Växjö University

School of Mathematics and Systems Engineering

Reports from MSI - Rapporter från MSI

Anti-phishing system

Detecting phishing e-mail

Yuanxun Mei

Maj
2008

MSI
Växjö University
SE-351 95 VÄXJÖ

Report 08047
ISSN 1650-2647
ISRN VXU/MSI/DA/E/--08047/--SE

Yuanxun Mei

Anti-phishing system

Detecting phishing e-mail

2008

Supervisor: Ola Flygt

Växjö
University

Abstract

Because of the development of the Internet and the rapid increase of the electronic commercial, the incidents on stealing the consumers' personal identify data and financial account credentials are becoming more and more common. This phenomenon is called phishing. Now phishing is so popular that web sites such as papal , eBay, MSN, Best Buy, and America Online are frequently spoofed by phishers. What’s more, the amount of the phishing sites is increasing at a high rate.

The aim of the report is to analyze different phishing phenomenon and help the readers to identify phishing attempts. Another goal is to design an anti-phishing system which can detect the phishing e-mails and then perform some operations to protect the users. Since this is a big project, I will focus on the mail detecting part that is to analyze the detected phishing emails and extract details from these mails.

A list of the most important information of this phishing mail is extracted, which contains “mail subject”, “ mail received date”, “targeted user”, “the links”, and “expiration and creation date of the domain”. The system can presently extract this information from 40% of analyzed e-mails.

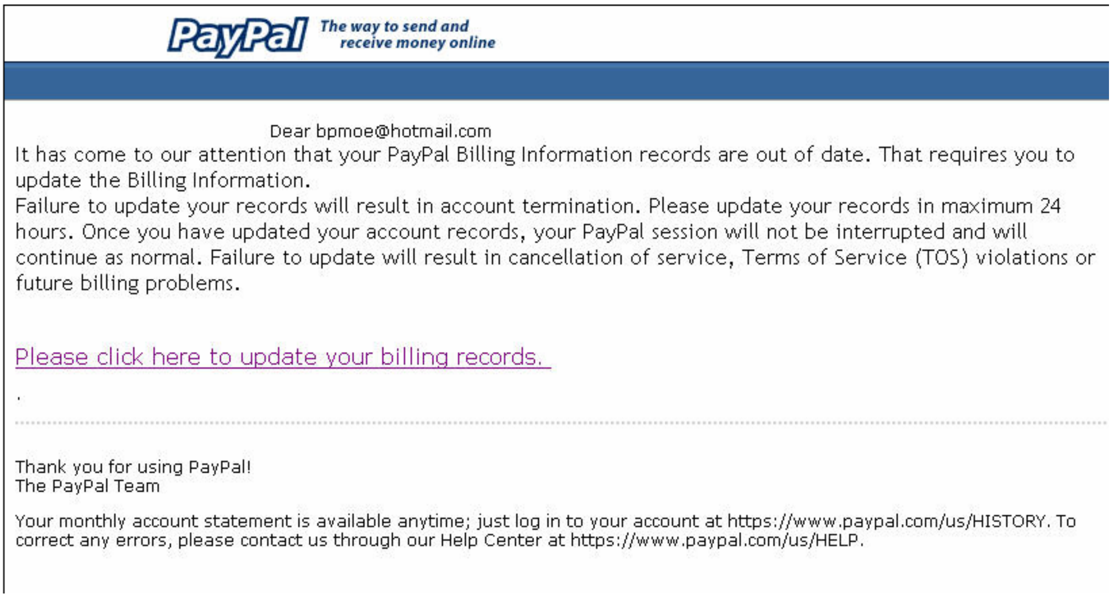
Keywords: Phishing, Anti-phishing, Pharming, Email, Domain name

1 Introduction

In this chapter I will introduce phishing phenomenon, the problem that phishing brings to the public.

1.1 Context

As the development of the Internet, more and more users are involving into the gigantic Internet Sea. We get many benefits through the Internet such as learning, buying stuffs online and so on. On the other hand, a potential threat is coming silently while the clients are enjoying the benefits of electronic commerce. One of them is called "Phishing" which takes advantage of different spoofing technologies to miss-lead the users to browse a webpage that is similar to the legitimate webpage. Then the users will be asked to input their accounts and passwords for some so-called emergency reasons that often showed as "If your account information is not updated within 48 hours then your ability to use it will become restricted."



Dear bpmoe@hotmail.com

It has come to our attention that your PayPal Billing Information records are out of date. That requires you to update the Billing Information.
Failure to update your records will result in account termination. Please update your records in maximum 24 hours. Once you have updated your account records, your PayPal session will not be interrupted and will continue as normal. Failure to update will result in cancellation of service, Terms of Service (TOS) violations or future billing problems.

Please click here to update your billing records.

.

Thank you for using PayPal!
The PayPal Team

Your monthly account statement is available anytime; just log in to your account at https://www.paypal.com/us/HISTORY. To correct any errors, please contact us through our Help Center at https://www.paypal.com/us/HELP.

Figure 1.1: A recent phishing example

In Figure1.1 we can see a typical phishing example. There are several traps waiting for the careless user, the icon is exactly the same as the official Paypal website, the context of the mail is also an official tone that makes it much more trustful. Many victims have been spoofed by these apparently correct e-mails.

1.2 Problem

In the report, I will explain the phishing attack to give the readers a clear idea on what's phishing? How does phisher success? How to avoid being phished? On the other hand, I will implement a sub system that can extract the details of a detected phishing e-mail, the extracted information will be used to alarm system which can shut down the faked web site, update the anti-virus products. I want to analyze phishing e-mails and work out their characteristics. Once I finish the analysis, I will make a conclusion of the main information of the mail and send the information to different organizations to block this phishing attack.

The damage caused by phishing ranges from denial of access to substantial financial loss. It is estimated that between May 2004 and May 2005, approximately 1.2 million computer users in the United States suffered losses caused by phishing, totaling approximately US$929 million [1]. In other word, there were almost 30 thousands victims every day. We could easily be involved in it. According to a survey carried out on behalf of Cloudmark [2], consumer confidence in brands would be severely dented by a phishing attack. Banks are most at risk, but ISPs, online shopping sites and even social networking sites would also see a fall in consumer confidence after a phishing attempt.

According to a survey [3] of more than 4,500 online U.S. adults in August 2007 (which was representative of the online U.S. adult population) the attacks were more successful in 2007 than they were in the previous two years. Of consumers who received phishing e-mails in 2007, 3.3 percent say they lost money because of the attack, compared with 2.3 percent who lost money in 2006, and 2.9 percent who did so in 2005.

As we know, the brand is the most valuable asset for a company. In a world full of competition, company can easily fall down for several incidents that will make the public lose confidence in them. Suppose you are an online client of a bank, if you are phished when you are using the service providing by the company, you will possible loose your confidence of the bank and never use it again. What’s more, you may tell your friends about your incident, and your friends will tell their friends and so on. The damage of the brand would be inconceivable and financial organizations and internet service providers should pay more attention on the “phishing” problem. If they can not decrease the attacks, the consumers’ trust in the online commerce will erode step by step. Eventually all the involvers in online commerce will be lost. Phishing not only damage the operations, but also bring a huge challenge to the clients’ trust in electronic commerce.

Phishing have caused its damage as showed in the survey by Pew Internet Life [4], the trust to the emails of the consumers have already fell into the lowest point. Cyota [5] did a survey on online bank account users recently. 74% percent of the people say they do not trust e-mails coming from the banks and the online commerce probably have already declined.

Except the trust loss, phishing will also make a direct loss for enterprises and consumers. If the phisher get the information credit card, the loss will be unavoidable. In addition, releasing a new credit card will cost 50 Dollars. If a lot of clients are phished, the cost will be extensive.

Anti-phishing Work Group is continuously making investigating about phishing. The total number of unique phishing reports submitted to Anti-Phishing Work Group in September 2007 was 38,514. However, the amount was about 22136 one year ago.

We can see the changes in Figure 1.2, the phishing reports amounts every month from September 2006 to September 2007. In the figure we see that January and June are the most active periods, perhaps because they are the holiday month, many users use the online service during these time.

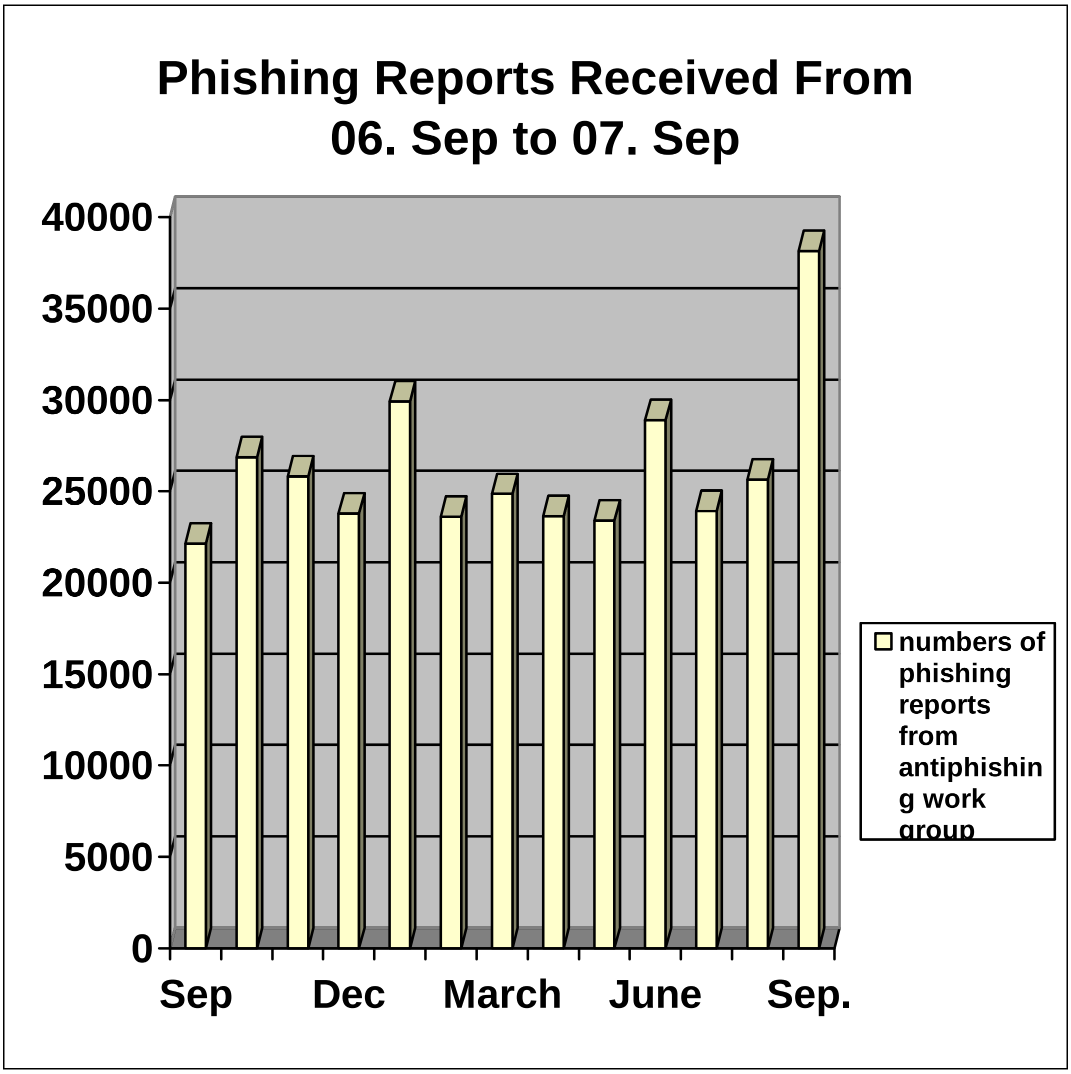


Figure 1.2 Phishing report from Anti-Phishing Work Group

1.3 Objective

When I was studying in China my bank account was attacked. It was in 2004, and at that time the online commerce just started. I am a person who is eager to try new things so I got an online account from ICBC (Industrial and Commercial Bank of China) which is one of the biggest banks in China running by government. At first, it gave me advantages indeed because I could buy products very cheap and spend shorter time on shopping through the online trading. But one month later, I found my account was drawn down. Then I heard this legitimate webpage was phished, that's why my money was lost together with a lot of other online users. Finally however ICBC returned my money. You can find this news on the internet [6]. From that time, I have paid attention to phishing. So when I found there is a subject about anti-phishing, I was very pleased to have a chance to learn more about it. This report is the result of a bachelor thesis work at Vaxjo University.

In Figure 1.2 we saw that phishing is becoming more and more common in the IT world. It is crucial for us to make much more efforts to prevent the spread. The cumulative lost is potentially huge so a system that could detect the phishing mails and prevent them to affect the users any more would be beneficial. More specific, I want to analyze phishing e-mails and work out their characteristics. Once I finish the analysis, I will make a conclusion of the main information of the mail and send the information to different organizations to block this phishing attack.

We already have a subsystem to detect whether the e-mails are phishing mails or not. What I intend to do is analyzing the detected phishing e-mails more closely, and extract the details such as the received date, the subject of the mail, the domain name and the registered information of the domain name. This information will be used in another subsystem called alarm system that will alarm the phishing incident to Internet and security companies or directly users.

1.4 Structure of report

In chapter1, several simple incidents were showed to give the reader a clear impression about phishing incident. Then we talked about the influence of phishing by listing various phishing report data. In chapter 2, we talk about the general theories and of phishing, antiphishing and related knowledge. Here we can find how the word “phishing” came out, how phishers successes by using different technology and in what ways we can detect phishing and prevent them. In the main part, chapter 3, the architecture of the phishing detecting system, the structure of program and the way to search the phishing information are included. In chapter 4, I reported about tests of the system and its success rate. We also discuss appropriate modifications of the system. We can also get the codes of the program in the appendix.

2 Phishing and Anti-Phishing

Quoting from Anti-phishing work group, [7]“phishing is a form of online identity theft that employs both social engineering and technical subterfuge to steal consumers’ personal identity data and financial account credentials. Social-engineering schemes use ‘spoofed’ emails to lead consumers to counterfeit web sites designed to trick recipients into divulging financial data such as account usernames and passwords.”

2.1 Background of Phishing and Antiphishing

Phishing has been around since 1995 but became more prominent in July 2003 when phishers began to actively target large financial institutions. Before 2003, phishing was almost unknown. It was generally used only in reference to steal AOL users’ credentials. As time went by, phishing started to be popular in 2004, almost 2 million U.S citizens had their check accounts raided by cyber-criminals. With the average reported loss per incident estimated at $1200, total losses were close to $2 billion. U.S. consumers were scammed out of roughly $3.2 billion over 2007 from phishing scams, a significant increase over last year, according to a survey, produced by Stamford Conn.-based research firm Gartner Inc. [8]

In recent years, the attack on online banking system has become more and more popular. One reason is that an increasing interest from financial institutes to offer online services. We can find one example in an online service provide company-Alibaba. Most people have never heard of Taobao - an online trade site owned by Alibaba in China. But in recent years, this company has gotten a rapid development. A survey shows that Taobao's market share increased from 9% to 40% in 2004. [9]

2.2 Phishing

As the definition suggests, phishers use social engineering and several technical tricks to achieve their aim that is spoofing the users. In this section, we will discuss the origins of “phishing” and some phishing-relative organizations.

2.2.1 Origins of the Word "Phishing"

Actually “Phishing” is made by two words, ”phreak” and “fishing”. Phreak, a word construct by phone and breaking, was coined by John Draper who is a famous hacker in history. Blue Box is an invention of his. He used Blue Box to hack the telephone system by sending a specific tone to the phone switches, and then the call would be free. Many hackers and hacker organization use ph, from phreak, as their nicknames. Talking about phishing, we know phishing is using a bait to spoof the fish, if it is beguiled by our bait, we will catch it as our prize. Here the email is the bait, the users who receive the emails are the fish. Therefore the two words can clearly explain the essence of stealing the online accounts and is now a well established term.

The first reported phishing incident is that phishers stole the American Online

(AOL) accounts from unsuspecting AOL users during 1990s. Just one year later, phishing attacks changed their target from the AOL to some financial companies and users for the purpose of stealing money. The users of online financial companies, Paypal, and EBay are the most popular targets for phishers now.

Phishing emails are actually just another form of spam. It is a subset of the category scam. The information stealing goal makes phishing special while spam is just useless information which is more or less harmless to the users. This means phishing are much more targeted and they usually target a bank or an online trading service. Social networking systems are also important targets for phishing. Experiments have historically show a success rate of over 70% for phishing attacks on social networks [10].

2.2.2 ISP

An Internet service provider (abbr. ISP, also called Internet access provider or IAP) is a business organization that provides consumers or businesses access to the Internet and related services [11]. As the companies are developing, they may provide a combination of services including Internet access, domain name registration and hosting, and web hosting.

There are thousands of ISPs all round the world, no matter big or small, they all provide Internet access service to their clients. If a phisher want to put their phishing website online, they must get this service from an ISP. So once we find a phishing site, the best way to stop the service is to inform the ISP. Further more we can find the registered information of the persons and send them to the court. Today most phishing sites are registered by the 10 largest ISPs in USA,

Rank	ISP	Market Share (%)
1	SBC	18.2
2	Comacast	13.1
3	America Online	10.2
4	Verizon	8.1
5	Road Runner	7.8
6	EarthLink	4.4
7	Cox	3.7
8	Charter	2.7
9	Qwest	2.6
10	Cablevision	2.3

Table2.1 Top 10 ISPs of USA

These ISPs have their own policies about domain registering. We need to know these policies to have better information about what kind of domains the phishers can use.

2.2.3 Anti-phishing organizations

Many organizations are concentrating on phishing monitoring and research. Below are several famous organizations in the anti-phishing field.

Anti-phishing Work Group:a group where you can report the phishing and get lots of information about phishing. They have an official website for collected phishing events and work together with many group companies [12].

CNCERT/CC is a functional organization under Internet Emergency Response Coordination Office of Ministry of Information Industry of China, who is responsible for the coordination of activities among all Computer Emergency Response Teams within China concerning incidents in national public networks. It provides computer network security services and technology support in the handling of security incidents for national public networks, important national application systems and key organizations, involving detection, prediction, response and prevention. It collects, verifies, accumulates and publishes authoritative information on the Internet security issues. It is also responsible for the exchange of information, coordination of action with International Security Organizations [13].

IRIS-CERT is Red IRIS' security service, and is aimed to the early detection of security incidents affecting Red IRIS centers, as well as the coordination of incident handling with them. Proactive measures are in constant development, involving timely warning of potential problems, technical advice, training and related services. [14]

The Messaging Anti-Abuse Working Group (MAAWG) is a global organization focusing on preserving electronic messaging from online exploits and abuse with the goal of enhancing user trust and confidence, while ensuring the deliverability of legitimate messages [15].

Many anti-virus software companies are also taking some efforts in the phishing research. If we find some suspicious incidents, they are also the places where we can report it.

2.2.4 Report of phishing

In a report on phishing coming from the Anti-phishing Group on October 2007,it is reported that the total number of unique phishing reports submitted to APWG in October 2007 was around 40 thousands, an increase of nearly 13,000 reports from the previous month. From the reports we can get useful information, such as the phishing country, new phishing features, and high frequent port for phishing.

Statistical Highlights for October 2007

Number of unique phishing reports received	31650
Number of unique phishing sites received	34266
Number of brands hijacked by phishing	120
Number of brands comprising the top 80% phishing	11
Country hosting the most phishing websites	United States
Contain some form in of target name in URL	29%
No hostname, just IP address	12%
Percent of sites not use port 80	0.82%
Average time online for site	3.1 days
Longest time online for site	31 days

Table 2.2: Statistical Highlights for October 2007

In Table 2.2, we find that United States is the most popular country for phishers. Maybe it’s because its high computer penetration and that much more Americans use online service than in most other countries. We also see that the survival time of phishing sites are all very short, normally a couple of days. Figure 2.1, we can see that many brands have been phished during 2006 to 2007, and phishers are inclined to famous brands which have a big amounts of users.

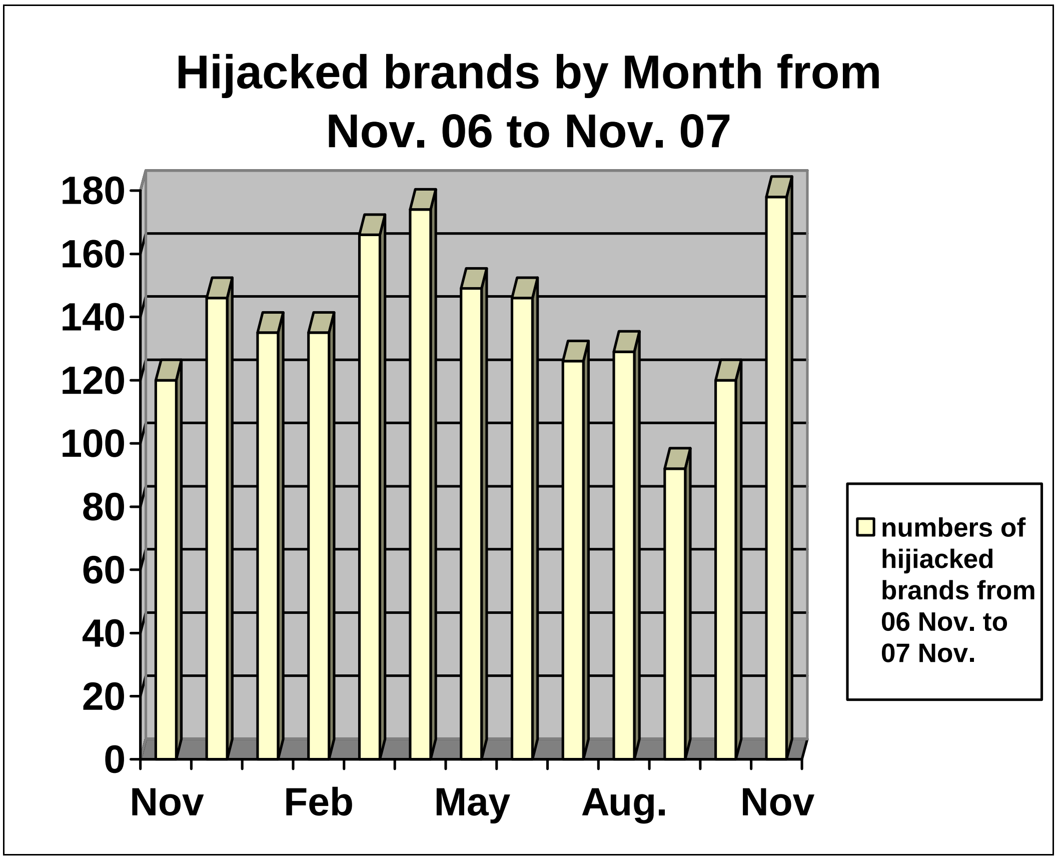


Figure 2.1: Hijacked Brands in Nov. from Antiphishing Group

2.3 Spoofing technology

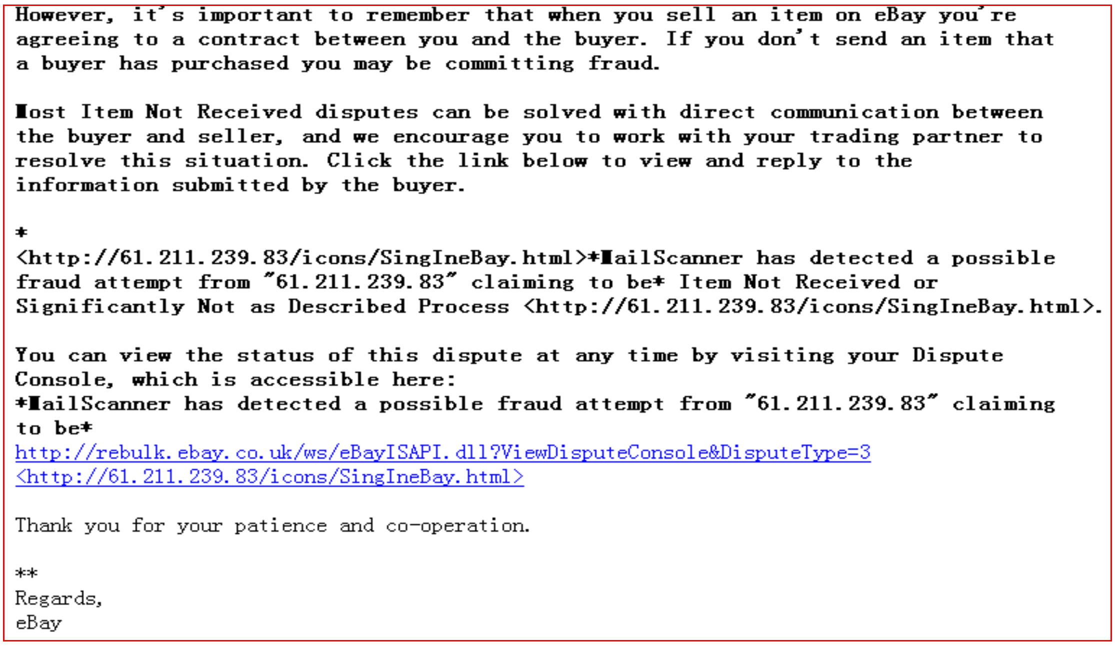
For achieving the aim of beguiling the users, phishers have their ways and technologies to success. The two main mediums for sending the phishing files are email and instant message. The reasons why phishers can succeed are the use of social engineering and the flaws of some internet protocols and tools. In the context of phishing, two spoofing technologies widely used are email and the web.

2.3.1 Email spoofing

Email is the most popular and basic way used for phishing, because it is very easy to send to a large amount of users.

People who send spam generally send millions of e-mails at a time. To maintain the high volume of e-mail generation, phishers use bulk-mailing tools. These tools generate unique e-mail headers and e-mail attributes that can be used to distinguish e-mail generated by different mailing tools. They will also set up a fake Web site to which the user is deceived to visit. The site contains images from the real Web site, or it can even be linked to the real site. On the Web site, the users are required to update their financial or personal information for some emergency reasons. After the users input their information, it will be sent to the phishers by mails.

The format of the email can be either text or HTML. Almost all scam emails are HTML based, simply put this means they have colors, pictures and other text formatting. The main advantage to the scammers of HTML emails is that it allows them to hide the real URL of a website behind one that looks genuine. Text emails are much simpler and plainer, and cannot hide URL's. Occasionally an HTML email is made to look like simple text to exploit awareness of this fact.



However, it's important to remember that when you sell an item on eBay you're agreeing to a contract between you and the buyer. If you don't send an item that a buyer has purchased you may be committing fraud.

Most Item Not Received disputes can be solved with direct communication between the buyer and seller, and we encourage you to work with your trading partner to resolve this situation. Click the link below to view and reply to the information submitted by the buyer.

*
<http://61.211.239.83/icons/SingIneBay.html>*MailScanner has detected a possible fraud attempt from "61.211.239.83" claiming to be* Item Not Received or Significantly Not as Described Process <http://61.211.239.83/icons/SingIneBay.html>.

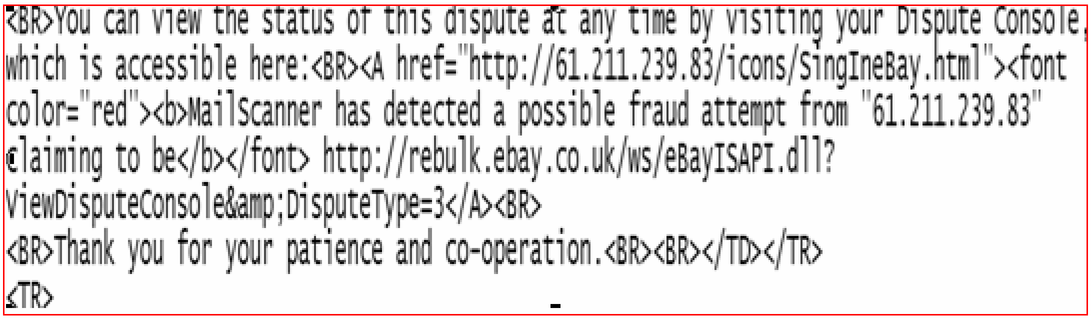
You can view the status of this dispute at any time by visiting your Dispute Console, which is accessible here:
MailScanner has detected a possible fraud attempt from "61.211.239.83" claiming to be
http://rebulk.ebay.co.uk/ws/eBayISAPI.dll?ViewDisputeConsole&DisputeType=3
<http://61.211.239.83/icons/SingIneBay.html>

Thank you for your patience and co-operation.

**
Regards,
eBay

Figure 2.2 a link which direct to another website instead.

Though the visible link in figure 2.2 is http://rebulk.ebay.co.uk, the actual link is the site: http://61.211.239.83 as seen in figure 2.2 .We can see that even for a faked site, the URL of the phishing site is also similar to a real URL. It raises the possibility of spoofing the users.



```
<BR>You can view the status of this dispute at any time by visiting your Dispute Console,
which is accessible here:<BR><A href="http://61.211.239.83/icons/SingIneBay.html"><font
color="red"><b>MailScanner has detected a possible fraud attempt from "61.211.239.83"
claiming to be</b></font> http://rebulk.ebay.co.uk/ws/eBayISAPI.dll?
ViewDisputeConsole&amp;DisputeType=3</A><BR>
<BR>Thank you for your patience and co-operation.<BR><BR></TD></TR>
<TR>
```

Figure 2.3 The HTML Codes corresponding to the email in Figure 2.2

A new technique is used by phishers, which is java script which is a hidden code plugged in the web page. Once the user clicks it, he could be misleading to the way to be spoofed. It is widely used and can not be detected easily.

2.3.2 Web site spoofing

A faked web site is almost identical with the real one. They have the same frame structure, and most of the frame could actually be copied from the legit web site.

However a tiny portion of one frame could be changed, the users can not recognize it even if they have been surfing the related legitimate web site for many times.

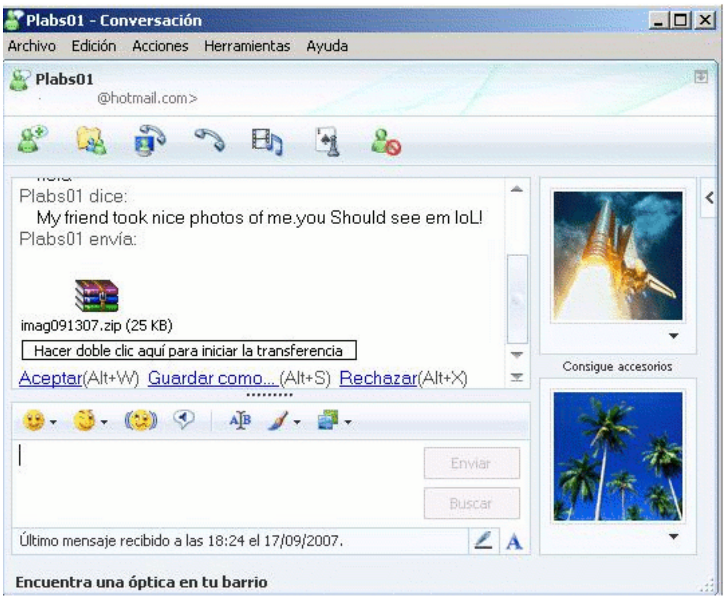
There is another attack which does not need to create a fake Web site, the phishers simply involve a redirect script that collect the data and forward the victim back to the real web site. As the protection to the finance web sites are quite strong, this skill is method is not used very common today. But a incident happened before in China, a phisher useed a Java Script on the ICCBC which is one of the biggest banks in China, causing many costumers' account to be compromised.

2.3.3 Instant messaging

As the usage in instant message software such as Window Message, Skype, and Yahoo Message have increased, the number of phishing incidents have also increased in the instant message word.

For this new form of phishing, Bakosto wrote ”It is important to understand that most instant messaging systems use only weak authentication schemes. Instant messaging is not a tool for exchanging confidential information. Only few instant messaging systems allow for encryption and sophisticated authentication. If you need instant messaging to communicate confidential information, use a system that allows you to control the server and provides for encryption and reasonable authentication. Jabber is an example of a free package [with these capabilities]." [16]

As it is hard to detect an instant message phishing attack, currently there is not an effective way to deal with it. What we can do is tell persons to be more careful about the information relative to their personal details. As one of the most famous instant messaging chatting software, MSN plays an important role for spreading phishing files. The users often receive some files sent by their buddies who they trust in, but they are actually dangerous virus files. Once the users download it, they may be monitored by phishers. We can see an example of a phishing file spreading on MSN in Figure2.4.



Plabs01 - Conversación

Plabs01
@hotmail.com>

Plabs01 dice:
My friend took nice photos of me.you Should see em loL!
Plabs01 envía:

imag091307.zip (25 KB)
Hacer doble clic aquí para iniciar la transferencia
Aceptar(Alt+W) Guardar como... (Alt+S) Rechazar(Alt+X)

Último mensaje recibido a las 18:24 el 17/09/2007.

Encuentra una óptica en tu barrio

Figure 2.4: An example of instant message phishing attack in MSN Live Messenger.

2.3.4 Phone phishing

Phone phishing was historically used by phishers even though at that time they didn’t know it was phishing. However it is still a medium used today. Phishers recreate a legitimate sounding copy of a bank or some finance institutes phone answerer. They will publish the phone number, usually by phishing mail, to the public. Once the victim is calling the number, the copied voice will be used to spoof the user to release information to the phisher.

2.3.5 Pharming

Normally we identify the web site on the Internet by their visible domain names, such as “google.com”. In the user’s computer, there is a stored list of domain names, called hosts files, which link an IP address to a domain. Once the user want to open a webpage, the computer will check the host lists first. If the domain of the webpage exists in the host list, the computer will direct the user to the IP address stored in the list. Else if the webpage does not exist in the host list, the computer will check the DNS server on the Internet to find the corresponding IP address of the webpage. For example, google.com is directed to 66.249.93.147. Though it brings us much convenience, a big flaw could be used by pharmers. The pharmers can access a user’s computer by means of a virus or Trojan horse and change the hosts file inside. Once the user wants to open one special web page, the computer will check host file to get the IP address. The pharmer can now misdirect the client’s web request to another faked web server. Another similar way is changing the information at the DNS sever on the Internet. Both attacks could be launch through malicious programs, such as viruses or trojan horses.

2.3.6 SMTP and HTML

Simple Mail Transfer Protocol (SMTP) is used for mail transfer. It was designed in 1982 and at the time it was intended to be used between limited and trusted users. So there were not many concerns about the security problem However by the time these security issues were exposed it was too late as the protocol had gained popularity and it still continues to be one of the most widely protocols [17].

Hypertext transfer protocol (HTTP) is used for the transfer of multi media documents on the internet. The HTTP is not inherently insecurity as SMTP, but it suffers from a lack of standardization and the heterogeneous usage of web browsers such as FireFox, Internet Explorer and Safari. Many attackers can use flaws in the browsers to spoof the users such as www.b1gbank.com which is a phishing site but most persons will see it as www.bigbank.com just because the number “1” is similar with the letter ”i”.

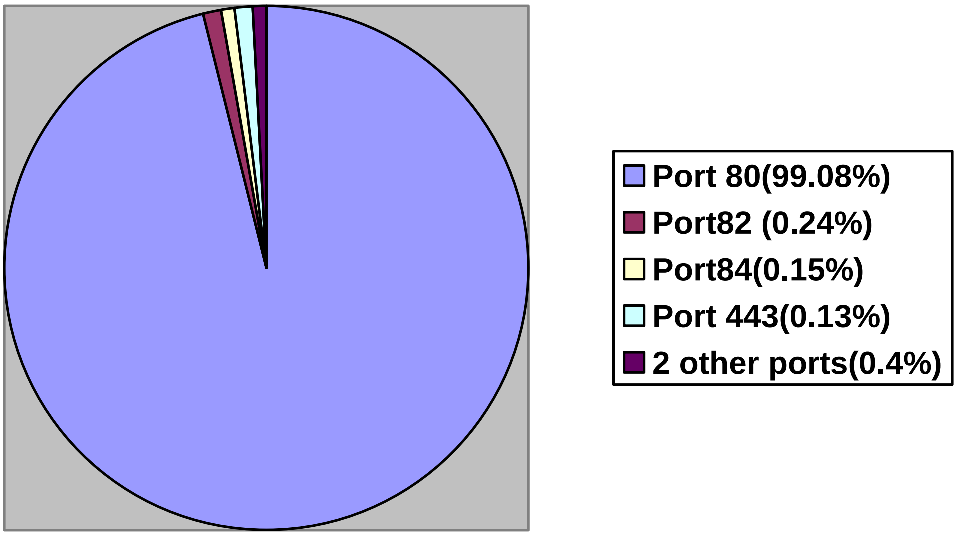
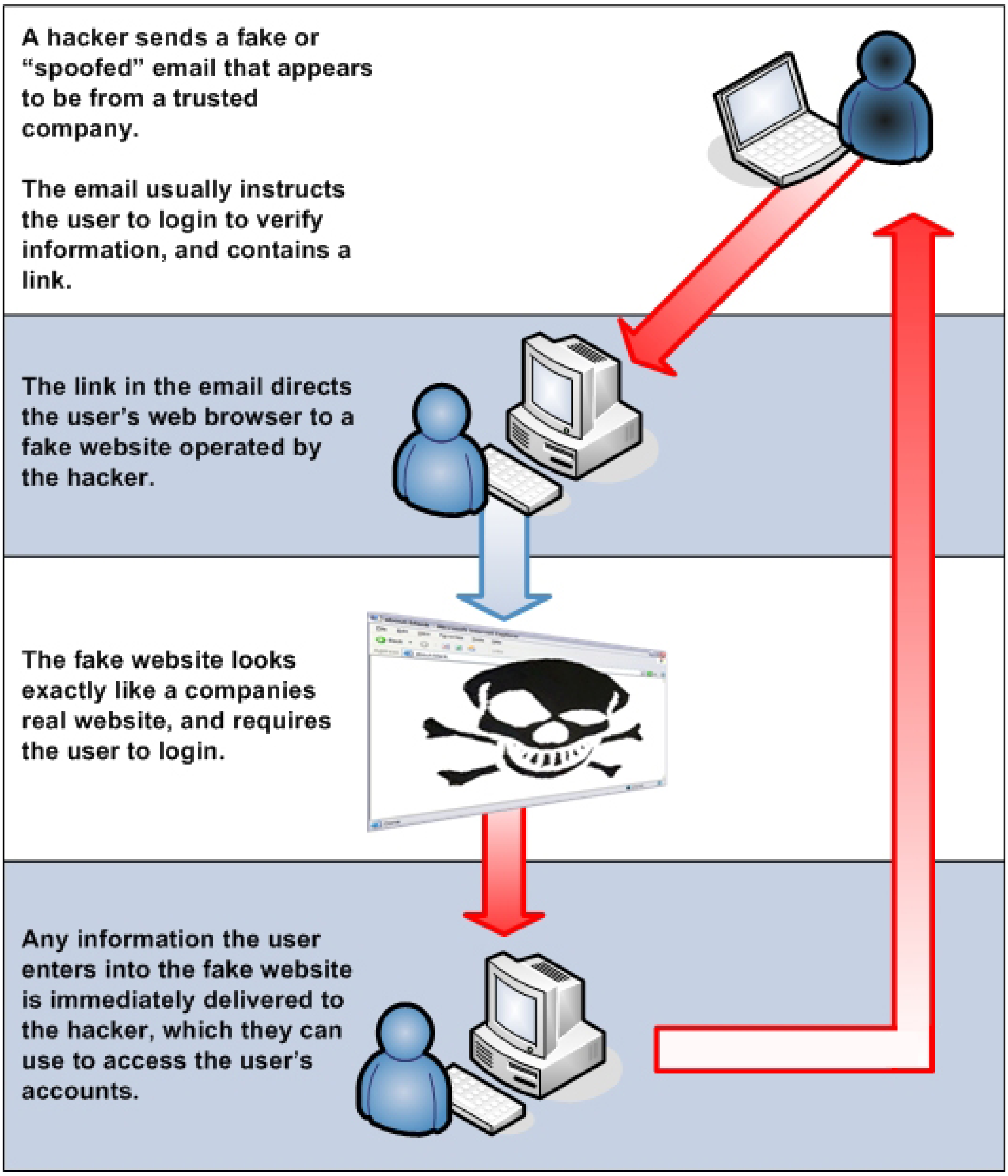


Figure 2.5: Top Used Ports Hosting Phishing Data Collection Servers [16]

We can see from figure 2.5 that HTTP port 80 being the most popular port used at 99.08% of all phishing sites reported. Port 80 is used for browsing webpages, and most information transferring on the internet is using this port.

2.4 How phishing works

Several mediums are used by phishers. They are instance message, phone phishing, and pharming. However the most effective and universal way is sending email. So here we will describe how phishing mails works.



A hacker sends a fake or "spoofed" email that appears to be from a trusted company.
The email usually instructs the user to login to verify information, and contains a link.
The link in the email directs the user's web browser to a fake website operated by the hacker.
The fake website looks exactly like a companies real website, and requires the user to login.
Any information the user enters into the fake website is immediately delivered to the hacker, which they can use to access the user's accounts.

Figure 2.6: A model for the processing of phishing. [18]

In general, phishing attacks are performed with the following four steps:

1) Phishers first create a faked web site in a web server. This web site would look similar or even the same as the legitimate web site. Then they will apply for a domain namewhich would have a very short survival time on some ISPs.

2) Using some tools, GroupMail is one example, they now send a lot of spoofed e-mails to target users in the name of those legitimate companies and organizations, trying to convince the potential victims to visit their Web sites.

3) Users receive the e-mails. When they open an e-mails there are some hyperlinks waiting for them to click. If they do click on the spoofed hyperlink, the link will direct them to a web page that is asking the users to input the required information.

4) Once the users input their information, the phishers will get them by email or some other means. Then phishers can do anything they want with this information, including drawing out the money from the users' account.

In our anti-phishing system, we use several technologies to collect phishing e-mails and analyzing them, and then we will take some actions to prevent the users from being beguiled.

2.5 Anti phishing

Normally we can find some anti-phishing plug-in systems in browsers or operating system. They can just alarm the users that it is probably a phishing attack, but the user could click the phishing link all the same. With our intended system, we can quickly shut down the site or distribute information used to update anti-phishing software. It is more secure than before.

According to the general idea about how phishing works in section 2.4, it's time to find a way to prevent it. There are many anti phishing products working on the Internet and the users' computers now. It is often integrated in web browsers and email clients. Internet Explorer (IE) and Firefox are the most common web browsers. Microsoft have a phishing filter in IE7; an access to a phishing sites will be blocked. In Firefox 2 there is a list of known phishing sites. The list is stored both in the software and on the Internet. Again the user will be blocked if he tried to access one of these sites.

Phishing/pharming filter, built-in anti phishing plug in web browsers, and augmenting passwords login are very popular techniques used in anti-phishing products. Phishing filter will detect the mail to see if it is a phishing attack or not. If it is, the e-mail will be blocked. Augmenting passwords login is used in some banks, such as ICCBC. The client has a unique card that has a series of numbers using when he/she want to login to the account. The server of the bank will send a random secure number to the users and the user will check the identification number by using the secure number and send an acknowledged number back.

Another effective way is education. Phishing attacks are often quite simple and its features are quite obvious. The best way to avoid being phished is to know what phishing is, how it looks like. Example are checking your toolbar to see if the web page is the link you want to open, pay more attention about the mails that requires your personal information and so on.

3. E-mail detecting system

In this part, I would like to introduce the whole system first. After giving you a general architecture of the anti phishing system, I will concentrate on my part that is e-mail detecting system. This specific part of the system will use an e-mail database as an input. After running the e-mail detecting system, it will work out the important and useful information about the phishing e-mails.

3.1 Architecture of anti-phishing system

In the whole system, we have an e-mail accounts creation system, mail detecting system, phishing mail analyze system, and warning system. In figure 3.1, we can see all the involved parties in the phishing and anti-phishing activity. They are phishers, anti-phishing systems, ISP, faked website, legitimate web site and target users.

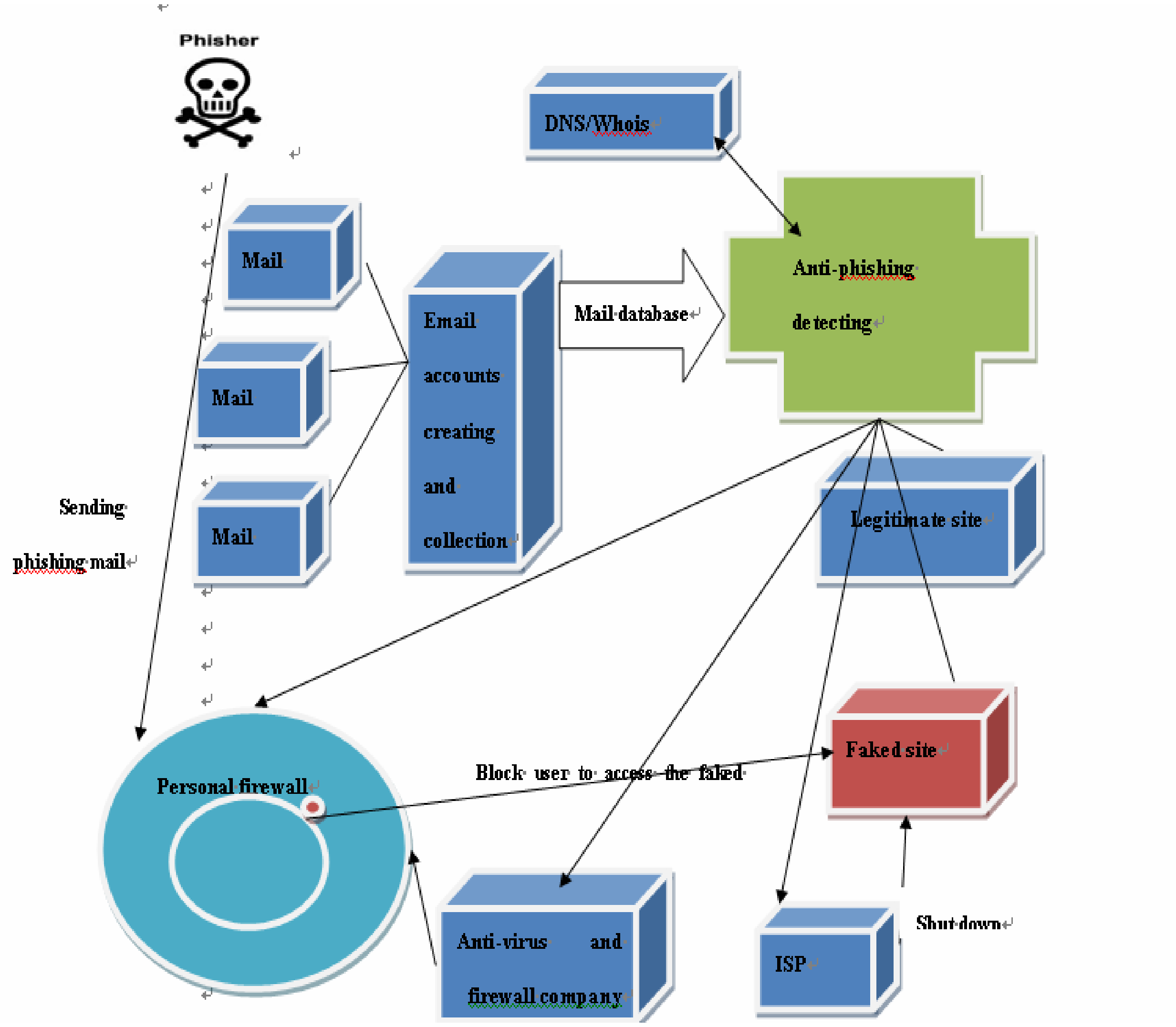


Figure 3.1: Architecture of anti-phishing system

Through the figure 3.1 we can set an overview the process of the system. Firstly, the accounts-creating system creates thousands of email accounts that will be harvested. E-mails will include normal ones, spam and phishing ones. We will put these e-mails into a database. The next part is the main part of the system, anti-phishing detecting. We will use features of the phishing e-mails to diagnose whether they are phishing or not. If there are phishing e-mail, we will process further to check the details of the mails including the domain information, the technique that phishers used, and the attacked websites. Then we will alarm the attacked legitimate websites to do some protecting options about the attack and inform the Internet Service Provider of the phishing sites to block the phishing site. We will also inform the anti-virus and firewall software companies to update their products to protect from this new phishing attack. There are now several ways to stop the phishing attack. First the site will be shut down by ISP. If that fails the phishing e-mail will be filtered of and lastly firewalls will block attempts to access the site.

3.2 Process of the system

1) Create a system to create email accounts automatically, and harvest the e-mails

At first, we build a system which can create new email accounts in many common mailbox providers such as Yahoo, Hotmail, etc. Then we will publish the addresses on Internet so that the phishers can find them. We keep these accounts activite and collected the e-mails which will be used in the analyzing part of the anti-phishing system. As this part is not very related to anti-phishing theory, it was not developed as part of this thesis.

2) Classify spam or phishing

After harvesting enough e-mails, it’s time for us to create a database for all the e-mails. The e-mails are stored as text files. Then we begin to analyze the mails to detect whether they probably phishing mails or not. Below are the features could be used to detect the phishing e-mail.

＊ Blacklist of the phishing websites. We collect and update the database of the phishing information on the phishing website like phishing IP address in a blacklist. We will simply search the database to see if they are inside the blacklist when we are scanning the e-mails. There are some blacklists on the internet we may collaborate with, one is http://www.spamcop.net/

＊ White list. We can also collect legitimate websites in a white list. If the e-mail’s list to a website exists in the white list, we can conclude that this e-mail is not a phishing e-mail.

＊ Age of domain. Normally the phishing mails will lead the users to a spoofed website. Here the users will be required to fill in their account information. The faked site cannot be active for a long time because the Internet Service Providers will learn about them and shut them down. In addition, many phishing sites have domains that are registered only a few days before phishing emails are sent out.

We measure age of this site, through table 2.2 we can tell the average online time of the phishing site is 3 days.

＊ The special symbols in URL. In some phishing sites, they use a few special symbols in its URL to spoof the users. For example: www.legitimatesite @ phishing site.com or IP-based URL: HTTP:// 192.168.1.1/paypal.

＊ Number of dots in URL. Phishing sites often use many sub-urls. The beginning part is similar to the legit site, so the clients may believe it is the legit site indeed. I found that phishing pages tend to use many dots in their URLs but legitimate sites usually do not.

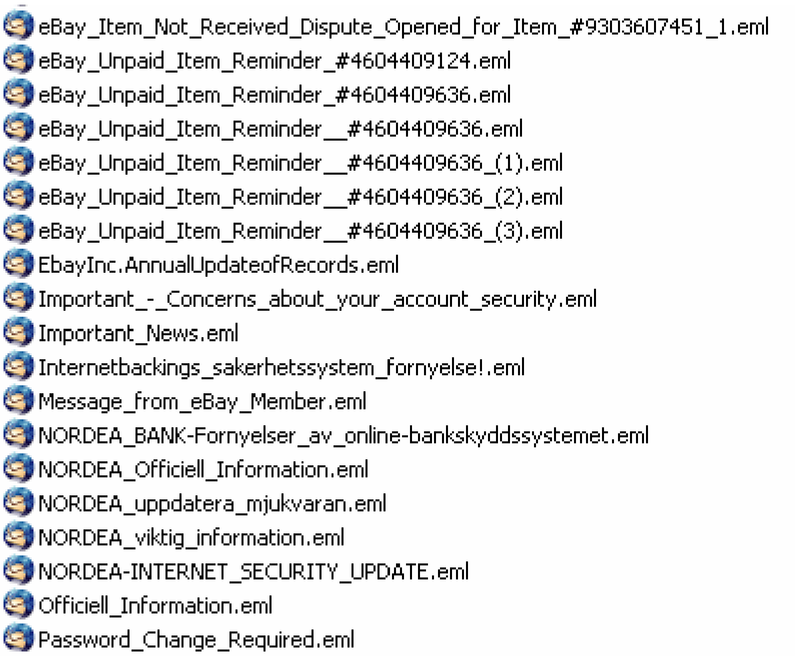
＊ Information in the content of the mail. For most phishing mails, they have the same purpose to acquire sensitive information. So they all have the input or hyperlink to lead the user to send out their information. Also the mail usually is HTML format; we can e.g. scan <Input>tags for "Credit card" or "password". If it contains this information, it probably is a phishing mail. As an example the hyperlink, <a href="badsite.com">paypal.com</a> will show paypal.com in the e-mail. But it will direct to the phishing site, badsite.com.

In previous work done by Nicklas Karlsson in Vaxjo University 2008, "System för uppt äckt av phishing", he classify the e-mails. The classifications that were made showed that it was possible to find up to100% pf the phishing emails with both Native bayes and with Support Vector Machine.

3) Output an information list of the phishing mails

When we get to the previous part of the system, most e-mails getting this far will be phishing e-mails. We make some further information from the emails. We have the database which contains information about all the phishing mails. Then we need to collect the information in the mails about domain name, the registered time, the expired time, the registered person or company. I use a program to read the text of the phishing mail and extract the link. Then I access the WHOIS.org, searching the information of the domain. Finally I format the information and output it for further processing in the system.

In figure 3.2 are a number of detected phishing e-mails that I will use as input in my e-mail detecting system.

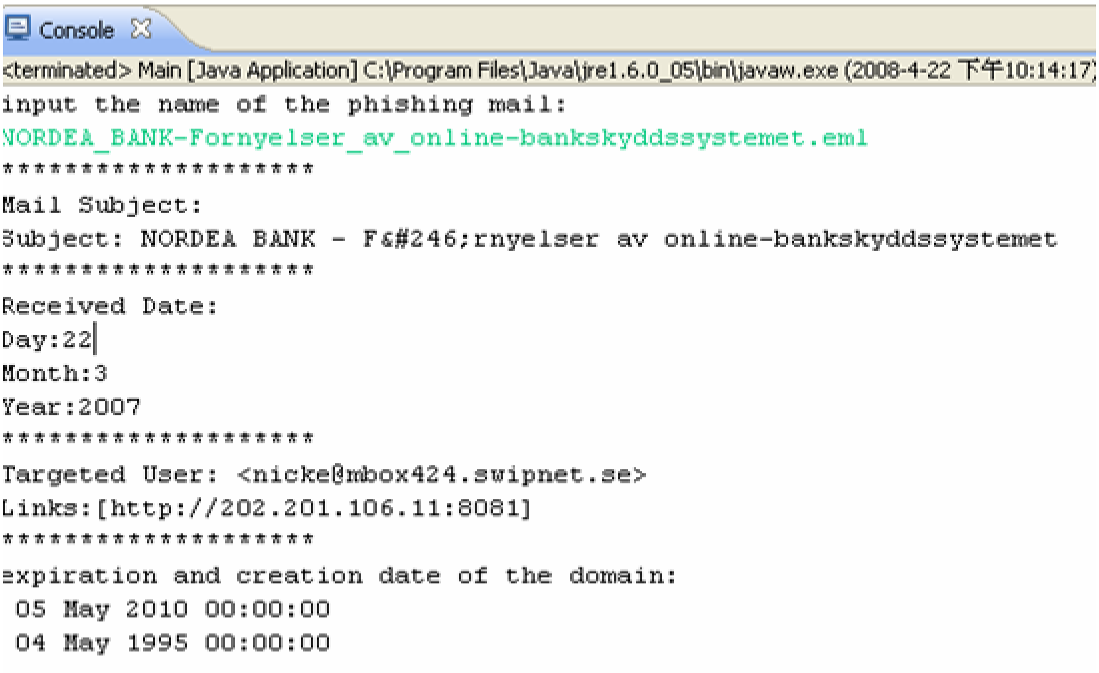


```
eBay_Item_Not_Received_Dispute_Opened_for_Item_#9303607451_1.eml
eBay_Unpaid_Item_Reminder_#4604409124.eml
eBay_Unpaid_Item_Reminder_#4604409636.eml
eBay_Unpaid_Item_Reminder__#4604409636.eml
eBay_Unpaid_Item_Reminder__#4604409636_(1).eml
eBay_Unpaid_Item_Reminder__#4604409636_(2).eml
eBay_Unpaid_Item_Reminder__#4604409636_(3).eml
EbayInc.AnnualUpdateofRecords.eml
Important_-_Concerns_about_your_account_security.eml
Important_News.eml
Internetbackings_sakerhetssystem_fornyelse!.eml
Message_from_eBay_Member.eml
NORDEA_BANK-Fornyelser_av_online-bankskyddssystemet.eml
NORDEA_Officiell_Information.eml
NORDEA_uppdatera_mjukvaran.eml
NORDEA_viktig_information.eml
NORDEA-INTERNET_SECURITY_UPDATE.eml
Officiell_Information.eml
Password_Change_Required.eml
```

Figure 3.2 Phishing mails

After inputting one file, the system will check the content of the e-mail and search for a phishing link. It will then search for information of the domain on the Internet. Rearrange the HTML information and output the required list.

In figure 3.3 we can see the entire process of the system. First I input the name of phishing e-mail, and then the system will read the e-mail to extract the main features, and then connect to Internet to find out the domain information.



```
<terminated> Main [Java Application] C:\Program Files\Java\jre1.6.0_05\bin\javaw.exe (2008-4-22 下午10:14:17)
input the name of the phishing mail:
NORDEA_BANK-Fornyelser_av_online-bankskyddssystemet.eml
********************
Mail Subject:
Subject: NORDEA BANK - F&#246;rnyelser av online-bankskyddssystemet
********************
Received Date:
Day:22
Month:3
Year:2007
********************
Targeted User: <nicke@mbox424.swipnet.se>
Links:[http://202.201.106.11:8081]
********************
expiration and creation date of the domain:
 05 May 2010 00:00:00
 04 May 1995 00:00:00
```

Figure 3.3 Input and output screen

4) Warning system which is used to inform the users , the Internet Service Provider and other companies

In this part of the system, we are planning to implement a function module which can alarm the users that there is a phishing site in the e-mail, inform the companies responsible for the domain registration that there is a probably phishing site exists in its server, tell the attacked legitimate website to do some preventive actions, and request security companies (e.g. anti-virus and firewall companies) to update their products.

In figure 3.4, we can see the companies used by phishers for registering the domain name. Phishers use a ISP to surf Internet and control slave PC to send phishing e-mail. Web hotel is used to set fake web site for anonymous reason, in this part a ISP is also used. Once I find a phishing e-mail, I will inform the responsible ISP to shut down the phishing site.

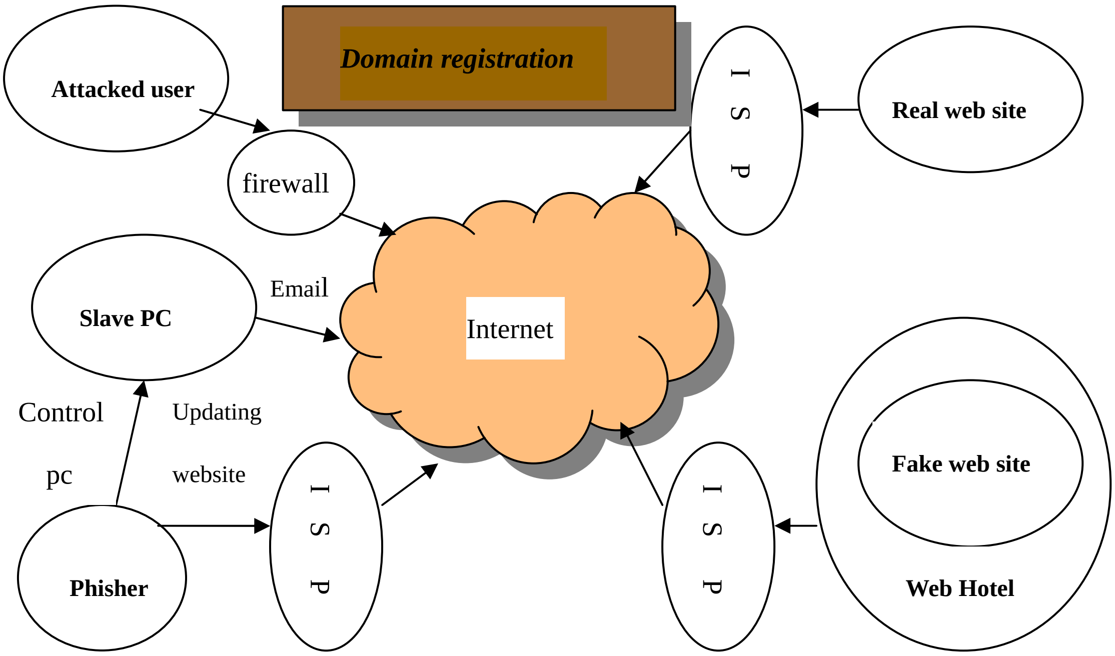


Figure 3.4 Responsible parties of domain registration

3.3 The architecture of the e-mail detection system

The e-mail detection system is the main part I worked with. Here we have plausible phishing e-mails as an input. We analyze them in detail, and then extract useful information from these phishing e-mails, such as the domain name, where they are registered, the expire time, which legitimate web sites they are phishing and so on.

From Figure 3.4 we can see the three main parts of our whole system.

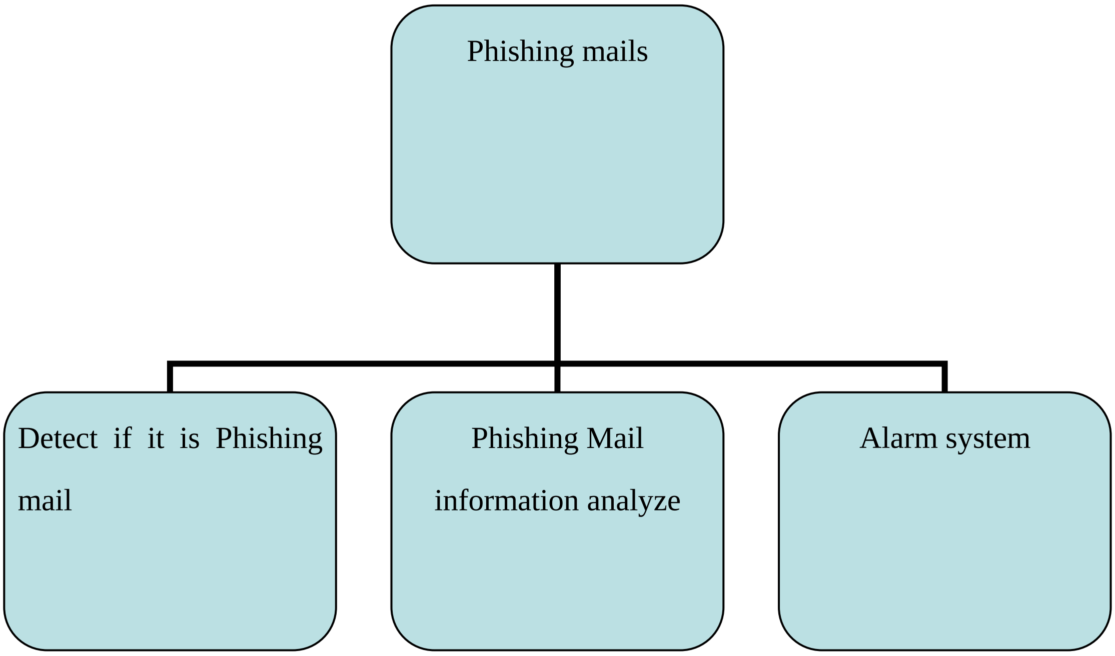

Phishing mails
Detect if it is Phishing mail
Phishing Mail information analyze
Alarm system

Figure 3.5 Three main parts of our system

3.4 Implementation

According to my programming experience, I chose to use Java language to implement the mail analyze part. I have three important function parts in my system. One is reading files, one is extracting the details of the e-mail, the last one is accessing Internet to extract the domain information through a relative webpage.

The barriers exist during the implement process are mainly two things. One is that there are several formats of e-mail text, so that I need to find out all the possibility to make the system run well. The other is how to find out the information of the phishing domain, in this part I chose to use a simple way that is a tailor made URL relative to the domain and WHOIS server.

3.5 Analyzing phishing e-mails

The practical part of the system was developed in Java Version 6 on Microsoft XP operating system. Both development and testing were performed on the same machine.

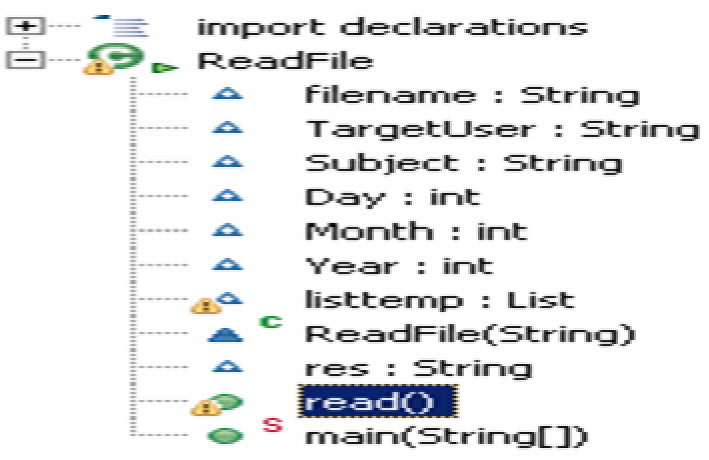
3.5.1 E-mail Analyze

My program consists of 4 java classes: ReadFile.java, URLTest.java, MothToInt.java, main.java

ReadFile.java. This class is used to read the e-mails database and extract the general information such as the sending data, the domain name of the faked site from the e-mail

The main part of the class is ReadFile (String) function. It requires a String attribute that is the name of the mail. Then it will get out the basic information of the text mail.

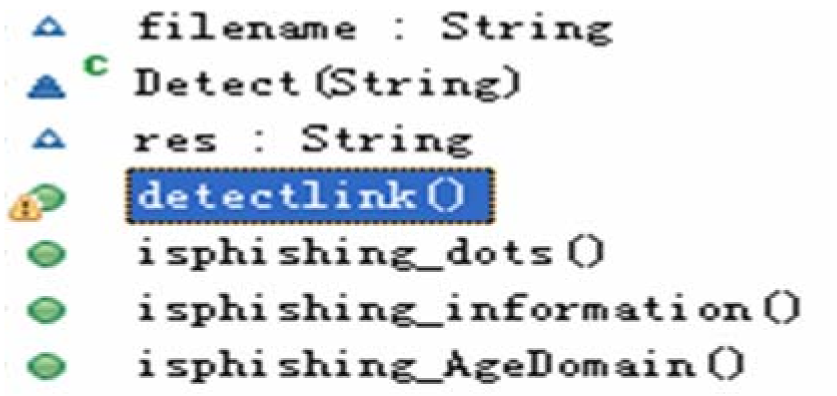
In figure 3.6, we can see the parameters and function of the class.



```
import declarations
ReadFile
    filename : String
    TargetUser : String
    Subject : String
    Day : int
    Month : int
    Year : int
    listtemp : List
  c ReadFile(String)
    res : String
    read()
  S main(String[])
```

Figure 3.6: The data structure and method in the class

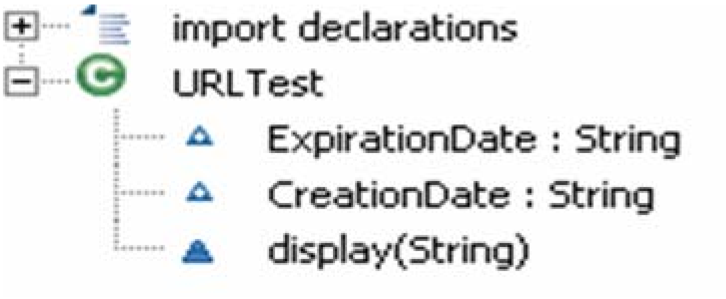
Detect.javaDetect.java: this class is scanning the e-mail to figure out whether it is a phishing mail or not. I use several different features to make this analyzing. They are the numbers of dots in URL, if it requires the user to input some financial information, and the age of the domain. If any function return true, it is phishing mail, otherwise it is considered to be a normal e-mail.



```
  filename : String
c Detect(String)
  res : String
  detectlink()
  isphishing_dots()
  isphishing_information()
  isphishing_AgeDomain()
```

Figure 3.7: The data structure and method in the class Detect.java

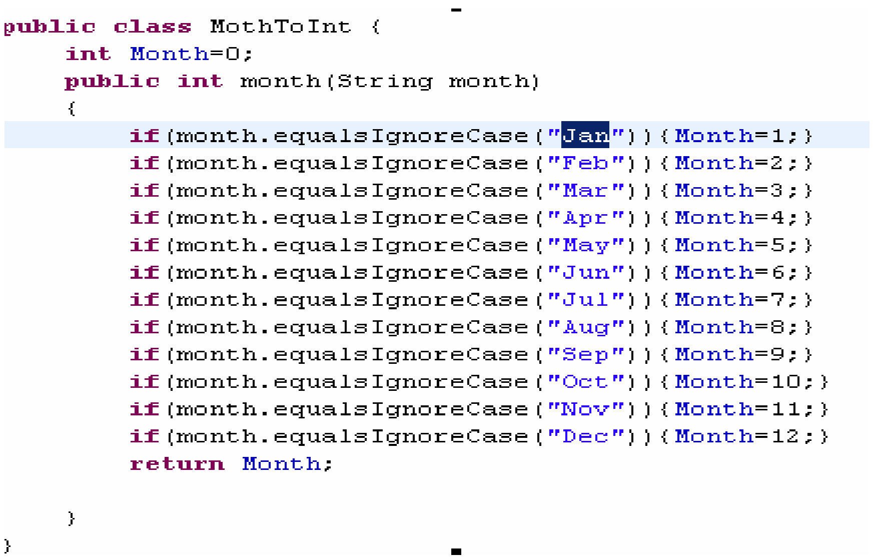
URLTest.java is a function to search for domain information through the WHOIS.org on the domain name that is output from ReadFile.java.



import declarations
URLTest
ExpirationDate : String
CreationDate : String
display(String)

Figure 3.8: The data structure and method in the class URLTest.java

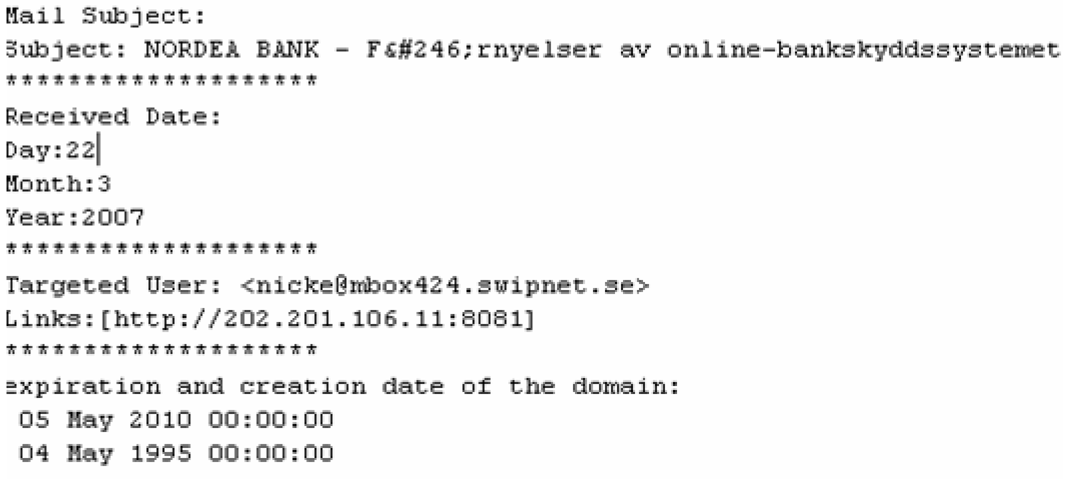
MothToInt.java: this class will translate the String month into Int month. For example, if the moth is”Jan”, the int month will be 1.



```
public class MothToInt {
    int Month=0;
    public int month(String month)
    {
        if(month.equalsIgnoreCase("Jan")){Month=1;}
        if(month.equalsIgnoreCase("Feb")){Month=2;}
        if(month.equalsIgnoreCase("Mar")){Month=3;}
        if(month.equalsIgnoreCase("Apr")){Month=4;}
        if(month.equalsIgnoreCase("May")){Month=5;}
        if(month.equalsIgnoreCase("Jun")){Month=6;}
        if(month.equalsIgnoreCase("Jul")){Month=7;}
        if(month.equalsIgnoreCase("Aug")){Month=8;}
        if(month.equalsIgnoreCase("Sep")){Month=9;}
        if(month.equalsIgnoreCase("Oct")){Month=10;}
        if(month.equalsIgnoreCase("Nov")){Month=11;}
        if(month.equalsIgnoreCase("Dec")){Month=12;}
        return Month;

    }
}
```

Figure 3.9: The data structure and method in the class MothToInt.java

Main.java: this class will display the result of the project. The result is showed in figure 3.10. It shows the details of one phishing mail that contains “Mail Subject”,” Received Date”, “Targeted user”, “Links”, and “expiration and creation date of the domain”.



```
Mail Subject:
Subject: NORDEA BANK - F&#246;rnyelser av online-bankskyddssystemet
********************
Received Date:
Day:22
Month:3
Year:2007
********************
Targeted User: <nicke@mbox424.swipnet.se>
Links:[http://202.201.106.11:8081]
********************
expiration and creation date of the domain:
 05 May 2010 00:00:00
 04 May 1995 00:00:00
```

Figure 3.10 Final output result of the e-mail detecting system

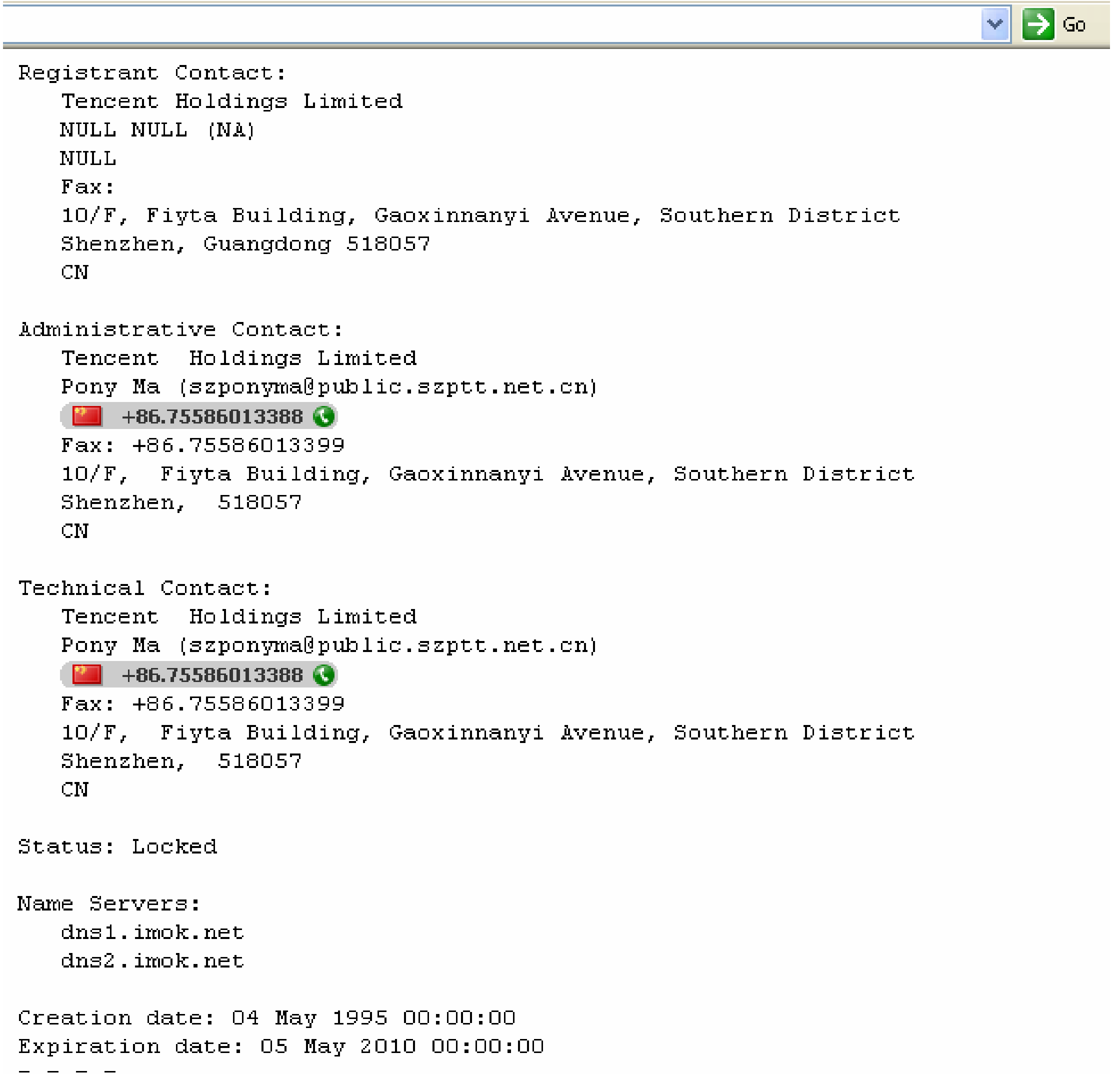
3.6 Domain information

It is very important to get the information of the domain. Here we use WHOIS which is a Internet-based service for domain information.

WHOIS service is an online “client/server” model. It will monitor the port 43. The WHOIS server will create a connection with the client when a user searches a domain name, it will then receive the request of users’ and look for the related information. If a record exists, it will send back the information to the user. Finally the connection with the user is closed.

In the system, we should search some domain information automatically. There are several ways to do it. One is as a Web Service with the limitation of the time for this project we did not chose this option. What we need is the returned information of the domain name. To attain this goal we send a tailor made URL to the server. For example, if we want to get the information on www.qq.com, we just need to open the link http://www.whois.org/whois_new.cgi?d=**qq**&tld=**com**, and then we will get the information of www.qq.com.

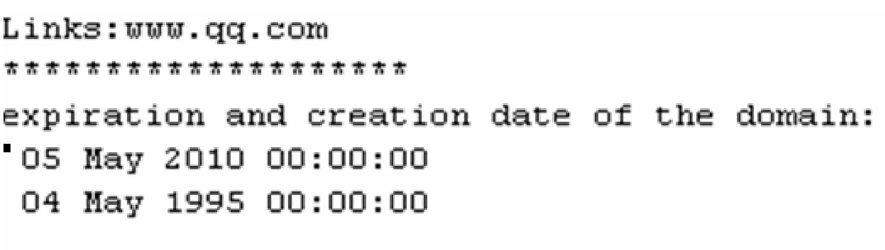
We can e.g. see that the page of the details of qq.com as a HTML document showed in figure 3.11.



```
Registrant Contact:
   Tencent Holdings Limited
   NULL NULL (NA)
   NULL
   Fax:
   10/F, Fiyta Building, Gaoxinnanyi Avenue, Southern District
   Shenzhen, Guangdong 518057
   CN

Administrative Contact:
   Tencent  Holdings Limited
   Pony Ma (szponyma@public.szptt.net.cn)
   +86.75586013388
   Fax: +86.75586013399
   10/F,  Fiyta Building, Gaoxinnanyi Avenue, Southern District
   Shenzhen,  518057
   CN

Technical Contact:
   Tencent  Holdings Limited
   Pony Ma (szponyma@public.szptt.net.cn)
   +86.75586013388
   Fax: +86.75586013399
   10/F,  Fiyta Building, Gaoxinnanyi Avenue, Southern District
   Shenzhen,  518057
   CN

Status: Locked

Name Servers:
   dns1.imok.net
   dns2.imok.net

Creation date: 04 May 1995 00:00:00
Expiration date: 05 May 2010 00:00:00
- - - -
```

Figure 3.11 Web page of domain details of qq.com following a request to http://www.whois.org/whois_new.cgi?d=qq&tld=com

Relevant information will be extracted from this web page and fed into our system. The creation and expiration time of the domain will then be output. This information is showed in figure 3.12



```
Links:www.qq.com
********************
expiration and creation date of the domain:
 05 May 2010 00:00:00
 04 May 1995 00:00:00
```

Figure 3.12 Extracted domain information of qq.com

3.7 Test of the system

After running the system on 100 phishing mails, it gave the following result:

Out of 100 phishing mails, this system could work out the complete correct information including the received date, subject, received person, and expiration and creation date of the domain for 47 of them. For some of the e-mails part of information was not correctly extracted due to different formats in the e-mails. Also in some e-mails there is more than one hyperlink, but only one of them is phishing link, the others are legit ones. I can not currently figure out which one is the actually phishing link. However, if I have a white list which includes legitimate domain names, the problem could be solved in future versions of the system.

4 Conclusions and Future work

In this chapter, I will make a conclusion for what I done in the system, what part could be improved of the system, and also several possible future work according to the new phishing trend.

4.1 Conclusion

My e-mail detecting system is running well. As the goal I mentioned in chapter1.2, I want to make a system that could extract the most valuable information of phishing e-mails. In the system, it reads the phishing e-mails and extracts the detail list that contains “Mail Subject”,” Received Date”, “Targeted user”, “Links”, and “expiration and creation date of the domain”. The list will be used in the alarm system in the future.

For achieving this aim, I have three functions in my system. One is reading phishing text, I use readline() function of Java to read the whole text, and make several marks when the scanning is meeting the keywords. In this way, this part can find out the “Mail Subject”,” Received Date”, “Targeted user”, “Links”.

The second one is dealing the special possibilities of the text and transfer the special text to the ones I need. For example, the months present with alphabet, I need to change it to a number for the time caculating, so that I use twelve options choosing to transfer it. The codes are in Figure 3.9.

After we found out a potential phishing site, we will go to the internet to search for the details of the domain. There comes the third function. WHOIS.ORG is a good place to search for domain information. Once we have received the information from WHOIS, we will select and rearrange it. List of most important information, “expiration and creation date of the domain”, will be output. The list will be used in the alarm system in the future.

At last, I tested the system. It can correctly extract 47 e-mails out of 100 e-mails. That is because it is hard to distinguish the phishing link and legit link in the emails. As I mentioned in the beginning of the report, I would like to find the characteristics of phishing e-mail and then extract the important information of the phishing domain. In the report, the four important features are given and the result of the system in figure 3.10 gives us the important information of phishing domain.

4.2 Future work

More and more phishing mails are using dynamic script language to make the faked site more close to the real site, sometimes they look excactly the same. It can even change the status bar of the browser. So if we find an e-mail including some script language like Java Script or PHP, we should pay more attention to this mail.

I can not solve the instant message phishing. Since the attacks occur in real time, we cannot track them easily. Maybe in the future, I can build some plug-in to monitor the information sent by instant message software.

And the images in the mails are a big feature that could be used in mail classification, but due to the limited time, we cannot find a good way to read information of the images. It is open for future implements.

Because of the limited time, we can not finish the warning part of the system. It will be very interesting since it is related to the communication of the internet and it also needs some security authentication.

Reference

[1] http://en.wikipedia.org/wiki/Phishing ,“Phishing”, 15th December 2007
[2] <<Phishing attacks damage consumer confidence: survey >>27th November 2007 By Steve Evans.
[3] http://www.gartner.com/it/page.jsp?id=565125 ”Gartner Survey Shows Phishing Attacks Escalated in 2007” 17th December 2007.
[4] http://www.pewinternet.org/PPF/r/155/report_display.asp “phsihing survey by Pew Internet” 1th January 2008.
[5] <<The Credibility of Enterprise’s Website and Its Evaluation in the Customer’s Perspective>>
[6] http://www.infosec.gov.hk/english/general/protect/ICBC_20040906.htm “ICBC (ASIA) - Verification of customer's username and password (with sample)” 4th January 2008.
[7] http://www.antiphishing.org/index.html
[8] http://en.wikipedia.org/wiki/Phishing
[9] http://acrossthepacific.rdvp.org/2005/11/taobao-vs-ebay-china.html
[10] http://en.wikipedia.org/wiki/Phishing
[11] http://en.wikipedia.org/wiki/ISP “ISP” 7th December 2007
[12] http://www.antiphishing.org
[13] http://www.cert.org.cn/english_web/index.htm 18 Jan 2008
[14] http://www.rediris.es/cert/index.en.html 18 Jan 2008
[15] http://www.maawg.org/home
[16] http://www.crn.com/security/23904957 Instant Messages Carry Latest Phishing Scams By Dan Neel, CMP Channel
[17]<<Phishing Exposed>> by Lance James
[18] http://edtechvalley.blogspot.com/2007_07_01_archive.html 2th January 2008

[19] http://security.tekrati.com/research/9780/

Appendix

Main Codes

Read() function of ReadFile.java:

```
read() throws IOException{
                FileReader read = new FileReader(filename);
                BufferedReader br = new BufferedReader(read);
                String temp = null;


                while((temp=br.readLine())!=null)
                {
                    {
                        //dectect the target user
                    int usermark=temp.indexOf("To:");
                        if(usermark!=-1)
                        {

                            TargetUser=temp.substring(usermark+3);
                        }
                        //detect the subjects of the mail
                    int subjectmark= temp.indexOf("Subject:");
                        if(subjectmark!=-1)
                        {
                            Subject=temp.substring(subjectmark+8);
                        }
                    int datemark=temp.indexOf("Date:");
                        if(datemark!=-1)
                        {
                            int daystart=temp.indexOf(",", datemark);
                            String day=temp.substring(daystart+2, daystart+4);
                            String month=temp.substring(daystart+5,daystart+8);
                            MothToInt mt= new MothToInt();
                            Month=mt.month(month);
                            Day=Integer.parseInt(day);
                            String year=temp.substring(daystart+9, daystart+13);
                            Year=Integer.parseInt(year);

                        }
                    //detect the hyperlink of the file
                    int start = -1;
                    int start1 =temp.indexOf("href");
                    int start2 =temp.indexOf("HREF");
```

```
                        if(start1!=-1)
                        {start=start1;}
                        if(start2!=-1)
                        {start=start2;}
                    if(start!=-1)
                        {
                        int end=-1;
                        while(end==-1)
                        {
                            temp=temp+br.readLine();
                            int end1=temp.indexOf("</a>",start);//A a?
                            int end2=temp.indexOf("</A>",start);//A a?
                                if(end1!=-1)
                                {end=end1;}
                                if(end2!=-1)
                                {end=end2;}
                        }

                                res=temp.substring(start+6,end);

                            int domainstart=res.indexOf("http://");
                            int domainend=res.indexOf("/", domainstart+7);
                            String restemp=res.substring(domainstart,
domainend-1);
                            if(!listtemp.contains(restemp))
                            {listtemp.add(restemp);}
                        }

                    }
                }
                br.close();
                read.close();
    }
```

Display function of URLTest.java- accessing into internet, then connect and rearrange the information of the webpage:

```
    display(String addr)
            {
      String link=null;
      String slip2=null;
      int start,end=0;
        URL url;
        try
                    {
                        start=addr.indexOf("www.");
```

```
                       if(start!=1)
                       {
                        end=addr.indexOf(".", start+4);
                        slip2=addr.substring(start+4,end);
                       }
                       else
                       {
                          start=addr.indexOf("http://");
                          end=addr.indexOf(".", start);
                          slip2=addr.substring(start+7,end);
                       }
                       String slip1="http://www.whois.org/whois_new.cgi?d=";
                       int last=addr.lastIndexOf(".");
                       String slip3=addr.substring(last+1);
                       link=slip1+slip2+"&tld="+slip3;

                          url = new URL(link);
                          InputStream ins = url.openStream();
                          BufferedReader bReader = new BufferedReader(new
InputStreamReader(ins));
                          String info = bReader.readLine();
                    while(info != null)

                    {
                    int timecreate=info.indexOf("Created on");
                    int timestart=info.indexOf("Creation date");
                    //int time_creation = -1;
                    if(timecreate!=-1)
                    {
                        int tempstart1= info.indexOf(":");
                        CreationDate =info.substring(tempstart1+1);
                    }
                    if(timestart!=-1)
                    {
                        int tempstart2= info.indexOf(":");
                        CreationDate =info.substring(tempstart2+1);
                    }

                     int timeexpire=info.indexOf("Expires on");
                     int timeexpiration=info.indexOf("Expiration date");
                     if(timeexpire!=-1)
                     {
                        int temp= info.indexOf(":");
                        ExpirationDate=info.substring(temp+1);
```

```
          }
          if(timeexpiration!=-1)
          {
              int temp1=info.indexOf(":");
              ExpirationDate=info.substring(temp1+1);
          }
                     info = bReader.readLine();
         }
           }
           catch(MalformedURLException e)
           {
           System.out.println(e);
           }
          catch(IOException e)
           {
               System.out.println(e);
           }
}
```

Växjö
University

Matematiska och systemtekniska institutionen
SE-351 95 Växjö

Tel. +46 (0)470 70 80 00, fax +46 (0)470 840 04
http://www.vxu.se/msi/