ABSTRACT

On the Internet, email communication has become widespread, but the exponential increase in Spam (intrusive messages) make the use of email for communication tedious and time consuming. Spamming is the process of flooding the internet with many copies of the same email message. Spam has become a persistent annoyance in the lives of the Internet user. It is considered to be one of the biggest computer problems that the users face today. People who send spam are called spammers. It is primarily used for commercial purposes. The main idea of this project is to study the techniques, methods, tools and strategies used by the spammers. During the course of the project, different categories of Spam are studied and some Spam analysis tools that are available are tested and the results are then tabulated. Finally, a recommendation is made to avoid the spam problem and block the spammers email addresses.
# TABLE OF CONTENTS

Abstract....................................................................................................................................................................................ii

Table of Contents.........................................................................................................................................................................iii

List of Figures..................................................................................................................................................................................vii

List of Tables................................................................................................................................................................................ix

1. Introduction and Background.........................................................................................................................................................1
   1.1 Introduction.............................................................................................................................................................................1
   1.2 Spam Definition......................................................................................................................................................................2
   1.3 Basic Characteristics of Spam...............................................................................................................................................3
   1.4 Spamware Software...............................................................................................................................................................8
   1.5 Reasons to avoid Spam............................................................................................................................................................9
   1.6 Project Overview....................................................................................................................................................................10
   1.7 Legal Action against Spammers........................................................................................................................................11

2. Spam Detection and Prevention..................................................................................................................................................13
   2.1 Spammers and their Categories.............................................................................................................................................14
   2.2 Various methods used by Spammers to collect email addresses.......................................................................................15
      2.2.1 Dictionary Attacks...........................................................................................................................................................15
      2.2.2 Spambots..........................................................................................................................................................................16
      2.2.3 Online Forms.................................................................................................................................................................16
      2.2.4 Web Harvesting..............................................................................................................................................................17
      2.2.5 Direct Harvest Attack Technique..................................................................................................................................18
      2.2.6 Domain Contact Points....................................................................................................................................................20
      2.2.7 IRC and Chat Rooms......................................................................................................................................................20
3.3 Tools used in this Research

3.3.1 K9

3.3.1.1 Working of K9

3.3.1.2 Advantages of K9

3.3.2 Power Email Harvester

3.3.3 Power Email Extractor

3.3.3.1 Working of Power Email Extractor

3.3.3.2 Advantages of Power Email Extractor

3.3.4 Spam Inspector

3.3.4.1 Working of Spam Inspector

3.3.4.2 Advantages of Spam Inspector

3.3.4.3 Limitations of Spam Inspector

3.4 False Positives and False Negatives

4 Evaluation and Results

4.1 Control Flow

4.2 Email Accounts Setup

4.3 Implementation of Baiting Method

4.3.1 Chat Rooms

4.3.2 Respond to Spammer

4.3.3 Newsgroups

4.3.4 Online Surveys
LIST OF FIGURES

Figure 1.1      Spammer hiding the identity..............................................................4
Figure 1.2      Working flow of Spamware Software.................................................8
Figure 2.1      General path of an email .................................................................13
Figure 2.2      Web Harvesting................................................................................17
Figure 2.3      DHA Technique...............................................................................19
Figure 2.4      Anti-Spam Techniques.....................................................................25
Figure 2.5      An Open-Relay System.......................................................................27
Figure 2.6      RBL Architecture............................................................................28
Figure 2.7      Organization’s Mail Administrator.....................................................28
Figure 2.8      Individual users................................................................................29
Figure 2.9      Spam Filters.....................................................................................30
Figure 2.10     Challenge-Response Filter Mechanisms.............................................33
Figure 2.11     SPF Methodology.............................................................................37
Figure 3.1      Screenshot of Power Email Harvester...............................................44
Figure 3.2      Screenshot of how bulk emails are sent to all valid email addresses......45
Figure 3.3      Email Extractions from Power Email Extractor.................................47
Figure 3.4      Add new News Server.....................................................................48
Figure 3.5      Group Selection Screen..................................................................48
Figure 3.6      Screenshot of keyword Search Tool...................................................49
Figure 3.7      Screenshot of Save Extracted Emails..................................................50
Figure 3.8      Screenshot of Domain Exclusion List..................................................51
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9</td>
<td>Screenshot for Auto-Reconnect Option</td>
<td>52</td>
</tr>
<tr>
<td>3.10</td>
<td>Flow Chart of False Positive and False Negative</td>
<td>59</td>
</tr>
<tr>
<td>4.1</td>
<td>Control Flow Diagram</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Breakdown of Spam by Baiting Method</td>
<td>73</td>
</tr>
<tr>
<td>4.3</td>
<td>Pie Chart of Spam Categorization</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>General Settings of Power Email Extractor</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Power email extractor before selection of any newsgroups</td>
<td>78</td>
</tr>
<tr>
<td>4.6</td>
<td>Selected newsgroup tab under ‘Selected Groups’ Option</td>
<td>79</td>
</tr>
<tr>
<td>4.7</td>
<td>Extracted email addresses</td>
<td>80</td>
</tr>
<tr>
<td>4.8</td>
<td>Multiple Newsgroup Selection</td>
<td>81</td>
</tr>
<tr>
<td>4.9</td>
<td>Keyword search tool</td>
<td>82</td>
</tr>
<tr>
<td>4.10</td>
<td>Three exclusion lists available</td>
<td>83</td>
</tr>
<tr>
<td>4.11</td>
<td>Spam Inspector Toolbar</td>
<td>84</td>
</tr>
<tr>
<td>4.12</td>
<td>Manage Friends list option</td>
<td>86</td>
</tr>
<tr>
<td>4.13</td>
<td>Manage Enemies list option</td>
<td>87</td>
</tr>
<tr>
<td>4.14</td>
<td>Spam Abuse Report</td>
<td>88</td>
</tr>
<tr>
<td>4.15</td>
<td>Graph: False Positives vs. False Negatives</td>
<td>89</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Spam Characteristics</td>
<td>7</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Spam Breakdown</td>
<td>39</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>List of email accounts created for this research</td>
<td>62</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Email Accounts that were used to register with chat rooms</td>
<td>64</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Email Accounts that were used to respond to spammers</td>
<td>64</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Email Accounts that were used to register in newsgroups</td>
<td>64</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Email Accounts that were used for filling online surveys</td>
<td>65</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Email Accounts that were used to respond popup requests</td>
<td>65</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>Email Accounts used to register on Websites</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>Email Accounts used for unsubscribing</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>Email Accounts that were used to register on Gambling Websites</td>
<td>66</td>
</tr>
<tr>
<td>Table 4.10</td>
<td>Email Accounts that were used in Obfuscation Method</td>
<td>67</td>
</tr>
<tr>
<td>Table 4.11</td>
<td>Email Accounts that were not used in Baiting Method</td>
<td>67</td>
</tr>
<tr>
<td>Table 4.12</td>
<td>Spam Categorization</td>
<td>69</td>
</tr>
<tr>
<td>Table 4.13</td>
<td>Friends and Enemies email addresses list used for Spam Inspector tool</td>
<td>85</td>
</tr>
</tbody>
</table>
1. INTRODUCTION AND BACKGROUND

1.1 Introduction

Electronic mail (E-Mail) is considered to be a powerful tool for communication among Internet users. E-mail became a vital tool in this current era as it is cheaper and faster but Spammers misuse these e-mails messages as a direct marketing tool. They send thousands of bulk messages every day, known as “Spam” which occupies high percentage of all email traffic received by a particular user. Spam is a slang name for unsolicited commercial e-mail sometimes also called as bulk mail or junk mail. Spam constitutes a major problem for both e-mail users and Internet Service Providers (ISP).

The reason behind the abundance of Spam is, its little start up expenses associated with it. All that a Spammer needs is a computer, a list of e-mail addresses and Internet connection. Spam is not only offensive and annoying, but also causes loss of productivity, decreases bandwidth and costs companies billions of dollars each year. Spam has become so persistent to modern computing. In this research, a wide variety of Spam is analyzed. Information is gathered to examine the tricks and techniques of the Spammers. Also various Spam detection tools and anti-Spam techniques that are available are studied in this research. With the above identified techniques, the various types of Spam that are currently on the internet are detected and their characteristics are studied. Based on these studies an overall contribution is made to the existing knowledge base in Spam research.
1.2 Spam Definition

During the early days of Internet, Spam did not exist. Plenty of emails were sent, out of which few were considered to be Spam. Spam is the general term for incoming e-mails that are sent to different e-mail accounts without the recipient’s prior consent. In general the word “Spam” means something unwanted. It normally refers to unwanted, junk e-mail messages or Usenet messages. These days unwanted Instant Messages (IM) and cell phone Short Message Service (SMS) are also referred to as Spam. The acronym UCE (Unsolicited Commercial E-mail) or UBE (Unsolicited Bulk E-mail) refers to the standard definition of Spam. “Unsolicited” means that the recipient has not granted permission for it to be sent [Anti-Spam 2005]. “Bulk” means that the same message is sent to many people at the same time. Generally Spam promotes Internet-based sales. It always markets commercial products and is sent in bulk to thousands of recipients.

The definition of Spam is not clear across many users or organizations. There are many interesting definitions of Spam that exist today. One of the standard definitions which is used in the Spamhaus project is:

An electronic message is “Spam” if

- The recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; and
- The recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent [Spamhaus 2005].

It is also defined as “the use of any electronic communications medium to send unsolicited messages in bulk”. So basically the overall meaning that can be derived from
all these definitions is an e-mail message is considered to be Spam if it is sent in bulk and without the permission of the recipient who has no benefit from it.

Spam is generally subjective. It is important to differentiate the boundaries between the e-mail messages that are Spam from those that are not Spam, even though they share some common traits. Sometimes email messages are mistaken as Spam, even though they are not Spam and the Spammers get blamed.

Some examples are as follows:

- Endless list of jokes or chain emails forwarded from friends or relatives. These are not considered as Spam as the sender is already known to the recipient.

- Trojan emails sent from an infected PC or Computer Worms or Viruses. Viruses propagate through the internet and spread by email. Once a PC gets infected, it will send copies of emails to all recipients stored in the address book. If a recipient opens an attachment from the email, it now propagates through the infected PC’s address book.

- Subscriber based or permission based marketing emails are not considered as Spam, because the recipient had requested to receive information. Many people sign up for mailing lists or newsletters but may not want to receive Spam. Since these people specifically “opted-in”, the messages are not considered as Spam.

In these situations anti-Spam measures cannot be used to control such illicit emails.

1.3 Basic Characteristics of Spam

In order to differentiate Spam from legitimate e-mails, it is necessary to identify some typical Spam characteristics. Spam characteristics generally appear either in email
headers or message content. Email headers are the ones that show the routing of an email message from its origin to its destination. Many Spammers hide their identity by using fake email headers so that they can hide the source of the message, as headers can be used to track the origin of Spam message.

![Fig 1.1 Spammer hiding the identity](image)

**Fig 1.1 Spammer hiding the identity [Adapted from [Hide 2006]]**

Figure 1.1 clearly shows the path how a spammer can send spam messages, hiding the identity. The Spammer starts sending the spam message by first hiding his identity using a fake header, no header or encrypted header. Then generally uses an open proxy server or an open relay or any unsecured wireless network to send Spam messages to the Email Server. Once the email server receives the spam messages, it then re-directs them to the client’s machine.

Apart from the headers, Spammers use some typical words like ‘free’, ‘earn money’, ‘click here’ etc in their messages. It is always important to filter words correctly
or else there might be chance of blocking genuine e-mails. Some of the general characteristics of a Spam are as follows:

- **Invalid Return Address**: Spam emails will not have a valid return address. When the recipient tries to reply to the e-mail message received, an error will occur. Instead of an empty ‘To:’ field, Spammers generally use some e-mail address which is either a nonexistent address or a fake email address which may or may not be a correct one.

- **Missing Reply Address**: In this type of Spam the ‘To:’ field will be empty. There are chances in which a recipient receives an email that may not have a ‘To:’ field in it. There may be a specific spamming reason behind these types of emails. As Spammers send bulk emails by entering the recipients email address in the ‘Bcc:’ fields, the ‘To:’ field is often empty. This is in general for all the Spam emails.

- **An eye-catching subject is presented**: Almost all the Spam emails have a specific and interesting subject in their message to attract the users.

- **Spam emails contain lot of advertisements i.e., website advertising, property offers, money making methods, special offers etc.**

- **Most of the Spam is in English and comes mainly from Asia and United States of America. Now a days Spam in Spanish is becoming common.**

- **Recipients email address is not in the ‘To:’ or ‘Cc:’ fields**: Spammers generally hide the recipients email address in the ‘Bcc:’ field in order to conceal the fact that the mail was sent to a large number of recipients.

- **‘From:’ field is same as the ‘To:’ field**: One of the most important trick Spammers use is to fill both ‘From:’ and ‘To:’ fields with the same email address, instead of
an empty ‘To:’ field or using a wrong email address in ‘To:’ field. Both ‘From:’
and ‘To:’ may be bogus email addresses.

- Missing ‘From:’ field: This is another common trap used by Spammers to hide the
  actual sender of the email message.

- Missing Message Id: As the message Id contains the information about the
  message, it is generally malformed or missing.

- Illegal HTML: If the Spam message contains illegal HTML, it may contain
  hidden Java script or VB script, which may open pop up windows containing
  advertisements.

- Message with hidden HTML: Usually an email contains both the plain text part
  and HTML format of that text part, but Spammers generally use HTML format
  without the text part, and force the readers to use the HTML format to read the
  message. These HTML formats have embedded links to another websites. If a
  user clicks the link provided, the Spammer can easily identify the true email
  addresses and send even more Spam to the live email addresses.

- The images used in Spam e-mails will link to some remote site. Some Spammers
  try to use images instead of simple text to attract the recipients. When the
  recipient tries to open the image by clicking it, the image links to some remote
  site.

- E-mail with an empty subject line: Some of the Spam emails don’t have a subject
  line. The Spammers just leave it empty.

- Hidden Headers: The headers of the e-mails are hidden in order to make it
difficult to trace the Spammers.
• Too many recipients: For example bulk mail. An email containing more than 10 recipients in the ‘To:’ or ‘Cc:’ field can be considered as Spam. However this can be used for legitimate emails too, but they are personal.

• The subject line has no connectivity with the body of message: Spammers randomly generate the subject lines for the Spam e-mails. Most of the time contents of the body are not related with the subject line.

• Fake Unsubscribe option: If a user tries to unsubscribe from the Spam email, it either opens an advertisement website or the link doesn’t work [Spam Characteristics 2003].

Once these characteristics are recognized, suitable steps can be taken to block these Spam messages. In a survey, Red Earth Software analyzed the headers of five hundred Spam messages and found the following Spam characteristics as shown in Table 1.1

<table>
<thead>
<tr>
<th>Spam characteristics</th>
<th>% of researched mails</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recipient address not in To: or Cc: field</td>
<td>64%</td>
</tr>
<tr>
<td>To: field is missing</td>
<td>34%</td>
</tr>
<tr>
<td>To: field contains invalid email address</td>
<td>20%</td>
</tr>
<tr>
<td>No message ID</td>
<td>20%</td>
</tr>
<tr>
<td>Suspect message ID</td>
<td>20%</td>
</tr>
<tr>
<td>Cc: field contains more than 15 recipients</td>
<td>17%</td>
</tr>
<tr>
<td>From: is the same as the To: field</td>
<td>6%</td>
</tr>
<tr>
<td>Cc: field contains more between 5-15 recipients</td>
<td>3%</td>
</tr>
<tr>
<td>To: field contains more between 5-15 recipients</td>
<td>2%</td>
</tr>
<tr>
<td>To: field contains more than 15 recipients</td>
<td>1%</td>
</tr>
<tr>
<td>Bcc: field exists</td>
<td>0%</td>
</tr>
<tr>
<td>To: field is empty</td>
<td>0%</td>
</tr>
<tr>
<td>From: is blank or missing</td>
<td>0%</td>
</tr>
</tbody>
</table>
1.4 Spamware Software

Spamware is a software tool that is designed by Spammers in the Spam business. This tool is used by the Spammers to send Spam e-mails. The software has different capabilities to generate random e-mail addresses and also uses a variety of mail servers to import hundreds of addresses. Another type of Spamware is used to search for e-mail addresses to build lists that can be used either for Spamming directly or to be sold to other Spammers.

The designed Spamware is then sold by different companies naming them as e-mail marketing packages [Spam 2003]. Most of the Spammers will use the “Fire and Forget” technique, where they will send the Spam messages to as many people as they can for a single time and then move to a different ISP provider.

![Figure 1.2 Working flow of Spamware Software](image)
The working flow of the Spamware software is depicted in the Figure 1.2. The Spammer uses the Spamware software to collect different email addresses from various Email servers. He then stores all these emails addresses in the form of a database. The Spammer then uses this database to send spam messages to different email accounts listed in the database.

1.5 Reasons to avoid Spam

In 2004, the California legislature found that Spam costs more than ten billion dollars merely in United States. The cost of spam may include both commercial and non-commercial emails. As the cost of email communication is cheaper than any other form of communication over the Internet, even small number of spammers can occupy the Internet with junk mail in no time and create a nuisance to the email users. Due to this, there is a lot of damage that mainly affects the community and the communication channels that are being spammed.

There are many reasons, why a user or an organization needs to avoid Spam. Some of the reasons are explained below:

- Loss of Production and Equipment: Many users are bothered by spam as it interrupts the amount of time they spend on reading their legitimate emails. They need some time to dismiss these unwanted messages, which indirectly effects the production of a business.

- Network Resources: Spammers attack different networks to different extents. Therefore, to measure the actual amount of Spam generated by a single person or by a group, many resources may be engaged. For e.g., the overhead and cost of
spam may include network bandwidth in order to acquire a host or a computer machine.

- Risks: Spammers often send Spam through insecure proxy servers and use false information. Using these falsify addresses, some spammers hack software to steal credit card numbers.
- Thefts: Spam affects the users either directly or indirectly. It may cause problems like financial theft, identity theft, data theft, property theft, virus bugs, child pornography (which is illegal in law), fraud, malware infection and unreliable marketing.

1.6 Project Overview

In this research project, large volumes of spam and their types are studied. Also, various approaches and techniques used by spammers to send spam are analyzed. The research project is mainly aimed at IT security professionals and users who are interested in the problem of spam. The research program starts with creating 40 email accounts with different email vendors. These email addresses are then registered in various newsletters, forums, advertisements and other malicious sites, which are more likely to send Spam to their registered users.

In the mean time while Spam is hitting the email addresses that were created, different tools and techniques are studied to detect and prevent Spam emails. The email addresses are then monitored in regular intervals to check for the volume of Spam. The Spam is then filtered and categorized according to its origin and characteristics.
Based on the results obtained, the Spam emails are then tested with different tools, thus studying the advantages and disadvantages of each tool that are used in this research. Finally analysis report is prepared in the form of tables and charts, which show the amount of Spam obtained in each category. At the end of the research, a detailed report including the contribution for the project is submitted which contains recommendations that can be followed to reduce the volume of Spam e-mails.

1.7 Legal Action against Spammers

Spammers should be treated legally. Not only they cause problems to individuals, they also harm the whole society. On January 1, 2004 CAN-SPAM (Controlling the Assault of Non-Solicited Pornography and Marketing) action came into force which allows Spammers to be imprisoned [Legal 2004]. It also prohibits many of the Spammers tactics used to hide their tracks and avoid their faults. The Can-Spam Act can be used only by net service firms and governments to tackle Spammers. AOL, Microsoft, EarthLink and Yahoo are the four firms that are taking legal action against the world's most cosmic Spammers [Control 2004]. The US firms states that the Spammers use fake email addresses to hide their identities as well as their domains and use open proxies through third party computer systems to send Spam [Sue 2004]. To avoid this, firms are making changes to e-mail software to ensure that messages come from the net domains that they want. ‘Sophos’, a world leader in protecting businesses against Spam and viruses, have published a report. This report explains from which countries have the Spam messages originated. They did a research on all Spam messages received and
revealed a 'dirty dozen' of offending countries in which the United States topping the chart. The 'dirty dozen' are as follows:

1. United States 56.74%
2. Canada 6.80%
3. China and Hong Kong 6.24%
4. South Korea 5.77%
5. Netherlands 2.13%
6. Brazil 2.00%
7. Germany 1.83%
8. France 1.50%
9. United Kingdom 1.31%
10. Australia 1.21%
11. Mexico 1.19%
12. Spain 1.05%

Others 12.23%

Sophos' findings are based upon an analysis of hundreds of thousands of emails examined by Sophos Spam researchers using honeypots and other measuring systems around the world [Sophos 2004].
2. SPAM DETECTION AND PREVENTION

This research covers the technical solutions that are presently available to solve the problem of Spam. As Spam is increasing day by day in its volume and becoming an annoyance, various anti-Spam techniques were developed to counteract it. Different tools have been developed by many professionals; in response to it, the Spammers evolved their techniques to increase the number of Spam. As different Spam filters are improved, Spammers designed different methods to bypass the filters. The process of developing both Spam and anti-Spam tools continues even today. Anti-Spam tools use a wide range of techniques, to decrease the Spam received by a user. Understanding the strengths and weakness of each tool and the methods used to defeat them is the basis of an effective Anti-Spam strategy.

![Figure 2.1 General path of an email](Adapted from [Connection 2006])

Figure 2.1 shows the general path of an email, which starts at the mail server and ends at a particular user Inbox. The Email Server receives all the emails which contain both spam messages and legitimate emails. The Email Server (for e.g., Yahoo) uses its
own filtering technique and rejects the emails which it considers as Spam, but that may not include all the spam messages. All the other emails which the server considers as true emails are sent to the Email Client (for e.g., Microsoft Outlook). Generally, Email Clients use some spam filtering software tool to filter the emails that are received, to a more intense level. The spam filtering software creates a special folder and sends all the quarantined spam to that folder, leaving the legitimate emails in the inbox. The user can now safely check the inbox, but if he wants to verify the spam messages he can do that by checking the Spam folder that is created.

2.1 Spammers and their Categories

People who send Spam are called Spammers. The motivation for almost all of the Spammers is money. The Spammers can be categorized as follows:

- Casual Spammer: These are the regular internet users who forward chain letters.
- Hobbyist Spammer: This kind of Spammer just sends Spam as a hobby. They send Spam to their normal clients using mass BCC mailing.
- Accountable Spammer: These Spammers are engaged in advertising authorized business.
- Hacker Spammer: These Spammers use tools that can bypass filters. They use some hacking techniques like cracking the password of certain users, to send Spam [Spam 2003].
- Small-Scale Spammer: The Spammers in this category send Spam using either Spamming toolkit or email address CDs that can be purchased or downloaded from internet.
• Large-Scale Spammer: These Spammers are well funded and have good knowledge of Spamming. They create the major harm on the Internet. These types of spammers generally use some sophisticated methods to gather email addresses and send spam. So these spammers are also called as sophisticated spammers.

2.2 Various Methods Used by Spammers to Collect E-mail Addresses

Many Internet users are tired of receiving Spam from Spammers. Spammers use various kinds of methods to collect true e-mail addresses. Most Spam softwares send out a robot or spider (just like a search engine) which "harvests" any e-mail addresses visible on company Web pages. Some Spam companies will simply collect Website addresses from the Internet and guess some of the generic e-mail addresses that may be in use, for example: info@, sales@, and enquiries@yourdomain.com. Alternatively hackers break in to existing subscribed e-mail lists, often newsletters or industry publications, and steal the client e-mail lists [Anti-Spam 2005].

Spammers can find email addresses in many different ways. Some of the known methods are described in the following sections.

2.2.1 Dictionary Attacks

These are also known as brute force attacks. They occur when Spammers use software that will generate e-mail addresses, which are the combinations of common first names and last names found in the dictionary, for example joerichard@msn.com. In this case, the Spammers have no idea of what actually the recipient email address is. Spammers just choose one of the many domains names available and then use some software (email generation software) to generate possible email addresses such as
joerichard@msn.com. To check whether the email addresses are true or not, Spammers use some validation software to test which addresses are legitimate and active. Invalid addresses produce a failed comment where as the active emails won’t [Spam 2003].

2.2.2 Spambots

Spambot is a piece of software which crawls through the ISP’s looking for e-mail addresses to send Spam. It is usually written in C programming language, for speed and portability reasons. Using these Spambots, a spammer can collect email addresses of innocent victims. A Spambot generally scans a Webpage for hyperlinks and email addresses. It stores the email addresses to the target list and follows the hyperlink and starts the process from beginning with the new Webpage. Some Spambots send spam to the email addresses as it finds them, whereas some other stores them for later use. Spambots usually vary in their intelligence and sophistication. A plain Spambot follows each hyperlink and stores the email addresses as it acquires, whereas some intellectual Spambots can identify dead links and can also track many pages at a time [Spambots 2005].

2.2.3 Online forms

Spammers generally get email addresses from various online forms that are filled by users. Users subscribe to many online merchants, guestbook’s etc, without their knowledge. But sometimes, vendors may add the email address to their database and sell to Spammers. Spammers then use this email addresses to further send spam messages.
2.2.4 Website Harvesting

This is another method a Spammer can wind up email addresses to the Spam lists. Spidering software or any other automatic programs are used to browse the Web and extract all the email addresses found on the Web sites, DNS records or online directories.

There are three different ways a spammer can extract the email addresses using Web harvesting techniques. A spammer can use one or all of these three techniques to gather email addresses. Figure 2.2 shows the different harvesting techniques used by Spammers to get required information.

![Web Harvesting Diagram](image)

- **Web Content Harvesting**: It is directly concerned with the content and description of the document, for e.g., images, email messages etc in a Web page. There are two approaches involved in this type of harvesting.
  1. **Web Page Content Harvesting**: It uses the general structure of the documents and also maps to some data model to gather required information for Spammers.
2. Search Result Harvesting: It performs like a search engine. It goes beyond the keyword extraction relating the words and phrases in the documents given by the Spammer.

- Web Structure Harvesting: In this type of harvesting model, the Web pages from different sources are linked, i.e., a Webpage may link to another Webpage and it may have links to some other page. A Spammer can use this harvesting method to collect more and more links which may further contain more email addresses in each link he visited.

- Web Usage Harvesting: It uses the data recorded by servers to evaluate the effectiveness of the Web structure. This Technique is rarely used by Spammers, as it is very difficult for a Spammer to attack a server due to security reasons. But, some sophisticated Spammers may use this technique by going beyond the servers to collect email addresses. It has two different approaches.

  1. General Access-Pattern Tracking: This tracking method will analyze Web logs to identify structural issues in Web pages by understanding there access patterns. It is used by large organizations or companies.

  2. Customized Usage Tracking: This tracking method is used to identify individual trends that can be customized based on information found on the website. This method is used by individual users [Harvesting 2004].

2.2.5 Direct Harvest Attack (DHA) technique

In this technique the Spammer generates email addresses that belong to a specific domain and sends message to all of the email addresses generated. The domain mail
server responds with an error to the invalid addresses that don’t exist, in this way a Spammer can notice which addresses are valid by deleting the non-valid email addresses from the list.

**Fig 2.3 DHA Technique [Adapted from [PEPC 2005]]**

Figure 2.3 shows the working of DHA Technique. The Internet Email Server can receive emails from a valid sender (who generally sends valid email messages) as well as from a Spammer (who sends spam email). The spammer attacks the Internet with DHA technique and sends emails to the Mail server. If the email address is valid, the email will be directed to the target systems, if not, the Email server responds to spammer notifying that the email address is not valid. In this way, the spammer can collect all the valid email addresses.
2.2.6 Domain Contact Points

This is one of the tempting methods used by Spammers. Every domain has one or more contact points that have the email addresses of all the contact persons. Spammers try to attack these contact points to collect a list of email addresses by using ‘whois’ Linux command. This is one of the most common methods that a spammer uses to identify email addresses.

2.2.7 Internet Relay Chat (IRC) and Chat Rooms

Users in chat rooms generally give their email addresses to anyone who asks. This makes the work of Spammer much easier. People who are novice to chat rooms are easily trapped. In this way Spammers gather the email addresses and confirm that those are the true addresses and send Spam to those email addresses.

2.2.8 Random Presumption

Sometimes Spammers send spam by guessing an email address from user’s first name and last name i.e., firstname.lastname@domainname.com. The Spammer doesn’t know in advance whether the particular email exists or not and waits for an error message from server if the email doesn’t exist.

2.2.9 Filling Online Surveys

These days there are lot of free online surveys available on the Internet. There are also online surveys where the user gets paid for completing them. These surveys just ask for the first name, last name and email address. Spammers directly harvest email addresses from these online surveys or the survey owner will sell their email lists to the Spammers and within no time the user’s inbox will be filled with Spam.
2.2.10 Yellow Pages

There are various ‘People Finder’ Websites available on the Internet. These Websites help different users, to locate their friends and relatives information including their mail in addresses, phone numbers etc. These pages gather addresses from various sources for e.g., UseNet. Most of these Websites prohibit email address harvesting using some special technique, so that it will be hard for Spammers to track the email addresses of the people. Spammers check these pages regularly in order to gather the names and create their own email addresses from the names obtained.

2.2.11 Social Engineering

A special trick used by Spammers to convince the people to give their email addresses is by forwarding a message from one user to another. In this way a Spammer can collect all the true email addresses. This is also called chain emailing. When a user sends this type of email address to all his friends and family in his contact list, the Spammer will collect the entire list and add it to his database.

2.2.12 Online Groups

Spammers can easily get the true email addresses by joining in many groups that are available on the internet, for e.g., Yahoo groups or Google groups. These groups make their members email addresses easily available to the members of the group. Thus by joining the group he will be able to get numerous amounts of emails. After he finishes collecting emails from one group the Spammer simply shifts to a different group and run his routine again. Some of the online groups these days are hiding the email addresses of their members, i.e. a Spammer can only send messages to the members in that group.
without actually seeing his email address. This makes the job of a Spammer little
difficult.

2.2.13 ‘Ident’ Daemon

In Unix computers, there is a program which runs in the background known as
daemon, which allows other computers in the domain or network to identify the users
who connect to the local machine. If a person from this machine connects to the internet
and surf the Web, there are chances that a Spammer can connect to the local computer
and request the daemon running in the background for the person’s email address to be
given out. This attack is only used by more intellectual Spammers who are paid for their
work as it needs more knowledge on programming languages.

2.2.14 Web Browser

Some Websites use different tricks to extract the user’s email addresses without
even user’s notice. For example, few Web pages allow email to be sent to the target
machine which may be controlled by Spammers when a mouse passes over some part of
the page and no warning message will be issued. Some other Web browsers pass the
header along with the email address to every Web page the user’s visits.

2.2.15 Mailing Lists

A very common trick used by Spammers to get into the mailing servers and
collect the list of e-mail addresses. This is perhaps the easiest way of obtaining valid
email addresses. The Spammers buy or hack the Websites which host the mailing list. In
spite, most of the mail hosting servers being secure Spammers somehow will look for the
loop holes in the designed software to hack into the servers to steal all the email
addresses stored on them. Membership into discussion forums also provides Spammers to acquire their email addresses.

2.2.16 Purchase User Databases

Even though purchasing the databases of the users from either individuals or companies is illegal, most Spammers will do that. If a Spammer is ready to spend some dollars for propagating his Spam he uses this easy technique. This technique can be used by all Spammers, no matter whether he has hacking skills or not.

2.2.17 ‘Finger’ Daemons

Spammers also use a query called ‘finger’ query, in which it will produce the list of all currently logged on users on a particular host. Spammers use this information to obtain a list of active user accounts. This is a Linux command that is commonly used by most of the technical geeks.

2.2.18 Newsgroups

One of the common methods for obtaining the email addresses is by scanning different newsgroups (Usenet) for true email addresses. If a recipient posts their e-mail address to a newsgroup, the address is exposed to the whole world. Spammers use special techniques discussed above to extract these addresses from the newsgroups. Spammers also check the body of the message for the ‘@’ symbol in an email address [FAQ 2004].

2.3 Reasons why different users receive different Spam

The type of the Spam that a user receives is based on the Spammer who harvests the user’s e-mail address. If a Spammer harvests an e-mail address and does not care about the content of the e-mail, different types of messages will be delivered to the user’s
address. As the Spam lists are shared among many Spammers, a harvested e-mail address receives the whole amount of topics generally associated with the Spam messages.

There are many reasons why different users receive different Spam. One reason may be that different people use different filters. These filters have their own methods of filtering the email messages. Some people may receive one type of Spam while others may receive different Spam messages. The second reason may be different Spammers may target different email addresses as there will be more than one list of email addresses that will be sold or acquired by Spammers. The Spammers may use their own methodology to send Spam to the users. There are many ways a Spammer can send the Spam, so if a particular Spammer uses one methodology, the other Spammer may choose another method to send Spam. That means that even the methodology used by a Spammer may result in varying Spam messages.

2.4 Anti-Spam Techniques

A valid sender sends a safe email and a spammer sends spam email to the mail server. If the email client uses one or if necessary all of the anti-spam techniques, the spam can be completely treated and placed in a separate folder. In this way, the user can get all the legitimate emails to the inbox, which is now spam-free. If the user wishes to view the spam mail, it can be done by safely viewing it in the quarantined mail folder. Figure 2.4 clearly depicts all the above process in a diagrammatic manner.
Figure 2.4 Anti-Spam Techniques [Adapted from [Quarantine 2005]]

As seen in the figure 2.4 both the valid and also Spam emails will be received by the email client server in the same fashion. Different anti-spam techniques which will be discussed later in this section will run against the received emails thus separating the spam from real email. The separated Spam is then kept in the Spam Quarantine folder which is called differently in different anti-spam tools. This folder is isolated from rest of the email client and one can easily check the messages inside the Spam folder with very minimum risk. There are various anti-spam techniques that are in use today. Few of them are listed below.
2.4.1 Closing Open Relays

Spam has become a difficulty where the ISPs are terminating the accounts of Spammers when they are being reported. Therefore, the Spammers search for Open Relays. Open Relay is also called as Spam relay, third-party relay or insecure relay. A mail server that is configured to transmit mail from end users to any destination address is known as an open relay. An open relay is used to deliver both regular emails and Spam mails but, all emails passing through an open relay may not necessarily be Spam. Using an open relay, an unscrupulous sender can send large volumes of Spam. Open Relays allow unknown or unauthenticated users to send emails. If a Spammer discovers that the mail server is in a vulnerable state, he then uses some software to send thousands of bulk mails in a short period of time [Open Relay 2003]. When a Spammer hijacks a mail server, the organization may suffer from system crashes, equipment damage and loss of finance and business. Spammer has a trick of delivering Spam messages by spoofing their identity to make their messages appear from a legitimate source. Running an Open Relay has serious problems too, especially for a system administrator. The server may become listed on one or more Open Relay Blacklists (ORBL). Once the server is listed, emails could be rejected considering as Spam by other systems. The main reasons why a Spammer uses an open relay are

- Many open relay mail servers are high-powered machines with fast internet connections.
- There are some dedicated Spam operations flooding the internet from static locations.
- A Spammer can hide their identity behind an open relay.
The concept of an open relay system is shown in the Figure 2.5. A spammer may either use Open Proxy or an Open Relay to send Spam. There may be one or many Open Proxies between a Spammer and an Open Relay. The Spam message send by the Spammer may pass through more than one Open Proxy finally passing though an open relay or he can avoid all these channels in the middle by directly hitting the open relay with the spam messages. The message is then transmitted to all the mail exchange (MX) servers that are connected to this open relay [Open Relay 2004].

### 2.4.2 RBL - Realtime Blackhole List / Relay Blocking List

Mail Abuse Prevention System (MAPS) created the RBL. MAPS are a part of the Threat Prevention Network. Trend Micro maintains these MAPS to make sure reputation ratings are correct and current. Every rating includes broad Spamming histories and Spam samples for complete transparency into the database [Threat 2006]. RBL is a list of Spammer IP addresses. The administrators of mail systems subscribe to RBL mainly to block Spam messages. Once a company or an ISP subscribe to the RBL, they will know from which IP addresses to block traffic from. The receiving end will check the RBL for
the connecting IP address. If the IP address matches the one on the list, then the connection gets dropped before accepting any traffic from the Spammer. Figure 2.6 depicts the RBL Architecture and shows how the Mail server and RBL communicate [Spam Problem 2003].

Different types of people are affected by RBL and they fall in two categories:

- An organization's mail administrator installs some anti-Spam mechanisms to reduce the amount of time the organization wastes in manually filtering and deleting Spam messages. This process is shown in the figure 2.7.
- **Individual users**, typically pay for service from a DSL, dialup or cable modem provider. An individual user has the authoritative position in this relationship.

This Process is shown in the Figure 2.8 [Spam problem 2003].

![Figure 2.8 Individual users [Spam problem 2003]](image)

There are nearly 30 commonly used RBLs; these are used by thousands of mail system administrators in order to reduce inbound Spam. These RBLs differ in their technological performance, redundancy, and accuracy.

### 2.4.3 Spam Filters

There are several kinds of Spam filters that are available today. Most of the filters are not perfectly designed; they cannot totally cut down the amount of Spam a user receives. Most popular solution to avoid Spam is to use keenly designed filters. Filters are easily installed and once they are setup, the user simply reads email from the filtered folder which makes all the background filtering techniques transparent to him.
There are seven powerful spam filters that are in use today. They are, Content Based Filters, Community Filters, Bayesian Based Filters, Challenge-Response filters, Whitelist Filters, Collaborative Filters and Signature Based Filters. Figure 2.9 depicts these seven spam filters along with the filtering process. When an incoming email passes through one or all of these spam filters, the clean email goes to the inbox and the spam emails are filtered and cut off from the valid emails. All these seven spam filters are studied later in this section.

The ideal Spam filter should produce zero false positives (This occurs when a Spam filter identifies a message as Spam when it is not) and zero false negatives (This occurs when a Spam filter fails to identify a message as Spam when it is). This section deals with some of them.
**Content-Based Filters** analyze the message subject, headers and content looking for some keywords associated with Spam like ‘free’, ‘win an Ipod’. These filters are flexible but they require some hands-on maintenance; so, they are also called as Rule-Based or Heuristic Filters. In this filtering, emails are filtered based on some predetermined words or phrases that appear in Spam mails but not in personal mail. The rules necessary for a filter to work should be initially declared by email user.

Content Filters are not fully effective, because Spam usually comes from different senders and also, it is very difficult for a user to identify the words that indicate Spam. The chances of false positive and false negative will be more in this case. Therefore, filters and rules should be constantly updated.

**Community Filters** communicate with a central server. When a user receives a message that is Spam, the user marks it as Spam. This information is posted to the central server and the message is added to the database. When adequate people have designated the message as Spam, it will be blocked from user's Inbox in the future. The decision here is taken based on number of people rather than one single user.

**Bayesian Based Filters** are based on this theory. “Bayesian filters have to be "trained" from known "good" and "bad" e-mails. Bayesian filters maintain a database of Spam and legitimate emails. Once the database is large enough, the system ranks the words by the probability they will appear in the Spam message. During training they extract "tokens" and store them in a database. When a new message is analyzed, the message is split into tokens and each token is given a value according to the following criteria:
The frequency of the token in Spam messages that the filter has been trained on.

The frequency of the token in good messages that the filter has been trained on.

The number of Spam messages that the filter has been trained on.

The number of good messages that the filter has been trained on.

After applying Bayes formula to these results, a value is extracted which gives the probability of the message being Spam or not. This value is called as "Spamicity" or “Spam Score”.

As the Bayesian filters learn based on user feedback, they minimize false positives. This is considered to be one of the effective filters used within any organization. However, Bayesian filters are not always perfect. If a Spammer tries to lower the message overall Spam score, the effectiveness of the Bayesian filters may decrease over time. The main disadvantage of Bayesian filters is that they need to be trained. This disadvantage is ignored when compared to the positive aspects offered by this filter [Filters 2005].

**Challenge-Response (CR) Filters** automatically send a response to a sender (if the sender is unknown) asking them to take some further action, to ensure their message will be received. In these filters, when an unknown sender sends an email to a user account, the system sends a challenge back to the sender. If the unknown sender replies back to the system, the system will then send the email to the recipient. Now, the sender and the recipient can exchange messages without the intervention of the system. Occasionally, the sender never replies to the challenge and the mail will be lost in certain amount of time. The advantage of these filters is that they never let Spam come into the
mail box where as the drawback is that, they are very rude and legitimate emails may be arrive late or lost.

Figure 2.10 Challenge-Response Filter Mechanisms [Adapted from [Procmail 2006]]

Figure 2.10 shows the working process of a CR Filter. Here, when tom@abc.com sends his first email message to tim@xyz.com, the system treats tom as an unknown sender and sends a challenge back to authenticate whether he is a valid sender or not. The target system will respond to this challenge if they recognize the email address tom@abc.com. If the source machine receives a response for the challenge sends, it accepts the response and forwards the message from the sender to receiver. Now, Tom and Tim can exchange the mails without any interaction of the system.
Whitelist Filters will not accept e-mail from any address, unless it is on a list of known good e-mail addresses. It is contrary to RBL. It is quite an aggressive technique for Spam filtering. A Whitelist Filter connects to a Mail Server and passes email only if it is from one of the explicitly accepted recipients on to the inbox. The user will have to update the list very frequently to avoid lose of any legitimate email [Filter II 2004].

Collaborative Filters detect messages sent to multiple recipients at the same time. This filtering method is rather a new approach to Content Based filtering. When a user identifies an email as spam, a signature is calculated on that particular email and is added to the group knowledge base. A signature is computed for every new mail received by the user and is compared to the created database of known spam. If the signature matches any email in the database, it is considered as a Spam message. The main advantage of this method is that any occurrence of spam email can be identified in a short period of time. An example of Collaborative Filters is Signature-Based Filters which is explained next [Filtering 2004].

Signature-Based Filters examine the contents of Spam, which are derived from honey pots or any other dummy email addresses which are set up specifically to collect Spam. Once a Spam message is received, the content will be examined and a unique identifier is given. This unique identifier is obtained by assigning a value to each character in the email. A Spam signature is created by totaling the values which are assigned to the character. The signature created is added to the signature database and is sent to the email service’s subscribers as an update. This signature is then compared with all the incoming emails and the matching messages are treated as Spam.
The advantage of these filters is that they seldom produce false-positives and false-negatives. The disadvantage of these filters is that they are very easy to defeat. They only deal with the Spam that has already been sent. The existing signature will be useless, even if there is a slight modification in the email message. As each individual filter works with varying degrees of effectiveness and are subjected to a certain number of false positives, most of the filters that are in use today are a combination of discussed previously filters [Filters 2005].

2.4.4 Honeypot E-mail Address Technique

A Honeypot is an email address which is hidden in the page where no one will click on it, but email harvesters can capture it. It is an information system resource whose value lies in unauthorized or illicit use of that resource. Honeypots are dummy emails that are created to attract Spam. It has not legitimate activity. Any emails received at the honeypot address will get the IP address of the sending mail server that are blackholed for a period of time, so that emails to other addresses on the receiving email server will not get delivered [RSS 2006]. Honeypots come in different shapes and sizes and are categorized into two types, low-interaction and high-interaction honeypots. Interaction is the level of activity it allows an attacker. Low-interaction honeypots have limited interaction whereas high-interaction honeypots are complex, involving real operating systems and applications [Honeypots 2003]. The problem with these honeypots is that they can filter out known Spam but they don’t help in preventing previously unknown Spam.
2.4.5 Blacklists and Whitelists

A general method to control Spam is to reject connections at the mail server based on the origin. This method is known as Blacklisting. In order to achieve this, the IP address is converted to the domain name and check whether this domain name is in the list of blacklisted addresses. Once an Internet Service Provider (ISP) is in the blacklist, there are several actions that need to be taken in order to remove it from blacklist. This method is not very effective solution to stop Spam as it not only blocks Spammers but also the ISP customers. This is not a friendly approach because it affects the users who do not send Spam and eliminates wanted messages. Moreover blacklists are not updated frequently. On the other hand blacklists can help to reduce many of the cost associated with Spam. Blacklists can be used if the Spammers repeatedly use the same ISP’s to send Spam and therefore, have some utility in stopping known Spammers.

Whitelists are contradictory to blacklists which are the databases of trusted email addresses, IP addresses or domains. Emails received from these Whitelists are allowed to pass to the user’s inbox. Whitelists can be built by adding the trusted sources manually. This method rejects messages that do not originate from an already known contact. Whitelists produce an excessive number of false positives and therefore is not widely used by email users or administrators. The disadvantage of this method is that it is impossible to generate a list of all possible legitimate email addresses, as true emails come from many sources [Overview 2004].
2.4.6 Sender Policy Framework (SPF)

SPF is an anti-Spam approach in which the domain of an email sender can be authenticated for that sender i.e., It allows a domain to specify which computers are allowed to send email for that particular domain. It is an open standard which specifies a method to prevent sender address forgery by keeping the Spammers from forging the domain names given in an email. By revealing the Spammer's true domain it easily blacklists the domain that sends Spam.

![SPF Methodology](image)

Figure 2.11 SPF Methodology [Adapted from [Procmail 2006]]

Figure 2.11 shows the working methodology of SPF. When a Spammer sends a message to a particular domain, the domain does a SPF check and looks up in the DNS
record to test whether the message is coming from an authorized domain sender or not. If the email does not come through the authorized Mail Exchanger, it sends a rejection message to the sender stating that the sender did not use the authorized domain [SPF 2006].

2.4.7 Reverse DNS Lookup

Reverse DNS (Domain Name Server) lookup determines the hostname for a given IP address. When an email is sent from one server to another, a network connection is established between the two servers. The mail server that is receiving the email can take the IP address of the sending server and do a DNS lookup to see if it matches with the header information of the email. This is one of the methods to find out whether the sender is a Spammer or not.
3. RESEARCH

This research project mainly deals in finding different categories of Spam and various methods that are available to avoid them. This section briefly describes the breakdown of Spam into diverse categories and using baiting methods to attract Spam send by Spammers. It also explains about different tools used in this research project like Power Email Extractor and Spam Inspector. The concept of False Positive and False Negative is also briefly described in this part of the research project.

3.1 Spam Breakdown

Spam can be breakdown into different categories depending on its type. Table 3.1 depicts the percentage of Spam breakdown according to a Symantec Probe Network in 2005 [Symantec 2005].

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>23%</td>
<td>Products</td>
<td>E-mail attacks offering or advertising general goods and services. Examples: Devices, Investigation services, Clothing, Makeup</td>
</tr>
<tr>
<td>15%</td>
<td>Financial</td>
<td>E-mail attacks that contain references or offers related to money, the stock market or other financial &quot;opportunities&quot;. Examples: Investments, Credit reports, Real estate, Loans</td>
</tr>
<tr>
<td>14%</td>
<td>Adult</td>
<td>E-mail attacks containing or referring to products or services intended for persons above the age of 18, often offensive or inappropriate. Examples: Porn, Personal Ads, Relationship advice</td>
</tr>
<tr>
<td>12%</td>
<td>Health</td>
<td>E-mail attacks offering or advertising health-related products and Services. Examples: Pharmaceuticals, Medical treatments, Herbal remedies</td>
</tr>
<tr>
<td>9%</td>
<td>Scams</td>
<td>E-mail attacks recognized as fraudulent, intentionally misleading, or known to result in fraudulent activity on the part of the sender. Examples: Nigerian investment, Pyramid Schemes, Chain letters</td>
</tr>
</tbody>
</table>
It can be observed from Table 3.1 that the analysis performed by Symantec Corporation shows the maximum number of Spam attacks are from online products, Financial and Adult content.

### 3.2 Baiting Method

One of the most popular ways a Spammer can gather e-mail addresses is to extract them from various Websites. The Baiting Method derives its definition from its name which is used in this research for attracting the Spammers to send Spam messages to the created email addresses. Baiting method comes into place only after creating a sample of email addresses. According to this method these created emails addresses are then registered in various websites which are prone to send Spam emails i.e., these e-mail

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7%</td>
<td>Fraud</td>
<td>E-mail attacks that appear to be from a well-known company, but are not. Also known as &quot;brand spoofing&quot; or &quot;phishing&quot;, these messages are often used to trick users into revealing personal information such as e-mail address, financial information and passwords. Examples: Account notification, Credit card verification, Billing updates</td>
</tr>
<tr>
<td>7%</td>
<td>Internet</td>
<td>E-mail attacks specifically offering or advertising Internet or computer related goods and services. Examples: Web hosting, Web design, Spamware</td>
</tr>
<tr>
<td>7%</td>
<td>Other</td>
<td>E-mails attacks not pertaining to any other category</td>
</tr>
<tr>
<td>4%</td>
<td>Leisure</td>
<td>E-mail attacks offering or advertising prizes, awards, or discounted leisure activities. Examples: Vacation offers, Online casinos, Games</td>
</tr>
<tr>
<td>1%</td>
<td>Political</td>
<td>Messages advertising a political candidate's campaign, offers to donate money to a political party or political cause, offers for products related to a political figure/campaign, etc. Examples: Political party, Elections, Donations</td>
</tr>
<tr>
<td>1%</td>
<td>Spiritual</td>
<td>E-mail attacks with information pertaining to religious or spiritual evangelization and/or services. Examples: Psychics, Astrology, Organized religion, Outreach</td>
</tr>
</tbody>
</table>
addresses are used to signup for newsletters, online forums and other online sites which involve in mass e-mailing.

The email addresses thus created for use in baiting method are called as bait emails. These bait emails can also be used in emailing back to the known Spammers or replying back to the existing Spam. This is one of the well-known techniques in baiting method.

3.3 Tools used in this Research

There are two tools used in this Research. From the two, one is an anti-spam tool and the other is an email extractor tool. The anti-spam tool is ‘Spam Inspector’ and the tool that is used to extract email addresses is ‘Power Email Extractor’. These tools are explained briefly in the sections below. The original proposal submitted in spring 2006 to the department detailed using an anti-spam tool ‘K9’ and an Email Harvester tool ‘Power Email Harvester’. Since the tools were not working properly in generating the required results, two other tools with the same functionality were selected. When the K9 tool was setup for configuration with Microsoft Outlook Express 2003, it never got configured. The problem with the Power Email Harvester was that whenever the tool was executed, it used to enter into an infinite loop thus finally going to a deadlock situation. Since K9 and Power Email Harvester were described in the project proposal submitted, the explanation of these tools is stated below.

3.3.1 K9

K9 is a precise, easy to use and fast learning Bayesian Spam filtering tool. K9 is an email filtering application that works in conjunction with user’s regular POP3 email
program and automatically classifies incoming emails as Spam (junk email) or non-Spam without the need for maintaining dozens of rules or constant updates to be downloaded. It uses intelligent statistical analysis that can result in extremely high accuracy over time. K9 learns from its mistakes and becomes better and better at being able to identify Spam. More importantly it learns to recognize what the user considers to be Spam.

This software is available for free download. One can choose to download either the minimalist ZIP file containing the K9 executable only or the full installation EXE that will enable the user to uninstall the application and all associated files and registry items if one wishes to. When installing over a previous version the user will not lose any information or "learning". K9 does not contain spyware, adware or any kind of malware material [K9 2004].

3.3.1.1 Working of K9

K9 is an email-filtering program that works in conjunction with most popular email applications that use the standard POP3 email protocol. Messages pass through K9 on their way to email program and as K9 processes them, it can learn to identify the difference between Spam and Good emails, marking Spam emails so that the email application can file them away or delete them.

Initially K9 will not know what is Spam or Good email so; the user needs to correct it when it gets things wrong. Over time a user will find that the user has to correct K9 less often as it learns from past mistakes.

In summary this is how it works.

- Configuring email application to retrieve emails via K9. Emails are retrieved by user’s regular email program but pass through K9 before reaching it.
• When passing through K9, each message will be analyzed and based upon previous emails it will determine if the message is Spam or good. K9 marks Spam emails in a special way.

• When the emails arrive in email application it uses its built-in filtering rules to look for the special Spam marker placed there by K9. If the marker is present the rule can automatically place the email wherever the user likes, for example in a Spam folder, or simply delete it. In this way the user will no longer have to manually sift through dozens and dozens of emails looking for the occasional good message amongst the junk [K9 2004].

3.3.1.2 Advantages of K9

K9 is a powerful tool and has many advantages when compared with other anti-Spam tools available in the market [K9 2004]. The positive features of K9 are listed below.

• A small and fast Spam filtering program
• Easy to install and configure
• Simple interface
• SSL can be used with the combination of any third party software
• Very good and fast learning methodology
• It can be easily transferred from one system to other. The two most vital files are found in the DB directory and are called good.dat and Spam.dat. These two files hold K9's knowledge.
3.3.2 Power Email Harvester

Power Email Harvester is simple to use. The system requirements for this software are Microsoft Windows 95/98/NT 4.0/ME/2000/XP. This software will query email hosts, such as yahoo.com, hotmail.com and many others, building an emailing list of valid addresses and sending bulk email to each address at the same time while it is searching. Figure 3.1 depicts the screenshot of ‘Power Email Harvester’ [Harvester 2006].

![Power Email Harvester](image)

**Figure 3.1 Screenshot of Power Email Harvester**

This software checks the validity of all possible addresses within the parameters, the user sets for it. Then it creates a list of all valid email addresses within the specified range, in alphabetical order. The program can be saved and a message can be sent to that
address while it is extracting. The list can be saved for later use. Bulk email can be sent while harvesting to all targeted and untargeted mails from a mailing list. It can find and send to any address on any server at the same time. Figure 3.2 depicts how bulk emails are send to all valid email addresses [Harvester 2006].

![Screenshot of bulk email options](image)

**Figure 3.2 Screenshot of how Bulk Emails are sent to all Valid Email Addresses**

### 3.3.3 Power Email Extractor

Power Email Extractor is a very powerful email address extraction program. It has the capability to download and extract 100,000 unique email addresses per hour. It extracts email addresses from newsgroups as the newsgroups are the fastest and easiest way to extract email addresses. This program works faster than the search engine
extraction program that collects email addresses from different search engines. In an unregistered version the only option that doesn’t work is the file save option. This program also has an inbuilt keyword search tool feature that allows finding newsgroups that match the particular given keyword. This tool requires an internet connection to collect the email addresses [Extractor 2006].

3.3.3.1 Working of Power Email Extractor

First the Power Email Extractor when connected to the internet will download the list of newsgroups from a newsgroup server. Each newsgroup can have thousands of email addresses included in them. One can select a newsgroup or multiple newsgroups that they want to extract email addresses from them. After selection of newsgroups is done and extract button is pressed, the Power Email Extractor automatically connects to the news server and extracts the email address form the newsgroups that were previously selected. Once the downloading is completed, the email addresses can be saved to a physical location on the system. The tool can also be used for the keyword search to find newsgroups that the user wants, based on some criteria. For e.g., to match a particular word or phrase in the newsgroup the user can type his keyword or phrase in the provided toolbox area of the tool. Figure 3.3 depicts the screenshot of basic power email extractor when emails are extracted from a newsgroup server.
Power Email Extractor has many in-built features. They are

- Adding new Newsgroup Server
- Group Selection
- Keyword Search
- Saving the Extracted emails
- Exclusion List
- Auto Reconnect

All the above features are described briefly in the following sub sections.

- Adding new Newsgroup Server: If there is a need to add a new news server, one can add it by entering the new Newsgroup Server name by the option provided in the tool. Figure 3.4 shows the dialogue box of adding the new newsgroup server.
• Group Selection: If the user wants to select more than one group at a time, it can be done by selecting the multiple groups that the user wants to extract email addresses from, by holding the Ctrl key. Figure 3.5 depicts the screenshot of group selection. It can be seen from the figure, the list of selected groups in the Selected Groups tab filtered from the entire list in the All News Groups tab.

• Keyword Search: There is a keyword search tool available with this program that allows the user to find news groups that match certain keywords. Power Email Extractor allows the user to perform a targeted keyword search from the
Newsgroup List. This can be done in three different ways. Figure 3.6 shows the screenshot of the keyword search tool by screening the three options which are as described below:

- Match anywhere in name: Use this option when there is a need to obtain as many matches as possible.
- Match at the beginning of name: Use this option whenever there is a need to obtain the exact match at the beginning of a newsgroup name.
- Match exact sub category name: This search option will search for an exact match from a news group.

![Keyword Search Tool](image)

**Figure 3.6 Screenshot of the Keyword Search Tool**

- Saving Extracted emails: One can save the contents of the Extracted Email List by clicking the Save button. But, in the trail version used in this research project this option is disabled. Figure 3.7 shows the screenshot of save extracted emails page.
Figure 3.7 Screenshot of Save Extracted Emails

- Exclusion Lists: If the user doesn’t want to extract the emails from some specific domains or doesn’t want to extract emails containing specific words, he can use this option provided in the tool. In this way, the user can modify the contents of the emails by using the exclusion lists. There are three exclusion lists which are provide by Power Email Extractor. They are depicted in the screenshot of exclusion list of Figure 3.8.
Figure 3.8 Screenshot of Domain Exclusion List

- Word/Phrase Exclusion List: If this option is enabled any email addresses that contains those phrases or words won't be extracted, for e.g., if the user includes the word ‘Spam’ in the Word Exclusion List, Power Email Extractor doesn’t collect emails which has the word ‘Spam’ in it.

- Top Level Domain Exclusion List: If this option is enabled any email addresses that belong to any of those top level domains won't be extracted, for e.g. .edu, .mil etc are the common top level domains. So if the user includes any of these domains in the list, the Power Email Extractor will exclude emails from these domains.

- Domain exclusion list: If this option is enabled any email addresses from those domains won't be extracted, e.g. aol.com, yahoo.com etc are some of the most popular domains. If the user doesn’t want email form these domains, he can exclude them by using this option.
• Auto Reconnect: In the event that the connection to the server is lost Power Email Extractor can be configured to reconnect at preset intervals. Figure 3.9 shows the screenshot for auto reconnect. If this option is enabled, it automatically connects to the server without the user’s interference.

![Auto-Reconnect](image)

**Figure 3.9 Screenshot for Auto-Reconnect Option**

### 3.3.3.2 Advantages of Power Email Extractor

• It automatically removes duplicate email addresses during the extraction process.

• It supports authentication by having the required User name and Password.

• It has Firewall support. Users can setup Firewall settings by entering Firewall Host, Firewall Port, and Firewall Type.

• It has an auto-save feature where it can automatically save email addresses as they are extracted.

• The Keyword Search Tool helps the user to search for specific newsgroup categories that match the user criteria.

• The Domain Exclusion Lists allows the user to exclude the email addresses from that particular domain.
Word/Phrase removal lists allows the user to exclude the email addresses that contain the customized phrases or words.

Auto-Reconnect feature will help the tool to reconnect to the server when the Internet connection is lost. If it is enabled Power Email Extractor won’t stop in the middle of an extraction process.

This tool helps the user to load the previously saved selected newsgroup lists. This helps the user some time by not reloading the same newsgroups again and again.

Even tough the user don’t have any permission to any newsgroup, this tool automatically finds public newsgroup server and connects to it [Extractor 2006].

3.3.4 Spam Inspector

Spam Inspector is one of the powerful Spam filtering tools that identifies and separates Spam emails from the legitimate emails. It uses both adaptive (Bayesian) and collaborative filtering to filter junk, while, after some training and configuration tweaks leaving good email alone. It is based on personal and global learning, which adapts itself to the email automatically, filtering all junk email with 99% accuracy. Overtime the Spam inspector automatically learns from the incoming emails, training it to become almost 100% accurate in stopping and blocking the Spam. It's a pity that Spam Inspector can't be used in all email programs and doesn't support all account types (IMAP) everywhere. Currently there are three versions of Spam Inspectors for AOL, Outlook Express and Microsoft Outlook. In this Research, the Spam Inspector is configured to Microsoft Outlook 2003 [Inspector 2005].
3.3.4.1 Working of Spam Inspector

Spam Inspector integrates with Microsoft Outlook 2003 and executes an internal feature inside the outlook. Once the user installs the Spam Inspector, all the incoming emails are automatically analyzed against Spam Inspector's heuristic library to determine if the message is Spam or not. If the spam inspector receives any junk mails, they are quarantined locally. If the user wants to review the junk mails later, they can be viewed from a special folder that is created by the spam inspector, in the mailbox. Different features of Spam Inspector are described below:

- Spam Inspector filters and quarantines Spam.
- Spam Inspector supports Windows 9x/ME/NT/2000/XP.
- It uses adaptive Bayesian filtering that learns with user’s mail and a global network of Spam reports. Bayesian Filters are the most accurate Spam filters available today.
- Additionally, Spam Inspector uses white and black lists of senders as well as language filters.
- One can easily report junk mail to relevant ISPs with Spam Inspector.
- Spam Inspector protects user’s privacy by removing tracking images and hiding offensive pictures.
- When the user blocks a legitimate newsletter sender, Spam Inspector suggest unsubscribing.
- Spam Inspector is designed for Outlook Express, Microsoft Outlook 2000/2003, Incredi Mail, Eudora, and Hotmail (in Internet Explorer).
During the Installation process, the user needs to select one of the available Email clients (Microsoft Outlook, Outlook Express) on the system to configure the Spam inspector.

Spam Inspector has three integrated features. They are as follows:

- Managing Friends List.
- Blocking Foreign Characters.
- In-built Customized Toolbar.

All the above features are described in the following sub sections.

- Managing Friends List: Spam Inspector has an automatic friend’s list management option where it allows focusing on the emails which are sent by the people who are in the Friends list. It will automatically approve the email coming from the senders list. During the installation process, the user will be prompted to add senders to the Friends List. It will also add all the emails from contacts in outlook to the friends list during the installation process.

- Blocking Foreign Characters: Spam inspector allows the user to block 40 foreign characters and 200 foreign country domains. Even, this feature can be configured during Installation process. The user has an option of selecting few or all of the domains and characters. If this option is enabled, the user won’t be receiving emails containing any of the foreign characters or domains that are blocked.

- In-built Customized Toolbar: This is the last, but not the least feature of Spam Inspector. The Spam inspector provides with its own customized toolbar which has a quick access to the commonly used Spam inspector features. The default toolbar has eleven shortcut buttons.
1. Clean Folder: Cleans the currently selected folder.

2. Delete Spam: Deletes the contents of the Spam folder.

3. Is Spam: If the user receives a message that is considered to be Spam, but the Spam inspector does not, the user should simply highlight the message and click the ‘Is Spam’ button. This is a part of learning process where Spam inspector will treat other similar messages like this one as Spam in the future. There are some options that could be configured with this, which includes: Add Sender to Enemies List, Bounce Email Back to Sender, Move Email to Quarantine Folder and Move Email to Deleted Items.

4. Not Spam: If the Spam Inspector has identified a legitimate email as Spam, the user can tell Spam Inspector that the message is not Spam by selecting the message and clicking the ‘Not Spam’ button. If the user wishes to further configure the ‘Not Spam’ he can use the following options: Add Sender to Friends List, Mark Email as Unread and Move Email Back to Inbox.

5. Bounce: Bounce is used to send out an undeliverable message from the ISP that notifies the Spammer that the email address does not exists.

6. Report: Report the selected messages to the ISP as Spam and get the Spammers blackout from ever sending an email to anyone.

7. Add to Friends: Add the selected senders to the Friends Lists. In the drop-down, the user can either select to the email addresses of the sender or domain name of the sender, to be added to Friends List.
8. Add to Enemies: Add the selected senders to the Enemies lists. In the drop-down the user can either select to the email addresses of the sender or domain name of the sender, to be added to Enemies List.

9. Manage your Enemies List: This will open the Enemies List window.

10. Manage your Friends List: This will open the Friends List window

11. Options: This will open the Spam Inspector Options window.

### 3.3.4.2 Advantages of Spam Inspector

Advantages of using Spam Inspector are as follows:

- Spam Inspector is easy and convenient to use.
- Combines adaptive and collaborative learning for a good Spam detection rate.
- Integrates well with a number of email clients and supports Hotmail in Internet Explorer.
- Protects Users privacy by removing tracking bugs from email.
- It hides and deletes offensive images from pornographic spam.
- It alerts the user about virus emails.
- It updates filters based on the global reports automatically.
- It automatically blocks foreign language spam.
- Powerful Spam reporting to ISPs and government agencies.
- It has the feature of bouncing back the spam email to sender.

### 3.3.4.3 Limitations of Spam Inspector

Negative features of Spam Inspector are as follows:

57
• Spam Inspector is a little overzealous initially, with default settings.

• Training Spam Inspector takes longer than teaching other Bayesian filters.

• Spam Inspector doesn't work with all email clients [Inspector 2005].

3.4 False Positives and False Negatives

This section deals with the e-mail messages that are received by the accounts and labeled as spam. There are two instances that need to be studied while comparing the filters. One is how much spam a filter misses and other is how much non-spam does the filter mark as spam. The former is known as the ‘Miss Rate’ also called ‘False Negative’ and the later is known as the ‘Success Rate’ also called as ‘False Positive’. Apparently these two are useless as the user needs to look in Spam and non-Spam folders for the true messages. False Negative is the percentage of Spam that is not reported as Spam whereas the False Positive rate is the percentage of non-Spam that is reported as Spam. All the filters that are in use today provide spam indicators that mark an e-mail as Spam and provide the total number of Spam messages in the brackets. The success rate of any filter depends on these False Positives and False Negatives. The Flow Chart of False Positive and False Negative are explained in Figure 3.10[Filtering 2004].
Figure 3.10 Flow Chart of False Positive and False Negative [Filtering 2004]

When the Inbox receives a new email, it either appears in the Inbox folder or Junk folder. If the mail appears in the Inbox folder and if the user feels that it is Spam, he can select the email and move it to the spam folder and the spam folder is updated with this change. This is considered to be False Negative. If the filter classifies the message as Spam, it is directly moved to the Spam folder. In the other situation, the user manually selects a message from the junk folder and moves to any other folder (which is not a spam folder). This is considered to be False Positive.
4. EVALUATION AND RESULTS

The evaluation of this Research was done in three different phases. First, forty email accounts were created and baiting method was used to receive different kinds of spam. In phase two, the spam received from all the email addresses was categorized and a pictorial representation was prepared. In the third phase of the testing process, the tools were tested on the email addresses that were created for this research. All the different segments are discussed in the following sections.

4.1 Control Flow

Figure 4.1 shows the control flow diagram for the research completed. It explains the flow of the research project diagrammatically, representing the different steps that were involved and also different tools that were used in analyzing Spam. This control flow diagram uses symbols to represent the starting point of the research, processes involved in research, various entities implicated in the process, and documentation.
Figure 4.1 Control Flow Diagram
4.2 Email Accounts Setup

This Research was implemented by creating forty unique email addresses under four domain names: Yahoo, Gmail, Lycos and Hotmail. The sole purpose of these email accounts was to attract Spam by using baiting method. These email accounts were created with general names like Scott, Rick etc. During the creation of the emails, the age, location and sex were clearly defined in the profile section for each account because some spammers will specifically look at the age group and other sections of the profile before sending Spam. During the account setup, the checkbox was selected to list the email address in various different marketing Websites. Table 4.1 shows the list of different email accounts created for this Research. The email account names were created in multiple formats which contained numerics, long strings of characters, dictionary names (combinations of first and last names) and random characters (i.e., ema.123ilme@gmail.com).

Table 4.1 List of Email Accounts Created for this Research

<table>
<thead>
<tr>
<th>Email Accounts Created for this Research</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:svjonathan@yahoo.com">svjonathan@yahoo.com</a></td>
<td><a href="mailto:veerraj.1@gmail.com">veerraj.1@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:prasadvlsprasad@yahoo.com">prasadvlsprasad@yahoo.com</a></td>
<td><a href="mailto:karismam@gmail.com">karismam@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:tweakertwe@yahoo.com">tweakertwe@yahoo.com</a></td>
<td><a href="mailto:timshareif@gmail.com">timshareif@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:twistytween@yahoo.com">twistytween@yahoo.com</a></td>
<td><a href="mailto:warreng.123@gmail.com">warreng.123@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:bubblebob1999@yahoo.com">bubblebob1999@yahoo.com</a></td>
<td><a href="mailto:warreng.123@gmail.com">warreng.123@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:abarghav@yahoo.com">abarghav@yahoo.com</a></td>
<td><a href="mailto:tomstanley.tom@gmail.com">tomstanley.tom@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:ksrsharma2000@yahoo.com">ksrsharma2000@yahoo.com</a></td>
<td><a href="mailto:ema.123ilme@gmail.com">ema.123ilme@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:abhignalle@yahoo.com">abhignalle@yahoo.com</a></td>
<td><a href="mailto:robrodrigoes@gmail.com">robrodrigoes@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:Villevan2006@yahoo.com">Villevan2006@yahoo.com</a></td>
<td><a href="mailto:jk.sumanth@gmail.com">jk.sumanth@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:preloader2006@yahoo.com">preloader2006@yahoo.com</a></td>
<td><a href="mailto:bridgetted.12@gmail.com">bridgetted.12@gmail.com</a></td>
</tr>
</tbody>
</table>
4.3 Implementation of Baiting Method

One of the most popular methods a spammer can collect email addresses of users is to crawl through the Web and extract different addresses from the Websites. In this research, baiting method was implemented in various ways like registering the email addresses in chat rooms, signing up for newsletters, online forums and some other online sites which involve in mass e-mailing. Two of these email addresses were not registered in any list servers or newsgroups to test if the spammers can track these email addresses to send Spam. Nine of the most common places where the spammers look for email addresses were identified and used in the baiting method. All created email addresses were scattered among these nine places. These approaches are described in the following subsections.

4.3.1 Chat Rooms

Chat rooms are considered to be the easiest source of email addresses which attract the spammers. Four email accounts were registered with Yahoo chat rooms. Table 4.2 shows the list of email addresses that were used for this category.
4.3.2 Respond to Spammer

Six email addresses were used to respond to Spammers. This is done by directly replying to the spammer. In some spam messages, there is a link provided within the text to reply to the message. In this research, both methods were used to respond to the spammer. Table 4.3 shows the list of emails used for this purpose.

Table 4.3 Email Accounts used to respond to Spammers

<table>
<thead>
<tr>
<th><a href="mailto:svjonathan@yahoo.com">svjonathan@yahoo.com</a></th>
<th><a href="mailto:warreng.123@gmail.com">warreng.123@gmail.com</a></th>
<th><a href="mailto:bridgetted.12@gmail.com">bridgetted.12@gmail.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:bubblebob1999@yahoo.com">bubblebob1999@yahoo.com</a></td>
<td><a href="mailto:jk.sumanth@gmail.com">jk.sumanth@gmail.com</a></td>
<td><a href="mailto:reygongalez@gmail.com">reygongalez@gmail.com</a></td>
</tr>
</tbody>
</table>

4.3.3 Newsgroups

Seven email addresses were used to register in various USENET newsgroups. Some of these newsgroups were related to financing, investing, health, banking, etc. Table 4.4 gives the list of emails used for this purpose.

Table 4.4 Email Accounts used to register in Newsgroups

<table>
<thead>
<tr>
<th><a href="mailto:prasadvlsparasad@yahoo.com">prasadvlsparasad@yahoo.com</a></th>
<th><a href="mailto:karismam@gmail.com">karismam@gmail.com</a></th>
<th><a href="mailto:lorilever@gmail.com">lorilever@gmail.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:ksrsharma2000@yahoo.com">ksrsharma2000@yahoo.com</a></td>
<td><a href="mailto:bridgetted.12@gmail.com">bridgetted.12@gmail.com</a></td>
<td><a href="mailto:rajkishore10@lycos.com">rajkishore10@lycos.com</a></td>
</tr>
<tr>
<td><a href="mailto:Hardinsteve0@lycos.com">Hardinsteve0@lycos.com</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3.4 Online Surveys

Eight email addresses were used for filling out various online survey forms. Some surveys advertised special products and offers like ‘Win a free I-pod’, ‘Win a free Laptop’, etc. These surveys prompted respondents to give their name and email address while registering, but after completing the survey, it continued by popping up different pages which contained more free offers. Table 4.5 shows the list of email addresses used for this purpose.

Table 4.5 Email Accounts used for completing Online Surveys

<table>
<thead>
<tr>
<th><a href="mailto:abhignalle@yahoo.com">abhignalle@yahoo.com</a></th>
<th><a href="mailto:jk.sumanth@gmail.com">jk.sumanth@gmail.com</a></th>
<th><a href="mailto:veerkumar@lycos.com">veerkumar@lycos.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:karismam@gmail.com">karismam@gmail.com</a></td>
<td><a href="mailto:bridgetted.12@gmail.com">bridgetted.12@gmail.com</a></td>
<td><a href="mailto:hardinsteve0@lycos.com">hardinsteve0@lycos.com</a></td>
</tr>
<tr>
<td><a href="mailto:rickrob1980@hotmail.com">rickrob1980@hotmail.com</a></td>
<td><a href="mailto:rubies4mel@hotmail.com">rubies4mel@hotmail.com</a></td>
<td></td>
</tr>
</tbody>
</table>

4.3.5 Popup Windows

Two email accounts were used to respond to popup requests on the different Websites that were visited. Table 4.6 shows the list of email addresses used for this purpose.

Table 4.6 Email Accounts used to respond popup requests

| tweakertwe@yahoo.com          | preloader2006@yahoo.com      |

4.3.6 Website Registration

Five email accounts were used for various Website Registrations. These email addresses are also added to various Websites mailing lists. Table 4.7 presents the list of email addresses that were used for this purpose.
Table 4.7 Email Accounts used to register on Websites

<table>
<thead>
<tr>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:Ksrsharma2000@yahoo.com">Ksrsharma2000@yahoo.com</a></td>
</tr>
<tr>
<td><a href="mailto:lorilever@gmail.com">lorilever@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:veerkumar@lycos.com">veerkumar@lycos.com</a></td>
</tr>
<tr>
<td><a href="mailto:Katheleen4@lycos.com">Katheleen4@lycos.com</a></td>
</tr>
<tr>
<td><a href="mailto:reyrob11@hotmail.com">reyrob11@hotmail.com</a></td>
</tr>
</tbody>
</table>

4.3.7 Unsubscribe Link

Three email accounts were used to request a subscriber to delete the email address from their mailing lists. This was done by using the unsubscribe link given in the email message. Table 4.8 shows the list of email addresses used for this purpose.

Table 4.8 Email Accounts used for Unsubscribing

<table>
<thead>
<tr>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:tweakertwe@yahoo.com">tweakertwe@yahoo.com</a></td>
</tr>
<tr>
<td><a href="mailto:lorilever@gmail.com">lorilever@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:rajkishore10@lycos.com">rajkishore10@lycos.com</a></td>
</tr>
</tbody>
</table>

4.3.8 Gambling Websites

Five email accounts were used to register at various gambling Websites such as poker.com, gambling.com etc. Table 4.9 shows the list of email addresses used for this purpose.

Table 4.9 Email Accounts used to register on Gambling Websites

<table>
<thead>
<tr>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:tweakertwe@yahoo.com">tweakertwe@yahoo.com</a></td>
</tr>
<tr>
<td><a href="mailto:tomstanley.tom@gmail.com">tomstanley.tom@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:rickrob1980@hotmail.com">rickrob1980@hotmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:veerraj.1@gmail.com">veerraj.1@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:sweetgirl8330@lycos.com">sweetgirl8330@lycos.com</a></td>
</tr>
</tbody>
</table>

4.3.9 Email Address Obfuscation

There were four email addresses used for obfuscation. Two methods were used to hide the email address from Spammer. In the first method, the word ‘at’ was substituted
for the ‘@’ symbol while registering in the Websites; in the second method, the word ‘dot’ was replaced with ‘.’ symbol. The sole purpose of this obfuscation method is to check whether the spammer can track the email addresses. Table 4.10 shows the list of email addresses used for this purpose.

Table 4.10 Email Accounts used in Obfuscation Method

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:rickrob1980@hotmail.com">rickrob1980@hotmail.com</a></td>
<td><a href="mailto:sr.ee@hotmail.com">sr.ee@hotmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:reyrobl1@hotmail.com">reyrobl1@hotmail.com</a></td>
<td><a href="mailto:rubies4me1@hotmail.com">rubies4me1@hotmail.com</a></td>
</tr>
</tbody>
</table>

4.3.10 Email Addresses that were not registered by baiting method

Two email addresses were used in this research which did not use any of the baiting method. The main purpose is to test whether the spammers can trace these email addresses and send Spam. Table 4.11 depicts the list of email addresses used for this purpose.

Table 4.11 Email Accounts not used in Baiting Method

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:kumarsingh11@hotmail.com">kumarsingh11@hotmail.com</a></td>
<td><a href="mailto:suksagarb@hotmail.com">suksagarb@hotmail.com</a></td>
</tr>
</tbody>
</table>

4.4 Monitoring Email Accounts

Once the email addresses were registered, they were constantly monitored for Spam messages. All the forty email accounts created were observed twice a week for a period of forty five days. The total number of emails, including the emails in the Inbox and Junk/Spam folder, were counted and recorded in an Excel spreadsheet which is shown in Appendix A. Some of the emails in the Inbox were Spam-related while others were from email providers and delivery notifications. All the junk mails (including those
in the Inbox) were classified according to their categories detailed in section 4.5.1. As the number of emails increased every week, the categorization chart differed according to the new values. False Positives and False Negatives were also considered and counted during this monitoring process. These values were recorded in a separate Excel spreadsheet which is shown in Appendix B. By the end of the forty five day evaluation period, all required data was collected and was ready for the statistical analysis of the Research Project. The generated results were then tabulated to check the percentage of Spam received from each registered category.

4.5 Email Address Scrutiny and Categorization

After the email addresses were analyzed, they are sorted into various categories. Email address categorization was divided into three different parts:

- Categorization of Email Addresses
- Breakdown of Spam by Baiting Method
- Pictographic representation of email address classification

4.5.1 Categorization of Email Addresses

Table 4.12 represents the total number of emails received by a particular email address, the total number of Spam messages for each account, and the categories of the spam messages. From a total number of 12945 emails collected in a forty five day period, 11958 emails were considered to be Spam messages, which is 92.37% of all emails. This is generally regarded as high end of the Spam scale. The amount of spam received by a particular email address depends on the usage of that address on the Website. Not all the
messages received by an email address can be considered Spam, because some were from subscription based lists.

Some findings from the Table 4.12 are as follows:

- When an email address uses email address Obfuscation technique, it is less likely to receive spam (e.g., ‘reyrob11@hotmail.com’ received zero spam).
- The type of Spam a user receives depends on the type of spammer who receives the user’s email address. If a spammer who sends spam regarding health topics obtains a particular email address, the user may receive only spam consisting of health issues.
- If a user’s email address is too big to read, it is less susceptible to receiving spam messages (e.g., ‘prasadvlsprasad@yahoo.com’ received zero spam, as the email address is too big and confusing for a spammer to read).
- Dictionary attacks generate more spam messages (e.g., ‘svjonathan@yahoo.com’ received 2418 spam messages out of 2604 which is 92.85 % of total emails).

<table>
<thead>
<tr>
<th>Email Address</th>
<th>Total Email Received</th>
<th>Total Spam Received</th>
<th>Spam Categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:svjonathan@yahoo.com">svjonathan@yahoo.com</a></td>
<td>2604</td>
<td>2418</td>
<td>C, B, H, G, Fr, S,O,W,A,JL</td>
</tr>
<tr>
<td><a href="mailto:prasadvlspasad@yahoo.com">prasadvlspasad@yahoo.com</a></td>
<td>51</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:tweakertwe@yahoo.com">tweakertwe@yahoo.com</a></td>
<td>107</td>
<td>103</td>
<td>B, H, Fr, S,W,A</td>
</tr>
<tr>
<td><a href="mailto:twistytween@yahoo.com">twistytween@yahoo.com</a></td>
<td>42</td>
<td>32</td>
<td>C, H, G, Fr, W,</td>
</tr>
<tr>
<td>Email/Username</td>
<td>Phone 1</td>
<td>Phone 2</td>
<td>Initials</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td><a href="mailto:bubblebob1999@yahoo.com">bubblebob1999@yahoo.com</a></td>
<td>609</td>
<td>564</td>
<td>C, G, Fr, S, O, W, JL</td>
</tr>
<tr>
<td><a href="mailto:abarghav@yahoo.com">abarghav@yahoo.com</a></td>
<td>108</td>
<td>56</td>
<td>H, G, O, W</td>
</tr>
<tr>
<td><a href="mailto:ksrsharma2000@yahoo.com">ksrsharma2000@yahoo.com</a></td>
<td>61</td>
<td>42</td>
<td>C, B, H, G, S, W, JL, A</td>
</tr>
<tr>
<td><a href="mailto:abhignalle@yahoo.com">abhignalle@yahoo.com</a></td>
<td>3</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:vilevan2006@yahoo.com">vilevan2006@yahoo.com</a></td>
<td>3</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:preloader2006@yahoo.com">preloader2006@yahoo.com</a></td>
<td>27</td>
<td>21</td>
<td>B, H, Fr</td>
</tr>
<tr>
<td><a href="mailto:veerraj.1@gmail.com">veerraj.1@gmail.com</a></td>
<td>80</td>
<td>55</td>
<td>B, H, S, O, W, JL, A</td>
</tr>
<tr>
<td><a href="mailto:karismam@gmail.com">karismam@gmail.com</a></td>
<td>2508</td>
<td>2424</td>
<td>C, B, H, G, Fr, S,O,W,A, JL</td>
</tr>
<tr>
<td><a href="mailto:timshareif@gmail.com">timshareif@gmail.com</a></td>
<td>57</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:warreng.123@gmail.com">warreng.123@gmail.com</a></td>
<td>548</td>
<td>471</td>
<td>C, B, H, G, Fr, S,O,W,A, JL</td>
</tr>
<tr>
<td><a href="mailto:tomtomstanley.tom@gmail.com">tomtomstanley.tom@gmail.com</a></td>
<td>5</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:ema.123ilme@gmail.com">ema.123ilme@gmail.com</a></td>
<td>2</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td><a href="mailto:robrodrigo@gmail.com">robrodrigo@gmail.com</a></td>
<td>844</td>
<td>821</td>
<td>B, H, G, Fr, S,O,W,A, JL</td>
</tr>
</tbody>
</table>
| jk.sumanth@gmail.com     | 1556    | 1386    | B, H, G, Fr,
| Email Address                    | Code 1 | Code 2 | Codes  
|---------------------------------|--------|--------|--------
| bridgetted.12@gmail.com         | 618    | 609    | B, H, G, Fr, S  
| reygongalez@gmail.com           | 72     | 69     | G, Fr, JL  
| kkprasad.34@gmail.com           | 2908   | 2824   | C, B, H, G, Fr, S,O,W  
| krishnanswamy.12@gmail.com      | 75     | 63     | B, H, G, S,W  
| lorilever@gmail.com             | 11     | 0      | N/A  
| sureshrv@lycos.com              | 4      | 0      | N/A  
| ramreddy0@lycos.com             | 4      | 0      | N/A  
| rajkishore10@lycos.com          | 5      | 0      | N/A  
| veerkumar@lycos.com             | 2      | 0      | N/A  
| banukanthpatel@lycos.com        | 7      | 0      | N/A  
| brainlawrence@lycos.com         | 3      | 0      | N/A  
| hardinsteve0@lycos.com          | 2      | 0      | N/A  
| sweetgirl8330@lycos.com         | 3      | 0      | N/A  
| gurupreethsingh@lycos.com       | 6      | 0      | N/A  
| kathleen4@lycos.com             | 7      | 0      | N/A  
| sr.ee@hotmail.com               | 1      | 0      | N/A  
| rickrob1980@hotmail.com         | 1      | 0      | N/A  
| kumarsingh11@hotmail.com        | 1      | 0      | N/A  
| suksagarb@hotmail.com           | 1      | 0      | N/A  
| reyrob11@hotmail.com            | 1      | 0      | N/A  
| rubies4me1@hotmail.com          | 1      | 0      | N/A  

71
The Spam breakdown abbreviations used in the Table 4.12 are as follows:

- C = Commodities
- B = Business
- H = Health
- G = Gambling
- Fr = Fraud
- S = Surveys
- O = Others
- W = Website Hosting
- A = Adult
- JL = Request to Join Lists

### 4.5.2 Breakdown of Spam by Baiting Method

In this section, a bar graph represents the breakdown of Spam by baiting method. The Spam was monitored and a chart was prepared from the acquired results. The total number of Spam emails is shown on the y-axis and the baiting method that was used to collect spam is shown on the x-axis of the bar graph. Figure 4.2 is a bar graph depicting the breakdown of the baiting methods and the total number of spam messages that were received.
Figure 4.2 Breakdown of Spam by Baiting Method

The total number of Spam emails received by using that particular baiting method is clearly labeled on the top of each bar (e.g., chat rooms baiting method received a total of 501 spam emails). It is not correct to assume if a greater percentage of Spam received by using one baiting method, that method is commonly used by spammers. In this Research, most of the spam received is from the newsgroup and the least is from chat rooms.

4.5.3 Pictographic representation of email addresses classification

Figure 4.3 depicts the Pie chart representation of the Spam categorization. The various categories of spam were clearly explained in Table 4.1. In this Pie chart, the percentages of the categories of spam were calculated and are shown clearly. The highest
percentage of Spam was received by the health and paid surveys categories, which are 27.10% and 27.50% respectively; the least amount of the Spam was received by the adult category at 0.85%. Even though email addresses were not registered in any pornographic Websites, some emails still contained pornographic content. It is not always true that the major part of any user’s Inbox will be filled with health related Spam messages.

Figure 4.3 Pie Chart of Spam Categorization
4.6 **Tools Used for Testing**

Two types of tools were used in this research project to test against the created email address. One is an email extractor tool known as ‘Power Email Extractor’ and the other is an anti-Spam tool known as ‘Spam Inspector’. Power Email Extractor is one of the prevailing spammer’s application tools used to extract email addresses, where as Spam Inspector is an anti-spam tool which is used to prevent Spam messages received by the user.

4.6.1 **Power Email Extractor Tool Experiment**

The tool Power Email Extractor was found at ‘http://tecsoftware.biz/extractor.htm’. A free trial of the latest version 4.1.0 was downloaded to the local research machine. The research computer used to install this tool ran windows XP media center as an operating system. This tool required a Cable/DSL connection or any other high-speed Internet connection. When the tool was initially started, the newsgroup list was automatically initiated. After initialization process was complete, it connected to the default server ‘freetext.usenetserver.com’ and extracted email addresses from that server. This was the default server used during the entire Research process. It keeps the history of this default server in memory and starts the newsgroups list automatically whenever the tool was initiated, without user’s intervention.

The download process took approximately five to ten seconds to connect to different newsgroups. Once the list of newsgroups or email addresses was obtained, the tool automatically disconnected from the internet to save the bandwidth/speed of the data connection. The configuration page was opened to setup the settings of the tool. Since
there was no Firewall on the research computer, only the general settings of the tool were configured. Figure 4.4 shows the screenshot of general settings in configuration page.

![General Setting of Power Email Extractor](image)

**Figure 4.4 General Setting of Power Email Extractor**

The ‘Auto Reconnect’ option was enabled; if the Internet connection is lost, the tool was reconnected automatically and any information that was not saved was not lost. Five ‘Max Attempts to Make’, was selected, which prompted the tool to attempt to connect five times before warning about the loss of Internet Connection. One minute was selected for ‘Minutes between Attempts’; between each attempt to reconnect to the server, the tool waits for one minute. With these settings, the tool must wait for five minutes before informing the user about the loss of network connection. The ‘Auto Save’ option was automatically disabled in the trail version.
Once the configuration settings were complete, Power Email Extractor was ready for testing and the results were documented. The series of steps used for this testing are briefly described in the sections below.

4.6.1.1 Power Email Extractor Outcomes:

Four features of Power Email Extractor were tested in this Research, as these were the only features available in the free trial that was installed. The tool has many more features available in the registered version such as auto-save, saving the email addresses in the form of a text document etc.

The four features tested using this tool were as follows:

- General extraction of email addresses from news groups
- Extraction of email addresses using group selection
- Extraction of email addresses using the Keyword search tool
- Extraction of email addresses using Exclusion Lists

All four tools are explained in detail in the following subsections.

**General Extraction of email addresses from newsgroups.** When the tool was opened, it downloaded various newsgroups from the server and displayed the list of newsgroups on the left side of the screen. Once the newsgroups were extracted, there was an option to select any newsgroup and retrieve the email addresses from that particular newsgroup. Figure 4.5 shows the screenshot of the Power Email Extractor before selecting any newsgroup.
Figure 4.5 Power email extractor before selection of any newsgroups

Form the list of newsgroups available, a newsgroup named ‘misc.legal.computing’ was selected to extract the email addresses. Figure 4.6 depicts the screenshot of the selected newsgroup tab under ‘Selected Groups’ Option.
Figure 4.6 Selected newsgroup tab under ‘Selected Groups’ Option

After the extraction process was completed, the selected newsgroup from the left pane is removed by the tool and displayed all the extracted email addresses from that newsgroup on the right pane of the screen. Figure 4.7 shows the screenshot of extracted email addresses from the selected newsgroup.
The collection of email address in the right pane is subdivided into five different fields. They are Email Address, Name, News Group, Date Posted and Subject respectively.

**Extraction of email addresses using group selection.** Using this type of selection, multiple groups were selected from the list of newsgroups displayed in the panel by using the Ctrl key. To test this tool, five newsgroups were selected. Figure 4.8 shows the screenshot of newsgroups with ‘Group Selection’ feature. The selection of newsgroups is highlighted.
During this extraction, the tool listed all email addresses starting from the first selected newsgroup to the last in order.

**Extraction of email addresses using the Keyword Search Tool.** Power Email Extractor’s Keyword search tool was tested by using all three different options that were available. Figure 4.9 depicts a screenshot of the Keyword search tool and the three options available: ‘Match Anywhere in Name’, ‘Match at Beginning of Name’, and ‘Match Exact (Sub) category Name’ respectively.
Figure 4.9 Keyword search tool

Extraction of email addresses using the Exclusion Lists. There are three types of exclusion lists available with this tool; all the three options were tested in this Research project. Figure 4.10 depicts the screenshot of all the three exclusion lists. They are ‘Word Exclusion List’, ‘Top Level Domain Exclude List’, and ‘Domain Exclude List’ respectively.
4.6.2 Spam Inspector Tool Experiment

Spam Inspector is one of the powerful anti-Spam tools. It used 4 MB of space for installation and was compatible with windows 95/98/Me/NT/2000/XP platforms. After the tool was installed on the Research computer, it was tested with the email accounts that were created. The email address ‘robrodrigoes@gmail.com’ was used to configure the Microsoft Outlook Express 2003.

After the configuration was complete, the Spam inspector ran a quick inspection on the email address (robrodrigoes@gmail.com). It created a special folder called ‘Spam Quarantine’ and checked the Inbox for any Spam emails. As Spam was detected, it was sent to the Spam Quarantine folder. The Spam Quarantine folder can be renamed if desired.
After the configuration was complete, the Spam Inspector’s customized toolbar was automatically installed and placed between the Microsoft Outlook Toolbar and the Inbox. Figure 4.11 shows the screenshot of the Spam Inspector Toolbar.

![Spam Inspector Toolbar Screenshot](image)

**Figure 4.11 Spam Inspector Toolbar**

### 4.6.2.1 Spam Inspector Tool Outcomes

Spam Inspector was tested on subject email address (robrodrigoes@gmail.com) using ten email addresses, which were taken from the email addresses created for this Research. From those ten, five emails addresses were used to manage the friends list and five email addresses were used to manage the enemies list. The friends and enemies email addresses list are shown in Table 4.13.
Table 4.13 Friends and Enemies email addresses list used for Spam Inspector tool

<table>
<thead>
<tr>
<th>Friends List</th>
<th>Enemies List</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:svjonathan@yahoo.com">svjonathan@yahoo.com</a></td>
<td><a href="mailto:tweakertwe@yahoo.com">tweakertwe@yahoo.com</a></td>
</tr>
<tr>
<td><a href="mailto:bubblebob1999@yahoo.com">bubblebob1999@yahoo.com</a></td>
<td><a href="mailto:preloader2006@yahoo.com">preloader2006@yahoo.com</a></td>
</tr>
<tr>
<td><a href="mailto:villevan2006@yahoo.com">villevan2006@yahoo.com</a></td>
<td><a href="mailto:kkprasad.34@gmail.com">kkprasad.34@gmail.com</a></td>
</tr>
<tr>
<td><a href="mailto:lorilever@gmail.com">lorilever@gmail.com</a></td>
<td><a href="mailto:hardinsteve0@lycos.com">hardinsteve0@lycos.com</a></td>
</tr>
<tr>
<td><a href="mailto:kumarsingh11@hotmail.com">kumarsingh11@hotmail.com</a></td>
<td><a href="mailto:sr.ee@hotmail.com">sr.ee@hotmail.com</a></td>
</tr>
</tbody>
</table>

Various emails were sent to ‘robodrigoes@gmail.com’ from the ten email addresses defined in friends list and enemies list. Neither friends nor enemies list was managed during the installation of Spam Inspector tool. Based on the Spam Inspector’s filtering technique, when email was sent, it was directed either to Inbox or the Spam Quarantine folder.

The testing was divided into four different parts:

- Cleaning the Inbox
- Managing the friends list
- Managing the enemies list
- Spam Abuse Report

**Cleaning the Inbox.** Cleaning of the Inbox was done by transferring emails received to their respective folders. The Spam Inspector scanned all new email sent to ‘robodrigoes@gmail.com’; if the sender’s email address was in the friends list, it was directed to the Inbox and if the email address was in the enemies list, it was sent to the
Spam Quarantine folder. If the sender is not defined on either the friends or enemies list, Spam Inspector used its own filtering technique and based upon the result, it directed the message to the Inbox, Junk Mail or Spam Quarantine folders.

**Managing the Friends List.** Five email addresses were used to manage the friends list. Whenever an email arrived from any one of these five email addresses, it was directed toward the Inbox. This was one of the learning techniques used by Spam Inspector. Figure 4.12 shows the screenshot of Spam Inspector’s Manage Friends list option.

![Figure 4.12 Manage Friends list option](image)

**Managing Enemies list.** Five email addresses were used to manage the enemies list. Whenever an email arrived from any one of these five email addresses, it was directed toward the Spam Quarantine folder. Figure 4.13 depicts the screenshot of Spam Inspector’s Manage Enemies list option.
Spam Abuse Report: Spam Inspector had an option to send an abuse report to SpamAbuse.org. If a user feels that a particular email addresses was sending all abusive content messages, the message can be reported to SpamAbuse.org. Figure 4.14 depicts the screenshot of Spam Abuse Report.
4.7 False Positive vs. False Negative

A False Positive is the percentage of non-Spam that is reported as Spam, while a False Negative is the percentage of Spam that is reported as non-Spam. In this Research, twelve email addresses received false positives and false negatives. These email addresses are shown on the x-axis and the corresponding number of emails that were considered to be false positives and false negatives are shown on y-axis of the graph depicted in figure 4.15.
Figure 4.15 Graph: False Positives vs. False Negatives

This graph tells us that in this Research project, the number of email addresses that received false negatives dominated false positives.
5. FUTURE WORK

It is hard to explain the scenario of Spam messages and the problems caused by it in the future. Many organizations and countries are trying to situate a periphery on dealing with this problem. In the future, this research can be expanded by adding some new features. As the spam problem is increasing, there will be more and more anti-spam techniques which may arrive that need to be studied. Spam filters may change their expertise and also the level of these filters may vary. At present, the anti-spam techniques are one step behind the current spamming tools. Therefore, anti-spam software may be developed based on the study in this report.
6. CONCLUSION

Even though, this report explains some of the chief aspects of spam, it may not be the perfect guide. Spam is gigantic subject today, and there are lots of tools available to avoid Spam. In this research project various analysis of spam has been studied. A brief document is presented which contains information about spam. On the internet, there are many research papers about various discussions on spam-specific topics, but there was no solid report that contains all the major aspects of spam. It explains the definition of spam and gives various reasons for a user to avoid spam. It is always important for users to know the characteristics of spam, so that they will know the harm caused by it and will also know the root cause of Spam messages. Some of the important Spam Filters and various anti-spam techniques are briefly discussed. Different types of Spammers, their categories and the techniques used by the spammers were highlighted. In accordingly, the legal action against Spammers was explained. Few tips to avoid Spam have been suggested towards the contribution of this research project.

Tips to Avoid Spam

There is no perfect way to fight against Spam which is now a days a major problem on the Internet. The amount of Spam received will depend on how the user responds to the Spam messages he receives. Based on this research, few tips are suggested that will help reduce the amount of spam the user can receive. They are briefly described below:
• Do not ever reply to spammers: Never respond to Spam messages. Even if the user requests to remove the email address from the mailing list, the spammer confirms that it is the true address and adds to the list of targeted emails.

• Just delete the spam message even without opening it. This is the easiest way for a typical user.

• Don’t use the unsubscribe link: Never click on any link in the spam message, even if it is unsubscribe link. By doing it, the spammer confirms that the spam has been delivered and that the recipient has responded.

• Do not post e-mail address on Websites or Usenet: If there is need to post the email address, complicate it so that it would be difficult for the spammer to harvest it. For e.g., name-at-domain name-dot-com.

• Don’t distribute e-mail address indiscriminately: The best way to avoid spam is to keep the email address private. Create two or more accounts. Choose one account for personal emails and the other for online shopping or subscribing to newsletters etc.

• Enter the email address in any of the search engines and check whether the email address is posted on any of the newsgroups. If so, try to remove it, so that there will be a chance of reducing the spam.

• Do not open spam messages: Spammers include ‘web beacons’ which is nothing but a small graphical image that is placed in an email to monitor the behavior of the user who receives the email. Spammers use this technique, to determine which emails are opened at the receiving side.
• Don’t do business with spammers: Never buy products from spammers or do any kind of business with spammers. Do not believe anything that spammers tell. It is highly risky to give credit card details to an unknown and unreliable source.

• Avoid Phishing: Never give bank account details or any confidential information even if the company requests it. Contact the company directly to check whether they requested the details. It is always better to check the website URL whether it is correct or fake.

• Use anti-spam software: If the ISP doesn’t provide any filtering technique to the inbox, use either spam filtering or anti-spam software to reduce the content of spam.

• Avoid using common first names and last names in the email address, as most spammers use the dictionary attack method to send spam, as it is one of the easiest methods for spammers to send emails.

• Use filtering systems: Many ISP’s and email services provide spam filtering. Although not all filters are 100% perfect, these spam filters reduce the incoming spam by filtering the good emails from the bad emails and sending them to two separate folders.

• Read the privacy policy when ordering online: When ordering online, check the websites privacy policy to make sure that the addresses are not sold. Some websites have checkboxes to opt out third party mailers.

• Pick an unusual user name: Select a user name which has either a numeric or special character. In this research, the user names which had numeric digits
received less spam. It is less likely to receive spam, if the user name is long and complex rather than short email addresses.

• Spam Poisoning: Spammers collect email addresses from various different WebPages. Replacing characters in human readable format can decrease the effect of spam. In this case, an e-mail address with all letters is preferred, for example the e-mail address king.kong@sci.tamu.edu is a general format used by many users. If the same notation is re-written as king dot kong at sci dot tamucc dot edu, it will be difficult for a spammer to trace the e-mail address. But once the email address is exposed by some way, all the efforts to hide the email address will be wasted. Obtained addresses are permanently added to the spammer’s lists.

• Masking email address: In this method the email addresses are masked by putting a word or phrase in the email address. For example, the email address “xyz@gmail.com” could be masked as “xyz@abc.gmail.com” and let the recipient know what the masking word or phrase is. In this way the spammers cannot harvest the user’s email address.

• Always use a separate screen name for chat: As chatting is one of the best way spammers can trap the email addresses, it is always better to use a separate screen name for online chatting.
BIBLIOGRAPHY AND REFERENCES


APPENDIX A: Email Addresses Monitoring

<table>
<thead>
<tr>
<th>Email Address</th>
<th>Total Email</th>
<th>Inbox Email</th>
<th>Spam Email</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:sjohanathan@yahoo.com">sjohanathan@yahoo.com</a></td>
<td>2694</td>
<td>186</td>
<td>2418</td>
</tr>
<tr>
<td>presadh@<a href="mailto:prasadh@yahoo.com">prasadh@yahoo.com</a></td>
<td>51</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:twoktorwu@yahoo.com">twoktorwu@yahoo.com</a></td>
<td>107</td>
<td>4</td>
<td>103</td>
</tr>
<tr>
<td><a href="mailto:twistyween@yahoo.com">twistyween@yahoo.com</a></td>
<td>42</td>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td><a href="mailto:bubblebob1999@yahoo.com">bubblebob1999@yahoo.com</a></td>
<td>6009</td>
<td>45</td>
<td>564</td>
</tr>
<tr>
<td><a href="mailto:dagarsh@gmail.com">dagarsh@gmail.com</a></td>
<td>108</td>
<td>62</td>
<td>66</td>
</tr>
<tr>
<td><a href="mailto:karsharma2000@yahoo.com">karsharma2000@yahoo.com</a></td>
<td>61</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td><a href="mailto:ash@gmail.egy">ash@gmail.egy</a>@yahoo.com</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:villevan2006@yahoo.com">villevan2006@yahoo.com</a></td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:preloader2006@yahoo.com">preloader2006@yahoo.com</a></td>
<td>27</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td><a href="mailto:veenaj_10@gmail.com">veenaj_10@gmail.com</a></td>
<td>60</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td><a href="mailto:kartasmj@gmail.com">kartasmj@gmail.com</a></td>
<td>2508</td>
<td>64</td>
<td>2424</td>
</tr>
<tr>
<td><a href="mailto:timharei@gmail.com">timharei@gmail.com</a></td>
<td>57</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:warnerj_123@gmail.com">warnerj_123@gmail.com</a></td>
<td>540</td>
<td>77</td>
<td>471</td>
</tr>
<tr>
<td><a href="mailto:tomstanley.tom@gmail.com">tomstanley.tom@gmail.com</a></td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:erma.123live@gmail.com">erma.123live@gmail.com</a></td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:rocchobigoo@gmail.com">rocchobigoo@gmail.com</a></td>
<td>844</td>
<td>23</td>
<td>921</td>
</tr>
<tr>
<td>jk <a href="mailto:sumonh@gmail.com">sumonh@gmail.com</a></td>
<td>1556</td>
<td>170</td>
<td>1386</td>
</tr>
<tr>
<td><a href="mailto:bridgeted.12@gmail.com">bridgeted.12@gmail.com</a></td>
<td>919</td>
<td>9</td>
<td>609</td>
</tr>
<tr>
<td><a href="mailto:reygonalr@gmail.com">reygonalr@gmail.com</a></td>
<td>72</td>
<td>3</td>
<td>69</td>
</tr>
<tr>
<td><a href="mailto:kkpriessd.34@gmail.com">kkpriessd.34@gmail.com</a></td>
<td>2908</td>
<td>84</td>
<td>2824</td>
</tr>
<tr>
<td><a href="mailto:kthomaswamy.12@gmail.com">kthomaswamy.12@gmail.com</a></td>
<td>75</td>
<td>12</td>
<td>63</td>
</tr>
<tr>
<td><a href="mailto:litiere@gmail.com">litiere@gmail.com</a></td>
<td>11</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:otnorlaw@ycos.com">otnorlaw@ycos.com</a></td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:narmedly@ycos.com">narmedly@ycos.com</a></td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:rakhshone10@ycos.com">rakhshone10@ycos.com</a></td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:veerkumar@ycos.com">veerkumar@ycos.com</a></td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:banukantphater@ycos.com">banukantphater@ycos.com</a></td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:brandlawrence@ycos.com">brandlawrence@ycos.com</a></td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:hardinestvo0@ycos.com">hardinestvo0@ycos.com</a></td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:sweetgirl3330@ycos.com">sweetgirl3330@ycos.com</a></td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:gutupaapithsingh@ycos.com">gutupaapithsingh@ycos.com</a></td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:kathleen4@ycos.com">kathleen4@ycos.com</a></td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>sr@<a href="mailto:ee@hotmail.com">ee@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:nickrob1980@hotmail.com">nickrob1980@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:kumarsingh11@hotmail.com">kumarsingh11@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:suksagar@hotmail.com">suksagar@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:roynob1@hotmail.com">roynob1@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><a href="mailto:rubofme1@hotmail.com">rubofme1@hotmail.com</a></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Total: 12945 997 11968
## APPENDIX B: False Positive and False Negative

<table>
<thead>
<tr>
<th>A</th>
<th>Email Address</th>
<th>B</th>
<th>False Positive</th>
<th>C</th>
<th>False Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><a href="mailto:sjonnathan@yahoo.com">sjonnathan@yahoo.com</a></td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><a href="mailto:prasadspresad@yahoo.com">prasadspresad@yahoo.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td><a href="mailto:tweakto@yahoo.com">tweakto@yahoo.com</a></td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td><a href="mailto:twstyyeen@yahoo.com">twstyyeen@yahoo.com</a></td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td><a href="mailto:bubblech1999@yahoo.com">bubblech1999@yahoo.com</a></td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><a href="mailto:abanghav@yahoo.com">abanghav@yahoo.com</a></td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td><a href="mailto:kashsharma2000@yahoo.com">kashsharma2000@yahoo.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td><a href="mailto:achari@uni.de">achari@uni.de</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td><a href="mailto:villevan2006@yahoo.com">villevan2006@yahoo.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td><a href="mailto:preleade2006@yahoo.com">preleade2006@yahoo.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td><a href="mailto:veeraj1@gmail.com">veeraj1@gmail.com</a></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td><a href="mailto:karismam@gmail.com">karismam@gmail.com</a></td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td><a href="mailto:timcharak@gmail.com">timcharak@gmail.com</a></td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td><a href="mailto:waaren.123@gmail.com">waaren.123@gmail.com</a></td>
<td>8</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td><a href="mailto:tomstanley_hen@gmail.com">tomstanley_hen@gmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td><a href="mailto:ajithp@gmail.com">ajithp@gmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td><a href="mailto:roboligo@gmail.com">roboligo@gmail.com</a></td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td><a href="mailto:jksunmony@yahoo.com">jksunmony@yahoo.com</a></td>
<td>2</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td><a href="mailto:bridgetted.12@gmail.com">bridgetted.12@gmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td><a href="mailto:myongalee@gmail.com">myongalee@gmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td><a href="mailto:kkprasad34@gmail.com">kkprasad34@gmail.com</a></td>
<td>3</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td><a href="mailto:krishnaoswamy.12@gmail.com">krishnaoswamy.12@gmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td><a href="mailto:toni@ymail.com">toni@ymail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td><a href="mailto:suresh@lycos.com">suresh@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td><a href="mailto:nammuky@lycos.com">nammuky@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td><a href="mailto:rajkisihore10@lycos.com">rajkisihore10@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td><a href="mailto:veeekuman@lycos.com">veeekuman@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td><a href="mailto:banukanthcato@lycos.com">banukanthcato@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td><a href="mailto:braillawrence@lycos.com">braillawrence@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td><a href="mailto:hardinsteve@lycos.com">hardinsteve@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td><a href="mailto:saeedir833@lycos.com">saeedir833@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td><a href="mailto:gujonpreetisingh@lycos.com">gujonpreetisingh@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td><a href="mailto:kathleen4@lycos.com">kathleen4@lycos.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td><a href="mailto:sr_en@hotmail.com">sr_en@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td><a href="mailto:nickn1980@hotmail.com">nickn1980@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td><a href="mailto:kumareign11@hotmail.com">kumareign11@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td><a href="mailto:suksangarb@hotmail.com">suksangarb@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td><a href="mailto:reyr11@hotmail.com">reyr11@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td><a href="mailto:rubies4me1@hotmail.com">rubies4me1@hotmail.com</a></td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>