, pipelines and other.

Moreover, former Enron's employees had lost nearly $2 billion from their pension funds. In May 2004, almost 20,000 Enron's retirees won a suit of $85 million compensation and each one was paid about $3,100. In 2005, investors were given a settlement of $4.2 from several financial institutions that were related to Enron's businesses.

Furthermore, shareholders and investors had filed a $40 billion lawsuit against financial institutions such as Bank of America, JPMorgan Chase & Co., Citigroup and others accusing them of actively participating in the accounting fraud that caused Enron to demise. In September 2008, a settlement of $7.2 billion was reached and
dispensed among the main plaintiffs, which were the University of California (UC) and 1.5 million individuals and groups. It is remarkable that UC's law firm received almost $688 million plus interest in attorneys fees, which is considered to be as the highest attorneys fees in US securities fraud cases.

Finally, after having described convictions and compensations, it seems significant to present the following figure, which is an ironic depiction of the aforementioned Enron's slogan, "ask why". Maybe if more people indeed asked more questions about the events included in the picture, the outcome could have been different.

Figure 16: Ironic depiction of Enron's slogan "ask why".
3 Problem Definition

This chapter, Problem definition, includes three sections, which are The Enron email dataset, Related work on email analysis and Dissertation’s problem definition.

3.1 The Enron email dataset

Emails have been characterized as a valuable resource for several research fields, such as textual and social network analysis. This is because emails have replaced the letter as the main written communication medium, both in business and in personal life (Haggerty, Karran, Lamb, Taylor, 2011). It is pointed out that the average employee sends approximately 43 emails every day. Hence, emails can provide a great deal of information and therefore investigators rely heavily on them throughout their investigations. Such a case is the Enron email dataset, which was used by investigators so that fraud is detected.

During the investigation of the Federal Energy Regulatory Commission (FERC), the Enron email dataset was publicly presented. Initially, it had several integrity problems, which were successfully solved by SRI-Artificial Intelligence Center researchers. After that, William Cohen, a professor at Carnegie Mellon University (CMU), posted the dataset on the web, so that researchers would be benefit. This version of the dataset is online on http://www-2.cs.cmu.edu/~enron/. It includes both personal and official emails from 3,500 folders and it doesn't include any attachments. Some messages were deleted from the dataset after requests from affected employees.

Obviously, the dataset is huge and this is why federal investigation lasted almost five years. It is extremely difficult to analyze such a large dataset and find significantly relevant information. Furthermore, the fact that emails contain both structured and unstructured data made things even worse. Emails are structured due to their format (RFCs 5321 and 5322). Structured data provide qualitative information to the forensics examiner and it is usually analyzed based on existing tools. Nevertheless, they also provide unstructured information due to their textual form of data. The unstructured nature of textual data makes it more complex to examine because it contains information associated with communication links and contacts that
form social networks. Also, relationships within the networks and identification of key actors are included into the social networks. Social network analysis with the aid of visualization techniques play a crucial role in evidence detection and collection as it provides a deeper insight of relationships between actors and their connection to events. Unfortunately, there are currently no standardized tools for the forensics analysis of unstructured data, but there several techniques that can be used for this.

3.2 Related work on email analysis

A lot of studies have been conducted so that the Enron email corpus would be fully examined. Each of these studies have proposed different approaches to deal with the analysis. For example, statistical analysis, data mining techniques and clustering algorithms to name a few.

As it is cited before, for the structured data analysis existing forensics tools are used. Examiners use computer forensics tools, such as EnCase and Forensics Toolkit (FTK) in order to reconstruct files and data from a suspect's computer. These tools are only used for structured data analysis and thus neither visualization of the evidence is provided nor evaluation of the importance of actors. However, Haggerty et al. (2011) proposed a tool that would automatically extract and visualize email data that exist in the hard drive. This tool was named Email Extraction Tool (EET).

In 2004, Bryan Klimt and Yiming Yang conducted email classification for the Enron dataset. Their goal was to explore how to classify messages as organized by a human. In order to accomplish it, SMV (Support Vector Machine) classifier was used and also they cleaned the data from duplicate messages. Moreover, in 2005, Jitesh Shetty and Jafar Adibi created a MySQL database for the dataset and they statistically analyzed it. Additionally, they derived a social network from the dataset and presented a graph of it as a Gower layout. Data cleaning was conducted in this paper as well. This cleaned version of the dataset is electronically available in http://www.isi.edu/~adibi/Enron/Enron.htm and it is used in this dissertation in order to perform the dataset analysis.

As for the text mining part, which is the extraction of knowledge from text, there are several tools proposed. For instance, Email Mining Tool (EMT), that was proposed by Li et al (2008), analyzes email archives which includes a graphical display to explore relationships between actors. Additionally, R. Al-Zaidy et al.
(2012) proposed a data mining algorithm to discover criminal networks from a collection of text documents and visualize this network. This paper also used Enron email corpus as a case study of real-life cybercrime.

Furthermore, a lot of research has been made in the field of Social Network Analysis (SNA). Haggerty et al. (2011) suggested a framework for the analysis of unstructured data, which focused on the triage and analysis of it so that key actors and relationships between them would be discovered. Also, social network and visualization techniques were applied. Enron email dataset was used to test the applicability of the approach. Moreover, SNA tools, like NetMap and Agna are developed to help investigators. SNA applications that visualize network information based on graph theory were created. Such applications are Pajek, Social Networks Visualizer (SocNetV) and Network Workbench (NWB). In addition, Chapanond et al. (2005) analyzed the Enron email dataset based on graph theoretical and spectral analysis. Finally, Matsuyama and Terano (2008) examined Enron email dataset based on an agent-based model.

3.3 Dissertation's problem definition

In this dissertation the main goal was basically to visualize the Enron email dataset and generate some useful statistics. In order to accomplish this goal several software technologies were used. These technologies will be listed below and they will be thoroughly described in the next section. It is important to be mentioned that the dataset used here is the dataset created by Jitesh Shetty and Jafar Adibi in 2005. This version of the dataset differs from the original one generated by William Cohen, as it is cleaned from duplicated messages. Additionally, in this section the problems faced when analyzing the Enron email dataset will be presented.

Firstly, this huge dataset was downloaded from the link mentioned before and then it was split in tables stored in a MySQL database. Secondly, with the aid of Python and MySQL the database was processed in order to provide meaningful results. Thirdly, Python was used in conjunction with two social network visualization tools, which are Social Networks Visualizer (SocNetV) and Network Workbench (NWB). The former is a flexible and user-friendly tool for the analysis and visualization of social networks that generates several useful network analysis statistics. The latter is a network analysis, modeling, and visualization toolkit mainly
for scientific research. This particular software was remarkably efficient in cases where a huge amount of data had to be processed and visualized. Consequently, in this dissertation the statistical analysis and the social network analysis of the Enron email dataset will be examined, but not the text mining part of the emails. Finally, some useful statistics were generated by using Python and MySQL (filter the database) and these statistics were visualized into figures and plots by using Microsoft Office Excel 2007.

Consequently, before starting describing the problems faced during the implementation part, it should be highlighted that this dissertation aims to provide the methodology that can be considered as the foundation when analyzing such large datasets. Furthermore, its goal is to act as a toolkit for further analysis that can be conducted in the future.

There were several problems while analyzing the Enron email dataset. However, the most significant one was undoubtedly the extremely large size of the dataset. In 2005, Jitesh Shetty and Jafar Adibi cleaned the data of the Enron email dataset and stored it in a MySQL database consisting of four tables (employeelist, message, recipientinfo and referenceinfo). However, this database was extremely large and it was very difficult to handle. The first step was to create a database and the four tables in MySQL environment. Then, after extracting each table of the database in a distinct file, it was attempted to store the data of each file in the corresponding MySQL table. This proved to be the first major problem due to the size of the tables. Particularly, the size of file "message" was huge and it couldn't be stored in MySQL without breaking it. Hence, a software (HJSplit) was used to break it into four smaller files in order to be able to store it. Another problem, encountered was the fact that the size of the files made it impossible to store them into the MySQL database through "MySQL Workbench" and therefore the command prompt was used instead.

Additionally, some problems occurred when using Python programming language to process the MySQL database. Again the size of the database made the execution of the code to run significantly slow. Therefore, some techniques to improve the efficiency of the program were used. Moreover, besides the size of the database there were some other unexpected problems when using Python. Particularly, it was quite time consuming to install some Python libraries that were required in order to proceed with the programming part of the dissertation.
Specifically, the libraries that created problems in their installation were "MySQLdb", "numpy" and "matplotlib". The first one is a Python DB API-2.0-compliant interface, that is used in order to connect to MySQL and process databases through Python. The second one is a package for scientific computing with Python and the third one is a Python 2D plotting library that helps programmers generate figures, plots, histograms, etc. Additionally, some libraries, such as "networkx" and "scipy", require the aforementioned libraries to be installed in the system before their own installation, which made things even worse. Finally, after several days of searching these problems, it was realized that the problem was basically because the computer used for the programming part is a 64-bit Windows, while these libraries are mostly well-defined for 32-bit computers. After a lot of searching on the internet the problem was solved.
4 Contribution

This chapter is called Contribution and it includes the implementation part of the dissertation. It consists of three sections, which are Methodology and technologies, Statistical analysis, and Social network visualization.

4.1 Methodology and technologies

The goal of this dissertation was mainly to produce a visualization of the Enron email dataset and generate some helpful statistics as well. A specific methodology and several technologies were used in order to accomplish this goal. In this section, the methodology and the technologies that were used before the statistical analysis and the visualization of the data will be fully described. As for the tools used in the statistical analysis and in the visualization process will be described in the next sections.

The very first step was to download the dataset from this link http://www.isi.edu/~adibi/Enron/Enron.htm. This dataset is a .sql file named "enron-mysqldump". It contained four tables, namely "employeelist", "message", "recipientinfo" and "referenceinfo". Secondly, before even installing MySQL, a simple code of Python was used in order to split the huge file (750MB) into these four distinct files (each file relates to a database table). Here is a screen shot of this code:

```python
import sys, os

openflag = False
infile = open("c:/test/enron-mysqldump.sql", 'r')

for line in infile.readlines():
    if line[:12] == "CREATE TABLE":
        tablename = line[13:line.index("(")-1]
        print "Found table:"+tablename+"..."
        if openflag:
            outfile.close()
            openflag = True
        outfile = open(outfilename+"."+tablename+".sql"
        drop = "DROP TABLE IF EXISTS "+tablename+";\n"
        outfile = open(outfile, "w")
        outfile.write(drop);
        if openflag:
            outfile.write(line)
        if openflag:
            outfile.close()
    if openflag:
        outfile.close()
    infile.close()
```

Figure 17: Screen shot of the code that split the dataset into the tables.
It should be cited that the version of Python used in this dissertation is Python 2.6. Although there are several upgraded versions (up to Python 3.3.2), Python 2.6 was selected because it is supported by a great deal of tutorials and it is considered to have a higher compatibility with more libraries than the latest versions. Initially, Python 3.3.2 was used, but due to some problems it was given up. Python is an open source software, that was downloaded from its official site http://python.org/download/. Also, it is an object-oriented language and it is considered to be a scripting language, which enables higher programmer productivity and improved code readability. Hence, since there was no advanced command of any other languages, Python was selected based on these characteristics.

After dividing the dataset into four files the next step was to create a local database in order to store and process the dataset. Therefore, MySQL 5.6 was downloaded from the official site of it http://www.mysql.com/downloads/. Then, the database "enron" was locally created by using MySQL commands. After that, tables "employeelist", "message", "recipientinfo" and "referenceinfo" were also produced in the MySQL 5.6 environment. The next step was to store the data of each file in the corresponding MySQL table. As it is mentioned before, due to the extremely large size of file "message", it couldn’t be stored in the database and therefore HJSplit 3.0 was used in order to split the file into four smaller ones. This particular software is also an open source software that was downloaded from its official page, http://www.hjsplit.org/windows/. Then, each file was successfully stored in table "message". Hence, after that database "enron" consists of four tables, namely, "employeelist", "message", "recipientinfo" and "referenceinfo". The following picture shows the what it is included in each table of database "enron".
At this point it is significant to describe the elements of each table. As it is shown in the above picture, table "employeelist" includes the following four fields:

- "eid", which is the employee id (primary key, auto increment),
- "firstName", which refers to the first name of the employee,
- "lastName", which refers to the last name of the employee,
- "Email_id", which is the email address of the employee,

Furthermore, table "message" contains seven fields:

- "mid", which is the message id (primary key),
- "sender", which refers to the sender's email address,
- "date", which is the full date of when the email was sent,
- "message_id", which is the internal message id from the mailserver,
- "subject", which refers to the subject of the email,
"body", which refers to the email body,
"folder", which refers to the email folder.

Moreover, table "recipientinfo" includes five fields:

- "rid", which is the recipientinfo id (primary key, auto increment)
- "mid", which is the message id (foreign key),
- "rtype", which gives us information about how the message was received by the recipient ("TO", "CC", "BCC"),
- "rvalue", which refers to the receiver's email address,
- "dater", which is a NULL field in all records. Probably, it shouldn't be in the database at all.

Finally, table "referenceinfo" includes three fields:

- "rfid", which is the referenceinfo id (primary key),
- "mid", which is the message id (foreign key),
- "reference", which relates to the part of the email that was forwarded or replied.

The following picture shows how many records are included in each table.

```
mysql> select count(*)
    -> from employeeList;
+----------+
<table>
<thead>
<tr>
<th>count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
</tr>
</tbody>
</table>
+----------+
1 row in set <0.05 sec>

mysql> select count(*)
    -> from message;
+----------+
<table>
<thead>
<tr>
<th>count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>252757</td>
</tr>
</tbody>
</table>
+----------+
1 row in set <14.76 sec>

mysql> select count(*)
    -> from referenceInfo;
+----------+
<table>
<thead>
<tr>
<th>count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>54778</td>
</tr>
</tbody>
</table>
+----------+
1 row in set <3.41 sec>

mysql> select count(*)
    -> from recipientInfo;
+----------+
<table>
<thead>
<tr>
<th>count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2064442</td>
</tr>
</tbody>
</table>
+----------+
1 row in set <1.59 sec>
```

Figure 19: The number of records included in each table.
As it is presented in the above picture, table "employeelist" contains 151 records. This means that the only 151 Enron's employees agreed to participate and have their data published in the dataset. The rest of the emails are exchanged either by Enron's employees who did not want to participate in this dataset for privacy or other reasons, or by people outside the company. There is an online list of these 151 employees that contains their status in Enron. This can be found on http://www.isi.edu/~adibi/Enron/Enron.htm.

After that, two more tables were created by using Python environment. The first one is called "output" and it contains the joined records "mid", "sender", "rvalue" and "date" from tables "message" and "recipientinfo", since these two tables share a common key ("mid"). It should be clarified that for each message, therefore for each "mid", there is only one sender, but there can be several receivers ("rvalue"). The two following pictures show the creation of table "output" and the insertion of values into it. The connection and the disconnection of the database and the creation of the cursor is only shown in the first picture since it is the same in both cases.

```python
import MySQLdb
import datetime

db = MySQLdb.connect(host = "localhost",
                     user = "root",
                     passwd = "root",
                     db = "enron")

cur = db.cursor()

cur.execute("''DROP TABLE IF EXISTS output'''")

cur.execute("''CREATE TABLE IF NOT EXISTS output (mid INT(10) NOT NULL,
                                                  sender VARCHAR(127) NOT NULL,
                                                  receiver VARCHAR(127),
                                                  date DATETIME)
            '''")

cur.close()
db.close()
```

Figure 20: Creation of table "output".
The second table that it was created is called "result" and it contains the values of "output" plus the id of senders ("senID") and the id of receivers ("recID") retrieved by table "employeelist" ("E_mail_id"). Both of these values are defined as INT(10). It is important to be mentioned that senders and receivers that are not included in the table "employeelist" have their "senID" and "recID" values set to 0 for convenient handling. The creation of table "result" is similar to the creation of table "output" and therefore it will not be displayed. The following image shows the part of code that inserts values into table "result".

```python
sql = """SELECT message.mid, sender, date, rvalue FROM message, recipientinfo WHERE message.mid = recipientinfo.mid""
try:
cur.execute(sql)
results = cur.fetchall()
for row in results:
    cur.execute("""INSERT INTO output (mid, sender, receiver, date) VALUES (%s,%s,%s,%s)"" % (row[0],row[1],row[3], row[2]))
db.commit()
except MySQLdb.Error, e:
    try:
        print "MySQL Error [%d]: %s" % (e.args[0], e.args[1])
    except IndexError:
        print "MySQL Error: %s" % str(e)
```

Figure 21: Insert values into table "output".
sql2 = """SELECT mid, sender, receiver, date FROM output"""

try:
    cur.execute(sql2)
    results = cur.fetchall()
    numrows = cur.rowcount
    for item in results:
        mid = item[0]
        sender = item[1]
        receiver = item[2]
        date = item[3]

        list_mid.append(mid)
        list_sender.append(sender)
        list_receiver.append(receiver)
        list_date.append(date)

    for i in range(numrows):
        if (list_sender[i] in list_email):
            x = list_email.index(list_sender[i])
            y = list_email.index(list_receiver[i])
            cur.execute("""INSERT INTO result 
VALUES (%s,%s,%s,%s,%s,%s)""" %
                (list_mid[i], list_sender[i], list_email[x],
                list_receiver[i], list_email[y], list_date[i]))
        elif (list_sender[i] in list_email):
            value = 0
            cur.execute("""INSERT INTO result 
VALUES (%s,%s,%s,%s,%s,%s)""" %
                (list_mid[i], list_sender[i], list_email[x],
                list_receiver[i], value, list_date[i]))
        elif (list_receiver[i] in list_email):
            y = list_email.index(list_receiver[i])
            value = 0
            cur.execute("""INSERT INTO result 
VALUES (%s,%s,%s,%s,%s,%s)""" %
                (list_mid[i], list_sender[i], value, list_receiver[i],
                list_email[y], list_date[i]))
        else:
            value = 0
            cur.execute("""INSERT INTO result 
VALUES (%s,%s,%s,%s,%s,%s)""" %
                (list_mid[i], list_sender[i], value, list_receiver[i],
                value, list_date[i]))

Figure 22: Insert values to table "result".
Consequently, database "enron" now consists of 6 tables and from now on mostly table "result" will be used in the processing of the data. Here is a screen shot of MySQL, which shows the tables included in "enron".

![MySQL screen shot showing tables](image)

Figure 23: Tables included in database "enron".

### 4.2 Statistical analysis

Statistical analysis of the Enron email dataset was not the primary purpose of this dissertation. However, after some time it was believed that some simple statistical analysis of the data would be very useful in order to have a better and deeper understanding of the dataset. Additionally, it should be pointed out that the figures provided in this section are only indicative in our attempt to comprehend the Enron email dataset. In the "Future work" section of the dissertation another approach is suggested.

In order to perform statistical analysis, Python code was used once again. Particularly Python was used along with MySQL to count the messages exchanged between senders and receivers listed in the dataset (not only those in table "employeelist"). Successively, the number of messages counted from the Python code was then inserted into Microsoft Office Excel 2007 in order to produce some figures about them. It should be noticed that in 1997 only 3 emails were exchanged and therefore this year will not be presented in a separate figure. Also, the Python code for each year is very similar in most cases and therefore only the MySQL commands will be presented when necessary. The following picture presents the part of code for generating the total number of emails per month during 1998.
for month in range(1,13):
    try:
        cur.execute("""SELECT COUNT(mid)
        FROM result
        WHERE year(date) = 1998 AND month(date)=%s
        GROUP BY month(date)""", (month))
        results = cur.fetchone()
        print (results)
    
db.commit()
    except MySQLdb.Error, e:
        try:
            print ("MySQL Error [%d]: %s  ") (e.args[0], e.args[1]))
        except IndexError:
            print ("MySQL Error: %s ") str(e)

Figure 24: Code for counting emails (1998).

After executing the above code, Python gave us the total number of emails exchanged in 1998 per month. These numbers were then inserted into Microsoft Office Excel 2007 and the following figure was created. As it is shown in the figure below, in 1998 only some hundred of emails were exchanged (271) and the majority of them were exchanged in the last months of this particular year (around 68%). Obviously, this is because emails was not one of the most frequently used technologies back then. This is why for many months in a row in 1998 there were no emails sent or received.

Figure 25: Total number of emails in 1998 per month.
Similarly, for the year 1999 almost the same code was applied and therefore it will not be presented. The total number of emails in 1999 is 22,631. The following figure shows the distribution of the emails every month. The number of emails increases as throughout the year reaching its peak in December (7,171). Approximately, 32% of the emails in 1999 were exchanged in December. However, there are two months (March and November) in which the number of emails is smaller than the one of the previous month.

![Total emails (1999)](image)

Figure 26: Total number of emails in 1999 per month.

Successively, for the year 2000 the number of emails exchanged is approximately more than four times bigger than the number of emails exchanged in the previous year. Specifically, the total number of emails exchanged in 2000 is 512,573. Hence, besides the total number of emails, another parameter is used to filter the data. This parameter is whether emails where exchanged between working hours or not. It should be cited that working hours are considered the hours between 09:00 am to 17:00 pm. Python code for this is shown below.
```
start_work = "09:00:00"
finish_work = "17:00:00"

for month in range(1,13):
    try:
        cur.execute("""SELECT COUNT(mid)
        FROM result
        WHERE year(date) = 2000 AND month(date)=%s
        AND time(date) NOT BETWEEN %s AND %s
        GROUP BY month(date)""", (month, start_work, finish_work))

        results = cur.fetchone()
        print (results)

        db.commit()
    except MySQLdb.Error, e:
        try:
            print ("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
        except IndexError:
            print ("MySQL Error: %s" % str(a))
```

Figure 27: Code for counting emails exchanged outside working hours range (2000).

The same code with a slight change in the MySQL command (without NOT) is used for calculating the emails exchanged within working hours. The total number of emails (without filtering) is estimated as previously. The following figure presents the total number of emails, the number of emails within working hours and the number of emails outside working hours.

According to the figure below in 2000 a significantly larger number of emails was exchanged outside working hours range. Specifically, approximately 75% of the emails exchanged in 2000 were exchanged outside the working hours range. Of course nothing can be proved based on this, but it is a little suspicious that most of the emails were exchanged by employees when they were not in their offices. Again the highest number of emails is exchanged in December and it reached 102,228. Probably, this is because throughout the year emails had been using more and more. Also, when using the working range filter the highest number of emails is exchanged in December as well (working hours: 20,178, not working hours: 82,050).
In 2001, there is the all-time highest number of emails exchanged, as it is doubled compared to the previous year. Particularly, the total number of emails exchanged this year was 1,350,785. At this point, emails had become more prevalent as a technology in the corporate world. The figure below shows the percentages of total exchanged emails per month, without any filtering. Python code was the same as before.

Figure 29: Percentages of total exchanged emails per month.
Based on the above figure, it is obvious that the majority of emails were exchanged in October 2001 (around 20.5% of the total emails of the year). Then 14.9% of emails were exchanged in November, 10.7% in April, 10.8% in May and 8% in March. In the following paragraphs it will be examined whether the above percentages are connected to key events that occurred in Enron at that time. However, before starting describing the connection between the number of emails and critical events occurred during 2001, it should be pointed out that also the same filtering as in 2000 is used here as well (working and not working hours). Hence the code is the same as in 2000 (only the year was changed). Consequently, the following figure presents the total number of emails, the number of emails within working hours and the number of emails outside working hours per month. As it is displayed below, the number of emails exchanged outside the working hours range is larger than those within the working hours range in the first semester of 2001. On the contrary, during the second semester of this year the number of emails exchanged within working hours is higher.

![Email statistics (2001)](image)

Figure 30: Statistics (2001).

In 2001 several changes were brought in Enron and it will be examined if the above figure confirms these changes. The first critical event occurred in February 2001, when Kenneth Lay announced his retirement and Jeffrey Skilling was assigned the positions of both president and CEO of Enron. However, this is not depicted in the
figure above since only 5.1% of the total emails were sent or received during this month. In other words, there is only a 7.2% increase in the number of emails exchanged in February compared to January. The second critical event occurred in March 2001, when it was announced that Enron and Blockbuster video-on-demand deal was cancelled. This event is obviously depicted in the number of exchanged emails, since the increase in the number of emails compared to the previous month is approximately 56%. Also, emails exchanged throughout March constitute 8% of the total number of emails of the year. Another significant increase in the number of emails is evident in April 2001 (almost 34% compared to March). This could be assigned to the publicly despondent and nervous behavior of Skilling which had a tremendous impact on Enron's stock.

After May 2001, the number of emails kept decreasing until September of the same year. It is quite peculiar that in August 2001, when Skilling's resignation was announced, there is no increase in the number of emails and only 4.1% of the total emails were exchanged in this month. Probably, this event is more depicted in the next month, since there is an almost 52% increase in the number of emails in September compared to August. Specifically 6.3% of the total emails of the year were exchanged during September.

In October 2001, there was a dramatic increase in the number of emails (224% compared to September). This could be justified, since there are several critical events occurred during this month. Firstly, Enron announced its first quarterly loss after more than four years. Secondly, the Securities and Exchange Commission (SEC) started an investigation about Enron's financial disclosures. Thirdly, it was announced that Enron legally locked employees investments for 30 days so that they couldn't sell their Enron stocks. Finally, Andrew Fastow, who was Enron's Chief Financial Officer, was fired during this month. After all these extremely critical events, it is not a surprise that the number of emails was tremendously increased during October and reached 276,699 emails, which constitutes 20.5% of the total emails of the year.

In November 2001, there was a decrease in the number of emails compared to the previous month. However, during November 201,020 emails were exchanged (14.9% of the total yearly emails), which is the second highest aggregated number of emails on a monthly basis. This large number of emails is justified because some significant events took place during this month as well. Firstly, Enron announced the
restatement of its financial statements of the previous years. Secondly, the merger
deal with Dynegy was announced and cancelled within the same month. Thirdly,
major credit rating agencies downgraded Enron's debt to junk bond status.
Consequently, the large number of emails during November is completely justified.
Finally, during December there is a huge decrease in the number of emails. This
decrease is around 76% and it is justified since Enron filed for bankruptcy on

Although Enron had filed for bankruptcy in December 2001, there is a
significant number of emails exchanged throughout 2002. Particularly, there are
177,757 exchanged emails during this year. Most of them were sent or received
during the January (around 42%) and February (around 45%). Naturally, after the first
two months, the number of emails follows a descending pattern and after August it
reaches zero values. The following figure shows the distribution of emails during
2002.

![Total emails (2002)](image)

Figure 31: Total number of emails in 2002 per month.

Moreover, another interesting filtering is whether emails were exchanged
between employees included in the table "employeelist" or not. As it is mentioned
before, senders and receivers that are not included in the table "employeelist" have
their "senID" and "recID" values set to 0 in table "result". So, the following Python
code prints the counted emails per year per month only for emails exchanged among
people in table "employeelist". The connection and the disconnection to the database
and the creation of the cursor is not shown below, since it is presented in another screen shot before.

```python
def print_emails_per_year_month(year, month):
    cur.execute("SELECT COUNT(mid)
                FROM result
                WHERE (senID<>0 OR recID<>0)
                AND (year(date)=%s)
                AND (month(date)=%s)
                GROUP BY month(date)",
                (year[y], month[i]))
    result = cur.fetchall()
    for row in result:
        print (row[0])
```

Figure 32: Code that prints the counted emails per year per month only for emails exchanged in table "employeelist".

The code that was used to produce the total number of emails exchanged per year per month will not be presented, since it is similar to this one without the filtering ("senID<>0 OR recID<>0"). Consequently, after inserting all the data into Microsoft Office Excel 2007, the following figure that shows the total number of emails and the filtered emails is produced.
The above figure converted to a yearly basis is the following:

Both figures show that most of the emails were exchanged among people who were not included in table "employeelist". Specifically, only 29% of the total emails are exchanged between employees included in "employeelist". This could be easily predicted since "employeelist" only contains 151 Enron's employees who agreed to participate and have their data published in the dataset.
Finally, the statistics of two employees from "employeelist" will be presented. These employees are Jeffrey Skilling and Kenneth Lay, who were two of the key players in the Enron scandal. Unfortunately, Andrew Fastow was not in "employeelist" and therefore he's statistics will not be presented. Python code will not be shown here, since it is very similar with the previous one, with some changes in "senID" and "recID" values. Skilling's id in the "employeelist" is 129, and Lay's 127 respectively.

![Jeffrey Skilling and Kenneth Lay statistics.](image)

These two figures above show the statistics of Jeffrey Skilling and Kenneth Lay respectively. They show the distribution of sent, received and total emails for each one. Both figures follow a similar pattern, which is actually the pattern of the total emails. In the early years only a few or even no emails were exchanged. After that there is an increasing pattern that reaches its peak in both cases in the late months of 2001 and successively there is a descending pattern after that. However, Kenneth Lay sent more emails (almost double) than Jeffrey Skilling. Particularly, Lay's total emails are 7,761 and Skilling's 3,934 respectively. Furthermore, there are some interesting facts concerning the statistics of these two Enron's executives. Firstly, Kenneth Lay sent and received 65% of his total emails in August 2001. During this month, Skilling resigned and Lay replaced him. Specifically, Lay sent 5,047 emails in August 2001 and these emails probably announced his decision to replace Skilling. Secondly, Skilling exchanged most of his emails (around 40%) in May 2001 (no major event recorded) and in August 2001 (Skilling's resignation).
4.3 Social network visualization

In order to perform social network visualization Network Workbench (NWB) was used. This software is a toolkit for large-scale network analysis, modeling and visualization. It was designed in order to provide an extremely helpful tool for network science researchers in many disciplines, such as biology, physics, computer science, bioinformatics and many others. Also, it was designed with efficiency in mind, since its goal is to provide efficient results so that scientific discovery can be accelerated. NWB was created after the cooperation of Indiana University, Northeastern University and University of Michigan. Moreover, NWB was built on Cyberinfrastructure Shell (CIShell), which is an open source software for integration and utilization of datasets, algorithms, tools, and computing resources. Additionally, NWB is an open source software as well and it was downloaded from its official web page, http://nwb.cns.iu.edu/download.html.

NWB is able to load, process and save files of seven different formats, which are NWB, GraphML, Pajek .net, Pajek .matrix, XGMML, TreeML, CSV and automatic conversion between these formats is supported. In this dissertation the format that it will be used is GraphML. The GraphML file format uses .graphml extension and it is an XML-based file format for graphs. The reason for selecting this kind of format is that it is also supported by NetworkX, which is also used in this dissertation. NetworkX is a Python library for creating, manipulating and studying complex graphs and networks.

In addition, SocNetV 0.81 was used as well. Social Networks Visualizer (SocNetV) is a cross-platform tool that enables the analysis and the visualization of social networks so that the researcher would be benefit. It is a user-friendly software that supports networks of various formats (GraphML, GraphViz, Adjacency (Sociomatrix), Pajek, UCINET, etc). Also, it is an open source software that it was downloaded from its official web page http://socnetv.sourceforge.net/downloads.html.

So, one might wonder why two similar social network tools are used in this dissertation instead only one of them. The answer is because both of them have their weak and strong points and therefore when using both of them it is easier to fully exploit their strong points and eliminate their weaknesses. Specifically, NWB has high computing capabilities, which enables it to easily produce the visualization of large networks. On the contrary, SocNetV couldn't handle large networks as well as
NWB. However, SocNetV provided us with significant network analysis statistics that NWB couldn't generate. Therefore, both of them were selected for different reasons and they were used in order to generate better results.

Moreover, some analysis terminology that will be used should be explained. Specifically, the degree of a node is the number of edges adjacent to the node and the average degree of the network is average of the degrees over all nodes in the network. Additionally, the shortest path is the smallest distance between two nodes in the network. Smallest distance means the fewer number of links that need to be gone through from one node in order to reach the other one. So, the average shortest path between a pair of nodes of the network is the average value of the aforementioned definition. Furthermore, the diameter of a network is the maximum length of all shortest paths between any two connected nodes. Also, the clustering coefficient of a node is the number of existing links connecting the node's neighbors to each other divided by the maximum possible number of such links. In other words, the clustering coefficient shows the completeness of the neighborhood of a node. Hence, the average clustering coefficient of a network is the average of the clustering coefficients of all the nodes. The density of a network is defined as the ratio of the existent edges to the potential edges. In addition, betweenness centrality measures how often a node acts as a bridge through the shortest path between two other nodes. Finally, closeness centrality measures how close a node is to all the other nodes in a specific group of nodes. This means, that it basically measures how fast information spreads from a given node to other reachable nodes in the network.

In order to generate visualization results about the networks that will be examined, certain layout algorithms are used each time and firstly they should be explained. The first algorithm used is the Kamada-Kawai layout, which is a force directed layout algorithm. This means that it is based on the philosophy of estimating a force between any two nodes by getting the derivative of the force equations. Additionally, this algorithm is an ideal one for quickly portraying undirected graphs, like the ones that are used here. In some cases after Kamada-Kawai layout algorithm another algorithm (e.g. Fruchterman-Reingold algorithm) should be used for better visualization results. So, the second algorithm used is the Fruchterman-Reingold algorithm, which is also a force directed algorithm and it is mostly used for visualizing very large undirected networks, like in our case. Another algorithm used
from NWB Guess visualization options is the Bin Pack algorithm layout. This algorithm's main goal is to pack objects of different volumes into a finite number of bins, so that the number of bins being used will be minimized. Moreover, GEM layout algorithm is used as well. GEM layout algorithm is a spring-embedder algorithm that is based on the concept of minimizing edge intersections in a great deal of contexts.

Regarding the Enron email dataset, firstly a relatively small network will be presented. This network contains all the emails sent and received between 1997 and 1998. Since this network is quite small, no filtering of the data is needed. In order to construct this network, Python code was used and produced a .graphml file that was then imported into NWB. The Python code is shown below. The connection and the disconnection of the database and the creation of the cursor is not presented because it is shown previously.

```python
sql = """SELECT GROUP_CONCAT(mid), sender, receiver
     FROM result
     WHERE year(date) BETWEEN 1997 AND 1998
     GROUP BY sender, receiver"

G = nx.Graph()

try:
    cur.execute(sql)
    results = cur.fetchall()
    for rows in results:
        G.add_node(rows[1])
        G.add_node(rows[2])
        e = (rows[1], rows[2])
        G.add_edge(e)

    nx.write_graphml(G, "net_1997_1998.graphml")
    db.commit()

except MySQLdb.Error, e:
    try:
        print("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
    except IndexError:
        print("MySQL Error: %s" % str(e))
```

Figure 36: Code to produce 1997-1998 networks.

This code basically creates a graph (G) and uses the result of MySQL command to add nodes and edges to this graph. Successively, this graph is transferred into a .graphml file and it is stored in the same directory with the executed code because no other directory is defined. After importing the "net_1997_1998.graphml" into NWB the following network is generated.
In order to produce the above network the Kamada-Kawai layout from the visualization options of NWB is used. Since in this case the network is not complex enough and Kamada-Kawai algorithm works properly and there is no need for the use of another algorithm.

According to Network Analysis Tool (NAT) of NWB, this network consists of 182 nodes and 377 links. In fact, one of the nodes, Jeffrey Skilling, is highlighted by presenting its label and by using blue color instead of red, which is the default one for nodes. However, during this time period Skilling only had one email exchanged, as it is clearly shown in the above picture. It should be stated that the other key person in the Enron scandal, Kenneth Lay, is not highlighted because during this time period he did not exchange any emails. Additionally, according to SocNetV some important results of the analysis are the network density and the network diameter which are 0.011 and 4 respectively. Moreover, the average degree of the network is 2.077, the average shortest path length is 2.051, the average clustering coefficient is 0.0234 and when using the weak component clustering 4 clusters are found. Furthermore, group betweenness centralization is estimated at 0.247 and the number of betweenness centralization is 8. Finally, group closeness centralization is 356.655 and the number of closeness centralization classes is 17.
Successively, in order to produce the network for the emails exchanged during 1999, Python code was used as shown below. This code creates seven different networks. Firstly, it creates a network for the first six months of this year. This is done because during the first semester of 1999 not a very large number of emails was exchanged and therefore the emails from these months could be grouped together. Secondly, it creates a different network for each of the remaining months (July to December).

```python
sqla = """"SELECT GROUP_CONCAT(mid), sender, receiver
    FROM result
    WHERE year(date) = 1999 AND
    month(date) BETWEEN 1 AND 6
    GROUP BY sender, receiver"""

Ga = nx.Graph()
m = range(7,13)
i = 0
q = 6

try:
    cur.execute(sqla)
    results = cur.fetchall()
    for rows in results:
        Ga.add_node(rows[1])
        Ga.add_node(rows[2])
        e = (rows[1],rows[2])
        Ga.add_edge(*e)

    nx.write_graphml(Ga, "net_1999_1_6.graphml")

while i<6:
    SQL = """"SELECT GROUP_CONCAT(mid), sender, receiver
    FROM result
    WHERE year(date) = 1999 AND month(date)=%s
    GROUP BY sender, receiver""",
    (m[i])

    G = nx.Graph()
    results = cur.fetchall()
    for rows in results:
        G.add_node(rows[1])
        G.add_node(rows[2])
        e = (rows[1],rows[2])
        G.add_edge(*e)

    nx.write_graphml(G, "net_1999_%s.graphml" % (m[i]))
    G.clear()
    i = i + 1
```

do.commit()
It should be stated that not all of the networks will be presented here. Firstly, the network with the aggregated emails for the first semester of 1999 will be shown (Figure 39) and secondly, the network that includes the month with the most exchanged emails, which is December, will be presented (Figure 40).

Again the name of Jeffrey Skilling is highlighted, but Kenneth Lay is not included in the network since he did not exchange any emails during the first semester of 1999. Also, Kamada-Kawai layout is used here as well and since the network is not large and complex enough there is no need for another algorithm to be used. Some useful network analysis statistics about this network are the number of nodes, which is 557, the number of links, which is 1,866, network density which is estimated at 0.006 and network diameter that is 7. Additionally, the average degree of the network is

```
except MySQLdb.Error, e:
    try:
        print("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
    except IndexError:
        print("MySQL Error: %s" % str(e))
```
3.357, the average shortest path length is 3.4, the average clustering coefficient is 0.116 and when using the weak component clustering 4 clusters are created. Furthermore, there are 80 betweenness centralization classes in the network and its group betweenness centralization is 0.353. Finally, there are 141 closeness centralization classes in the network and its group closeness centralization is 1,105.34.

As it is stated before, the network of December will be shown, since it is the month with the highest number of emails exchanged. This network is shown in the following picture below:

![Network of December 1999 (Kamada-Kawai and Fruchterman-Reingold layout)](image)

Figure 40: Network of December 1999 (Kamada-Kawai and Fruchterman-Reingold layout).

In the picture above the network of December 1999 is shown after using two different layout algorithms. Firstly, Kamada-Kawai layout and then Fruchterman-Reingold layout were used. Also, Jeffrey Skilling and Kenneth Lay are highlighted with blue and yellow color respectively. Some useful network analysis statistics about this particular network are the following. Firstly, the number of nodes is 1,604, the number of links, is 4,939, network density is estimated at 0.002 and network diameter is 10. Additionally, the average degree of the network is 5.010, the average shortest
path length is 3.091, the average clustering coefficient is 0.067 and when using the weak component clustering 18 clusters are created. Furthermore, there are 294 betweenness centralization classes in the network and its group betweenness centralization is 0.388. Finally, there are 449 closeness centralization classes in the network and its group closeness centralization is 3,175.5.

In order to filter the emails exchanged in 2000, the filtering of working and not working hours was used. Python code used for working hours is shown below.

```python
start_work = "09:00:00"
finish_work = "17:00:00"

for month in range(1,13):
    try:
        cur.execute("""SELECT GROUP_CONCAT(mid), sender, receiver
        FROM result
        WHERE year(date) = 2000 AND
        month(date)=%s AND
        time(date) BETWEEN %s AND %s
        GROUP BY sender,receiver"");
        (month,start_work,finish_work))

        G = nx.Graph()
        results = cur.fetchall()
        for rows in results:
            G.add_node(rows[1])
            G.add_node(rows[2])
            e = (rows[1],rows[2])
            G.add_edge(*e)

        nx.write_graphml(G,"net_2000_%s.graphml" % (month))
        G.clear()

        db.commit()

    except MySQLdb.Error, e:
        try:
            print ("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
        except IndexError:
            print ("MySQL Error: %s" % str(e))
```

Figure 41: Code to generate networks of working hours in 2000.

The above code creates 12 networks for 2000, one for every month of the year. Also, data is further filtered by specifying the time that emails were either sent or received. In this case, the time range is between 9 am and 17 pm, which is considered to be the basic working hours in a company. The largest network created by this filtering is in November 2000. This network is presented below (Figure 42) and some useful network analysis results are the number of nodes which is 5,176, the number of
links which is 15,547, network density which is estimated at 0.00058 and network diameter which is 11. Additionally, the average degree of the network is 3.015, the average shortest path length is 6.844, the average clustering coefficient is 0.060 and when using the weak component clustering 72 clusters are created. Furthermore, there are 821 betweenness centralization classes in the network and its group betweenness centralization is 0.224. Finally, there are 1,198 closeness centralization classes in the network and its group closeness centralization is 10,188.9.

Figure 42: Network for November 2000 during working hours (Bin Pack algorithm).

As for the algorithm used here is the Bin Pack algorithm and the result of it is like a network consisting of different clusters (bins). Also, Jeffrey Skilling seems to be actively participating in the exchange of emails during this time period within this particular time range. On the contrary, Kenneth Lay is not an active participant during November 2000 and between working hours.

Additionally, since the number of emails is increased outside working hours range, more filtering is needed. Firstly, the range is classified into messages
exchanged from "17:00:00" to "23:59:59" and from "00:00:00" to "08:59:59". In the first group the networks produced are quite small and therefore insignificant and they will not be presented. Also, Python code is the same as before, changing only the time range. In the second time range there are some quite large networks created that couldn't be handled by SocNetV and the visualization with NWB was not done properly. Hence, more filtering was needed and it was about days of the week and the weekend. So, the following code presents the creation of networks for each month of 2000 for emails exchanged between "00:00:00" and "08:59:59" at weekends.

```python
start_work = "00:00:00"
finish_work = "08:59:59"

for month in range(1,13):
    try:
        cur.execute("""SELECT GROUP_CONCAT(mid), sender, receiver FROM result
WHERE year(date) = 2000 AND
month(date)=%s AND
dayofweek(date) BETWEEN 6 AND 7 AND
time(date) BETWEEN %s AND %s
GROUP BY sender,receiver"",
(month,start_work,finish_work))

G = nx.Graph()
results = cur.fetchall()
for rows in results:
    G.add_node(rows[1])
    G.add_node(rows[2])
    e = (rows[1],rows[2])
    G.add_edge(e)

nx.write_graphml(G,"net_2000_%s.graphml" % (month))
G.clear()

db.commit()

except MySQLError, e:
    try:
        print ("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
    except IndexError:
        print ("MySQL Error: %s" % str(e))
```

Figure 43: Code for the creation of networks for each month of 2000 for emails exchanged between "00:00:00" and "08:59:59" at weekends.

From the above filtering the month with the largest network created is for December 2000. This network is presented below (Figure 44) and some useful network analysis results are the number of nodes which is 4,989, the number of links which is 16,534, network density which is estimated at 0.00066 and network diameter which is 14. Additionally, the average degree of the network is 3.323, the average
shortest path length is 7.299, the average clustering coefficient is 0.064 and when using the weak component clustering 43 clusters are created. Furthermore, there are 906 betweenness centralization classes in the network and its group betweeness centralization is 0.161. Finally, there are 1,290 closeness centralization classes in the network and its group closeness centralization is 9,871.64.

Figure 44: Network for December 2000 during "00:00:00" and "08:59:59" time range at weekends (GEM and Bin Pack layout).

The algorithms used here were firstly GEM layout algorithm and then Bin Pack. Again the network is depicted classified into clusters. Also, Jeffrey Skilling is shown as a blue node and Kenneth Lay as a yellow one. Both of these nodes are appeared bigger than the others in order to differentiate from them. As it is shown above, both of them actively participated in exchanging emails during this time period.

In 2001, the number of exchanged emails increased dramatically. Therefore the filtering here is not done for each month, but for every six days. This sounds peculiar, but when more days were used as a filter then the networks that were generated could not be processed. Again, the filtering for working and not working
hours is used. Although in previous years more emails were exchanged outside the working hours range, during 2001, specifically in the second semester of 2001, more emails were exchanged within the working hours range. This is important because the most critical events of the Enron scandal occurred during the second semester of this year. Therefore, these critical events during 2001 will be examined within the working hours range. Python code for working hours is shown below.

```python
start_day = 1
end_day = 6
start_work = "09:00:00"
finish_work = "17:00:00"

while end_day<=366:
    try:
        cur.execute("""SELECT GROUP_CONCAT(mid), sender, receiver
                    FROM result
                    WHERE (year(date) = 2001) AND
                    (dayofyear(date) BETWEEN %s AND %s) AND
                    (time(date) BETWEEN %s AND %s)
                    GROUP BY sender,receiver""",
                    (start_day,end_day,start_work,finish_work))

        G = nx.Graph()
        results = cur.fetchall()
        for rows in results:
            G.add_node(rows[1])
            G.add_node(rows[2])
            e = (rows[1],rows[2])
            G.add_edge(*e)

        nx.write_graphml(G,net_2001_%s.graphml % (end_day))
        G.clear()

        start_day = start_day + 6
        end_day = end_day + 6

        db.commit()
    except MySQLException, e:
        try:
            print ("MySQL Error [%d]: %s" % (e.args[0], e.args[1]))
        except IndexError:
            print ("MySQL Error: %s" % str(e))
```

Figure 45: Code for filtering year 2001 for every 6 days and during working hours.

After the above code execution, 61 networks were produced. However, here the most important will be presented. The importance of the network is based on the critical events of the Enron scandal. Therefore, three networks that involve some critical events will be presented.
The first network shown is the one between August 11 and August 16, 2001. It should be reminded that in August 14, Jeffrey Skilling announced his resignation, claiming personal reasons. The following picture depicts this network.

Figure 46: Network August 11-16, 2001, working hours (Physics layout).

The algorithm that was used to produce the above network is the Physics layout. Some useful network analysis results are the number of nodes which is 2,036, the number of links which is 5,659, network density which is estimated at 0.00137 and network diameter which is 14. Additionally, the average degree of the network is 2.794, the average shortest path length is 5.009, the average clustering coefficient is 0.0429 and when using the weak component clustering 37 clusters are created. Furthermore, there are 363 betweenness centralization classes in the network and its group betweenness centralization is 0.182. Finally, there are 544 closeness centralization classes in the network and its group closeness centralization is 3,987.74.
Secondly, the network between October 22 and October 27 will be shown. During this time period SEC began its examination over Enron's financial statements. Also, a few days earlier Andrew Fastow was fired. The following picture displays this network. It should be pointed out that this is the largest network produced by using the aforementioned filtering.

![Network October 22-27, 2001, working hours (GEM layout and Bin Pack layout).](image)

Obviously, Kenneth Lay exchanged more emails than Jeffrey Skilling. Actually, Jeffrey Skilling only received 4 emails. He did not send any emails, since he was resigned and probably did not use this account. In order to create the above network GEM and Bin Pack algorithms layout were used. Some useful analysis results about this network are the number of nodes which is 5,923, the number of
links which is 23,973, network density which is estimated at 0.00068 and network diameter which is 11. Additionally, the average degree of the network is 4.071, the average shortest path length is 8.007, the average clustering coefficient is 0.078 and when using the weak component clustering 115 clusters are created. Furthermore, there are 1540 betweenness centralization classes in the network and its group betweenness centralization is 0.209. Finally, there are 2,111 closeness centralization classes in the network and its group closeness centralization is 11,575.1.

Thirdly, the network between November 28 and December 2 is shown. During this time period some extremely critical events occurred. Enron's debt was downgraded by major agencies to junk bond status and the merge deal with Dynegy was cancelled. Also, on December 2, 2001 Enron filed for bankruptcy. The picture below displays this network.

Figure 48: Network November 28 - December 2, 2001, working hours (GEM layout, Bin Pack layout).
As it is obvious from the above picture, Jeffrey Skilling did not exchanged any during this time period. This is because Jeffrey Skilling had already resigned and he did not use this account to send messages. Also, no one sent him an email during this time period. However, Kenneth Lay actively participated by exchanging 36 emails in a six-day period. Again, in order to create the above network, GEM and Bin Pack layout algorithms were used. Moreover, some network analysis statistics are the number of nodes which is 3,467, the number of links which is 12,994, network density which is estimated at 0.00108 and network diameter which is 14. Additionally, the average degree of the network is 3.768, the average shortest path length is 4.695, the average clustering coefficient is 0.092 and when using the weak component clustering 82 clusters are created. Furthermore, there are 858 betweenness centralization classes in the network and its group betweenness centralization is 0.217. Finally, there are 1,199 closeness centralization classes in the network and its group closeness centralization is 6,729.97.

Regarding 2002, only a few months are interesting concerning the amount of emails exchanged. Therefore, only one month, which is March, will be shown. March was selected because it is the last month that a relatively high number of emails was recorded. Python code will not be presented because since there is no further filtering it is similar to the previous ones.

The aforementioned network is presented in the picture below (Figure 49) and the layout used here is the Physics layout. As it is displayed below, neither Jeffrey Skilling, nor Kenneth Lay exchanged any emails during this time period. Also, some helpful network analysis statistics are created by SocNetV and NWB. According to them the number of nodes is 3,088, the number of links is 10,341, network density is estimated at 0.00108 and network diameter is 11. Additionally, the average degree of the network is 3.364, the average shortest path length is 5.028, the average clustering coefficient is 0.078 and when using the weak component clustering 39 clusters are created. Furthermore, there are 423 betweenness centralization classes in the network and its group betweenness centralization is 0.352. Finally, there are 610 closeness centralization classes in the network and its group closeness centralization is 6,109.46.
Consequently, based on the above results it could be mentioned that during critical events in the Enron scandal the number of emails increased dramatically which means that the social networks became evidently larger at that time. However, since this dissertation does not include text mining techniques to prove that the context of emails is directly linked to the critical events, it does not aim to further describe the connection between the networks and these events. As it is cited before, it aims to provide the methodology and act as a toolkit basis for analyzing large datasets, like Enron's.
5 Conclusions

This chapter, Conclusions, describes the dissertation results and it includes four sections, which are Overview of the dissertation, Results evaluation, Personal reflection and Future work.

5.1 Overview of the dissertation

In this section of the dissertation a summarization of what was described before will be presented. It should be cited that this dissertation mainly involved two essential parts. The first one was the literature review of the topic, while the second one was the problem definition and its implementation part.

Literature review consumed a lot of effort and time in order to be completed. It was a demanding task that included a great deal of research on a topic that was previously unknown to the writer. Moreover, it was classified into four main categories. The first category was computer forensics. It was a topic that required a thorough research on the field in order to provide computer forensics definition, goals, process, rules, methods and eventually its importance. To make this happen several scientific papers and parts of two books were carefully studied. The second category was digital crimes. Regarding digital crime, research was again exigent in order to present digital crime definition and classification and the types of digital data and digital evidence. The third category was financial crimes, which included its definition and its categories. In this category as well several scientific papers were studied and part of a book as well. The fourth category was the Enron scandal. This was the most demanding category, since it required a meticulous research of an event that occurred twelve years ago. In order to fully understand the scandal and its effects numerous scientific papers were studied. Additionally, a documentary film about the particular scandal was watched, so that a scrupulous approach could be applied on the topic. Therefore, a great deal of information about Enron was presented. It included the outline of the scandal, the rise of the company, innovations and success keys, the collapse of the company and the reasons behind it, the federal investigation on the scandal, ex-employees convictions and compensations to victims.
Problem definition and implementation was also an extremely demanding process that had to be followed. However, in this part the difficulty was not due to extent studying of the research conducted in the field. The complexity of this part had to do with the practical issues while programming the code. As it is mentioned before, there was not a lot of experience in the programming field and therefore it highly consumed effort and time in order to produce some meaningful results. Regarding problem definition, the Enron email dataset and related work in its analysis were presented. Additionally, a description of this dissertation's problem was defined. Regarding the implementation part, methodology and technologies, statistical analysis and social network visualization were provided.

5.2 Results evaluation

This section includes the evaluation of the results that have been generated by this dissertation. Additionally, it includes a comparison of the objectives that were initially set and those that were actually achieved.

The key result of this dissertation is the creation of a computer forensics methodology that can be considered as the foundation when analyzing extremely large datasets, like the Enron email dataset. Hence, this dissertation aims to act as a computer forensics toolkit for further analysis that can be conducted on this subject in the future and for financial crimes in general. However, this key result has one significant difference than the objective that was initially defined. Initially, it was determined to develop an automated decision support system tool in order to handle the Enron dataset, which involves an extremely large amount of data. Nevertheless, the extremely large size of the Enron email dataset combined with time constraints and the lack of experience in handling such datasets made us alter the key objective of the dissertation. Therefore, it was decided to provide a computer forensics approach on financial crimes, which involve such large datasets, that can be regarded as the computer forensics methodology when analyze them. Consequently, the Enron email dataset was processed and statistical analysis and social network visualization results were produced, but not as an automated tool.

Moreover, there were some additional results generated by this dissertation. These results involve the statistical analysis and the social network visualization of the Enron email dataset. The statistical analysis provided us with some indicative
figures about the Enron email dataset. These figures showed that when the critical events of the scandal occurred the number of emails in most cases increased. Unsurprisingly, October 2001, which is considered as the month of the scandal, is the month that includes the highest number of emails. Additionally, social network visualization showed that networks became larger and more complex when critical events occurred. It is remarkable that in order to process the dataset a lot of data filtering was required. Again October 2001 provide us with the biggest network, since it is the month that included the highest amount of emails. Moreover, both in statistical analysis and in social network visualization two major actors of the scandal are highlighted. These employees, Jeffrey Skilling and Kenneth Lay, played a significant role in Enron's rise and fall as well. Also, they were convicted due to the scandal and that is why they are highlighted.

5.3 Personal reflection

When a dissertation is completed it always has an impact to the writer. This impact becomes even greater when the dissertation topic was previously unknown to the writer. In this dissertation, the Enron scandal was utterly unknown to me. On the contrary, there was some previous knowledge in the information security part of it. Regarding the Enron case, the lack of previous knowledge about the scandal made me study harder in order to deeply comprehend it. After conducting a meticulous studying about this particular scandal, it provided me with a remarkably strong impact. It made me realize that the current economic model may no longer fit in our modern society. This is because while studying the Enron case, the ethics of the corporate world and generally of the capital market were questioned several times. The philosophy behind capitalism causes the overexploitation of resources in order to create surplus or growth. That was exactly what Enron did. Its employees were solely evaluated based on their quarterly results and therefore they always had to present profits or otherwise they were fired. This equals to overexploitation of human resources in order to create surplus.

Moreover, it should be argued that Enron introduced one of the roughest employee ranking system (PRC) in the world due to Skilling's firm belief that people are only motivated by money and fear. As a result, this created a highly competitive environment in the company. However, it is proven that progress in general is not
only based on competition, but also on collaboration between people, who are not only motivated by financial gain, but also by purpose and passion to achieve a goal. Furthermore, before studying the Enron case, I believed in the concept of free markets. However, now I do not consider them as applicable to all cases. I agree with Robert Kuttner who suggested that other fields of the economy can be run as free-market, but electricity cannot, since opportunities for marker manipulation are created (Kuttner, 2002). California electricity crisis in the summer of 2000 is the best example that confirms Kuttner's suggestion.

Consequently, our modern society needs a market system that could fit in its ethical and social ambitions and not contradict with them. Therefore, it is time for a more ethical system to be created or maybe capitalism could alter in order to become more ethical.

Regarding the information security part of the dissertation, there was some previous knowledge on the subject, but not at the extent that there is now. This part also strongly affected me. It made me realize that there is no such thing as a completely secure system. There will always be system vulnerabilities. These vulnerabilities are discovered by attackers at a faster pace these days due to high computing capabilities. This means that nowadays information security becomes exceptionally difficult to maintain due to the sophistication and the speed of contemporary attacks. Therefore, there is an intense need for information security specialists who can provide a high level of security in the systems. Also, it influenced me in a way that I intend to learn more about information security and computer forensics after the completion of this dissertation and maybe attempt to work on this field at a professional level too.

Generally, this dissertation helped me in developing independent learning. Regarding the literature review, numerous scientific papers were studied in order to accomplish a deeper knowledge on the subject. As for the implementation part, which was the most challenging one, a lot of research via the internet was conducted in order to be able to handle Enron email dataset. Hence, independent learning helped me realize the way that scientific research is conducted in general. Additionally, it made me want to conduct a further research on the topic in the future.
5.4 Future work

The Enron email dataset is an extremely large dataset that could not be easily processed. However, processing was eventually accomplished by using several filtering techniques. Nevertheless, such a large dataset has a great deal of information still hidden. Therefore, in this section suggestions about future work on the topic will be provided.

Firstly, a thorough text mining of emails body should be conducted. There have been some attempts to this direction, but a more scrupulous approach should be applied. The suggested idea is to develop an algorithm that performs data mining on the body of each email in order to determine whether there was some information about the scandal exchanged or not. Undoubtedly, this approach would take a lot of time to be completed, mainly because of the incredibly large amount of emails. Also, it would require high computing capabilities in order to accelerate the process.

Secondly, another suggestion is the development of an interactive tool that the user would be able to specify the filtering of MySQL commands in order to visualize the networks. Particularly, the user could specify the nodes of interest, the time period range, the hours range and many other options should be available.

Finally, while processing the Enron email dataset some employees were recorded to have more than one Enron email accounts. For example, the email account of Kenneth Lay that was recorded in the table "employeelist" is kenneth.lay@enron.com. However, in the dataset there are more accounts that he probably used, like ken_lay@enron.com and kenneth_lay@enron.com. Additionally, some emails in the dataset were exchanged by employees' other email accounts (e.g. jeffreyskilling@yahoo.com). Consequently, the last suggestion is to develop a tool that would be able to accumulate all the accounts of each employee in one unique account in order to increase results accuracy.
Bibliography

- Cory Altheide, Harlan Carvey, 2001, "Digital Forensics With Open Source Tools".
- Rabeah Al-Zaidy, Benjamin C.M. Fung, Amr M. Youssef, Francis Fortin, 2012, "Mining criminal networks from unstructured text documents".
- Australian Institute of Criminology, 2011, "CyberCrimine: Definitions and General Information".
- Anurat Chapanond, Mukkai S. Krishnamoorthy, Bulent Yener, 2005, "Graph Theoretic and Spectral Analysis of Enron Email Data".
- Sayan Chatterjee, 2003, "Enron’s Incremental Descent into Bankruptcy: A Strategic and Organisational Analysis".
- Steven C. Curral, Marc J. Epstein, 2003, "The Fragility of Organizational Trust: Lessons From the Rise and Fall of Enron".

Digital Forensics Research Workshop (DFRW), 2001, "A Road Map for Digital Forensic Research".


Peter Gottschalk, 2010, "Investigation and Prevention of Financial Crime".


Paul M. Healy, Krishna G. Palepu, 2003, "The Fall of Enron".

Hulst, R.C. van der, Neve, R.J.M., 2008, "High-tech crime, different crime types and perpetrators".
Information Security and Forensics Society (ISFS), 2004, "An Introduction to Computer Forensics".

W. Chan Kim, Renee Mauborgne, 1999, "Strategy, Value Innovation and the Knowledge Economy".

Bryan Klimt, Yiming Yang, 2004, "The Enron Corpus: A New Dataset for Email Classification Research".


Robert Kuttner, 2002, "Free Markets are Great-but not for electricity".

Wei-Jen Li, Shlomo Hershkop, 2008, Salvatore J. Stolfo, "Email Archive Analysis Through Graphical Visualization".


Shinako Matsuyama, Takao Terano, 2008, "Analyzing the ENRON Communication Network Using Agent-Based Simulation".

Rodney McKemmish, 1999, "What is Forensic Computing?", Australian Institute of Criminology trends and issues No. 118.

Bethany McLean, Peter Elkind, Alex Gibney 2005, "Enron: The Smartest Guys in the Room", DVD.

Kerry-Anne Mendoza, 2012, "Is there such a thing as ethical capitalism?", viewed 2 November 2013, &lt;http://www.opendemocracy.net/ourkingdom/kerry-anne-mendoza/is-there-such-thing-as-ethical-capitalism&gt;.


Aaron Philip, David Cowen, Chris Davis, 2010, "Hacking Exposed Computer Forensics, Secrets & Solutions".
William C. Powers, Raymond S. Troubh, Herbert S. Winokur, 2002, "Report of Investigation by the Special Investigative Committee of the Board of Directors of Enron, Corp.".


Karen Ryder, 2002, "Computer Forensics - We’ve had an incident, who do we get to investigate?".

Jitesh Shetty, Jafar Adibi, 2005, "The Enron Email Dataset Database Schema and Brief Statistical Report".


The Council of Europe, 2012, "Convention on Cybercrime".


C. William Thomas, 2002, "The Rise and Fall of Enron".


United States Computer Emergency Readiness Team (US-CERT), 2008, "Computer Forensics".

Appendices

Appendix A

Table 5: Order of digital evidence volatility (Matthew Braid, 2001).

<table>
<thead>
<tr>
<th>Order of Volatility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Registers and Cache</td>
</tr>
<tr>
<td>2. Routing tables</td>
</tr>
<tr>
<td>3. Arp Cache</td>
</tr>
<tr>
<td>4. Process Table</td>
</tr>
<tr>
<td>5. Kernel statistics and modules</td>
</tr>
<tr>
<td>6. Main memory</td>
</tr>
<tr>
<td>7. Temporary file systems</td>
</tr>
<tr>
<td>8. Secondary memory</td>
</tr>
<tr>
<td>9. Router Configuration</td>
</tr>
<tr>
<td>10. Network Topology</td>
</tr>
</tbody>
</table>

Appendix B

Table 6: Enron Segment and Stock Market Performance, 1993 to 2001 (Healy and Palepu, 2003).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pipelines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenues</td>
<td>$1,466</td>
<td>$976</td>
<td>$831</td>
<td>$806</td>
<td>$1,416</td>
<td>$1,849</td>
<td>$2,032</td>
<td>$2,955</td>
</tr>
<tr>
<td>Earnings^a</td>
<td>382</td>
<td>403</td>
<td>359</td>
<td>570</td>
<td>580</td>
<td>637</td>
<td>685</td>
<td>732</td>
</tr>
<tr>
<td>Domestic:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trading &amp; Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenues</td>
<td>$6,624</td>
<td>$6,977</td>
<td>$7,269</td>
<td>$10,858</td>
<td>$16,659</td>
<td>$23,668</td>
<td>$28,684</td>
<td>$77,031</td>
</tr>
<tr>
<td>Earnings^a</td>
<td>316</td>
<td>359</td>
<td>344</td>
<td>332</td>
<td>766</td>
<td>403</td>
<td>592</td>
<td>2,014</td>
</tr>
</tbody>
</table>
Earnings are measured before subtracting interest and taxes.

Appendix C

Table 7: Timeline of Critical Events for Enron in the Period August 2001 to December 2001 (Healy and Palepu, 2003).

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 14, 2001</td>
<td>Jeff Skilling resigned as CEO, citing personal reasons. He was replaced by Kenneth Lay.</td>
</tr>
<tr>
<td>Mid- to late August</td>
<td>Sherron Watkins, an Enron vice president, wrote an anonymous letter to Kenneth Lay expressing concerns about the firm's accounting. She subsequently discussed her concerns with James Hecker, a former colleague and audit partner at Andersen, who contacted the Enron audit team.</td>
</tr>
<tr>
<td>October 12, 2001</td>
<td>An Arthur Andersen lawyer contacted a senior partner in Houston to remind him that company policy was not to retain documents that were no longer needed, prompting the shredding of documents.</td>
</tr>
<tr>
<td>October 16, 2001</td>
<td>Enron announces quarterly earnings of $393 million and nonrecurring charges of $1.01 billion after tax to reflect asset write-downs primarily for water and broadband businesses.</td>
</tr>
<tr>
<td>Date</td>
<td>Event Details</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>October 22, 2001</td>
<td>The Securities and Exchange Commission opened inquiries into a potential conflict of interest between Enron, its directors and its special partnerships.</td>
</tr>
<tr>
<td>November 8, 2001</td>
<td>Enron restated its financials for the prior four years to consolidate partnership arrangements retroactively. Earnings from 1997 to 2000 declined by $591 million, and debt for 2000 increased by $658 million.</td>
</tr>
<tr>
<td>November 9, 2001</td>
<td>Enron entered merger agreement with Dynegy.</td>
</tr>
<tr>
<td>November 28, 2001</td>
<td>Major credit rating agencies downgraded Enron's debt to junk bond status, making the firm liable to retire $4 billion of its $13 billion debt. Dynegy pulled out of the proposed merger.</td>
</tr>
</tbody>
</table>