CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

Informatics

Department of Information Technologies

Diploma Thesis

SPAM FILTERING

Author: Bc. Kateřina Selingerová

Supervisor: Ing. Miloš Ulman, Ph.D.

Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technologies

Academic year 2008/2009

DIPLOMA THESIS ASSIGNMENT

Kateřina Selingerová

specialization of the study: Informatics

In accordance with the Study and Examination Regulations of the Czech University of Life Sciences Prague, Article 17, the Head of the Department assigns the following diploma thesis to

Thesis title: **Spam filtering**

The structure of the diploma thesis:

1. Introduction
2. Objectives of thesis and methodology
3. Basic characteristics of spam filtering
4. Spam filtering in practice
5. Conclusions
6. Bibliography
7. Supplements

The proposed extent of the thesis: 50 - 60 pages

Bibliography:

Goodman, D. Spam wars: our last best chance to defeat spammers, scammers, and hackers. New York: SelectBooks, 2004. 330 s. ISBN 1-59079-063-4

Feinstein, K. How to do everything to fight spam, viruses, pop-ups, and spyware. New York: McGraw-Hill/Osborne, 2004. 234 s.+ 1 CD ROM. ISBN 0-07-225655-9

Costales, B., Flynt, M. Sendmail milters : a guide for fighting spam. Upper Saddle River: Addison-Wesley, 2005. 329 s. ISBN 0-321-21333-5

Polčák, R. Právo na internetu - spam a odpovědnost ISP. Brno: Computer Press, 2007. 150 s. ISBN 978-80-251-1777-4

The Diploma Thesis Supervisor: **Ing. Miloš Ulman**

Deadline of the diploma thesis submission: April 2010

.....
Head of the Department

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE
Provozně ekonomická fakulta

.....
Dean

In Prague: 22th December 2008

Declaration

I declare that I have worked on my diploma thesis titled “Spam Filtering” by myself and that I have used only the sources mentioned at the end of the thesis.

In Prague, ...

Kateřina Selingerová

Acknowledgement

I would like to thank to Ing. Miloš Ulman, Ph.D. for his advice and supervision of my diploma thesis. Further thanks belong to my Erasmus friends who helped me to distribute my questionnaire, to Mgr. Jan Čihák, Mr. Miroslav Chocholouš, Mr. Petr Chocholouš and Mr. David Fridrich for their help with the interviews for my research and to Mr. Mathias Sajovitz for English corrections.

Spam Filtering

Spamové filtrování

Spam filtering

Summary

Nowadays, spam is, without any doubt, one of the biggest and most common Internet threats. Almost every user who has an email account has already received some kind of spam email. This thesis deals with the spam problem, especially with the techniques tackling spam. These techniques are called spam filtering.

The first part of the thesis deals with the well known techniques from a basic theoretical viewpoint. The second part includes the analyses of different spam filtering approaches in four chosen companies and also a short summary, advantages and disadvantages of these techniques. The third part is based on the results of questionnaires sent to the citizens of five European countries. The answers of 232 respondents demonstrate the situation in connection with the content and the volume of spam in chosen countries and provide a comparison of spamming in these countries.

Keywords

Spam, spam filtering, blacklist, greylisting, Bayesian filtering, malware

Spamové filtrování

Souhrn

Spam je v dnešní době jednou z největších hrozeb Internetu. Téměr každý uživatel vlastnící emailovou adresu již někdy obdržel nějakou formu spamového emailu. Tato diplomová práce pojednává o problematice spamu, především o možnostech obrany proti spamu. Tyto obranné techniky a přístupy se nazývají spamové filtrování.

První část diplomové práce pojednává o základních technikách spamového filtrování z hlediska již zveřejných poznatků o této problematice. Druhá část zahrnuje analýzu přístupů k spamovému filtrování ve čtyřech vybraných společnostech. Nabízí také shrnutí, výhody a nevýhody těchto technik. Třetí část je založena na výsledcích dotazníku rozeslaného do pěti vybraných evropských států. Zpracované odpovědi od 232 respondentů ukazují zajímavosti a poznatky týkající se obsahu a množství spamových emailů ve vybraných zemích a nabízejí výsledné porovnání zemí z hlediska spamu.

Klíčová slova

Spam, spamové filtrování, blacklist, greylisting, Bayesianské filtrování, malware

Contents

1. Introduction

One of the most common Internet threats is without a doubt spam. All over the world, 14, 5 billions of spam messages are sent every day and almost every owner of an email account has already received some kind of spam email. Spam emails are not just annoying, they can also include other dangerous threats such as computer viruses, worms, Trojan horses or spyware software.

Nowadays, every email provider and every company administering its own email domain and email boxes is using spam filters to protect themselves against spam. Worldwide, the amount of spam messages creates 98, 5% of all email messages. There are different techniques for spam filtering, different levels of protection and different software that can be used. Each company and each provider uses different settings for these spam filters to make them accurate for their purposes. Usually, a special administrator takes care of these settings, controls the amount of spam, the percentage of false negative and false positive emails, blacklists to be used or keywords to be found in the subjects or contents of the email messages.

There are different companies dealing with spam problems. Some of them have the users (employees) well educated in IT and on the Internet issue, some of the users are common, customary users of email boxes. Some of the company's administrators are responsible for just a few email boxes; some of them take responsibility for millions of them. Some of the companies have financial means for securing their email boxes, some of them have to use just free anti-spam software. So what is the difference between spam filtering in these companies? Is it true that the company with the largest responsibility for its users is the best secured? Is it true that commercial software is better than free software? Can this problem be generalized?

2. Objectives and Methodology

2.1 Objectives

In the diploma thesis **Spam filtering** will be summarized well known and less known information about the spam filtering problem. The first objective includes the organisation of the basic information into a brief summary, comparing some important definitions and demonstrating the most important facts about Internet security and spam problem.

The objective of the second part, the part of the author's own research, is to compare spam filtering in four different companies. The author has chosen different companies in terms of how they use company email. In the first one the users are usually elderly and not well educated in computer usage. In the second one, there are a lot of young people with experience concerning computer usage and internet knowledge. The third company is an IT company and the fourth company provides email services to millions of different users all over the Czech Republic. The main objective in this part is to find out if there are important differences in the amounts of received spam in these companies, to make some basic analyzes in order to find the advantages and disadvantages of used spam filters and to compare the approaches of email security in these companies.

The third part is the international questionnaire, sent by the author of this diploma thesis to the citizens of five chosen European countries. This questionnaire should show the differences between the content of spam emails received by the respondents, the amount of false positive and false negative emails depending on the email provider and the amount of received spam in different countries.

2.2 Methodology

Information and facts for this thesis was especially retrieved from technical literature, journals and Internet resources. The sources for the research part are the rewritten dialogues between the author and the company's employees. Also the results from the questionnaires are used in this thesis.

The first step was to find all possible information about spam filtering for the first part of the thesis. The next step was to arrange the meetings with the creators and administrators of the company email accounts in four chosen companies. It was necessary to prepare the questions and the themes for the conversation in order to be sure that almost the same questions will be answered. After these interviews, the conversations were rewritten, sorted and used for the research part of the thesis. The comparison of these spam filtering approaches was made in order to show some basic facts and ideas about the spam filtering problem in the companies.

Another step was the creation and the distributing the questionnaires. After receiving all the answers the author had to sort the answers, compare them with worldwide results and think about the consequences. Some graphs and tables were created to allow for the better comparison of the results. The final step of the research part was to point out some problems of spam filtering, to think about better solutions and to provide an overview of the possibilities of spam filtering in practice.

3. Basic characteristics of spam filtering

3.1 Internet threats - malware

Malware is the general expression for all kind of threats they can “attack” computers. It can be computer a virus, a worm or a Trojan horse, but also adware, spyware, rootkits, pop-ups and so on. It is not unusual that some types are cooperating together to take control over the computer. *“Malware is malicious software that is installed on your PC usually without your knowledge and it can enter your PC as a result of surfing the Internet and in a variety of different ways. Malware is capable of spying on your surfing habits, logging your passwords by observing your keystrokes, stealing your identity, reading your email, hijacking your browser to web pages that "phish" for your personal information, and a variety of other invasive tactics.”* (www.spamlaws.com) Malware is usually spread through the Internet; there is also the possibility to infect a computer with some infected USB drivers, CD or DVDs or external hard drivers. The most effective ways how to infect a computer via Internet connections are emailing, having the user click on unwanted advertisements or pop-ups and the downloading of free software or games.

When malware is created, it is usually inserted just into a small number of web sites or emails. After that, redirection is used to “bring” user to the infected pages. Another possibility to spread malware is creating proxy web pages. These pages are able to hide the malware for a longer time and they can also look like some official and well-known web pages, the users are usually safely visiting. (www.spamlaws.com)

3.1.1 Computer virus

The real meaning of the word computer virus is that a computer virus can copy itself, infect a computer and spread to other computers using the infected files. However, nowadays the word virus is used like the common term for all unwanted software like Trojan horses, computer worms, spyware, adware and most of the rootkits.

Nowadays, almost every computer user, especially with a connection to the Internet, already found some kind of virus on his computer. The main signs of the infection are:

- *Computer functions slow down.*
- *Computer reacts slowly and is frozen more often.*
- *Sometimes computer restarts itself.*
- *Uncommon error messages or dialog boxes are shown unexpectedly.*
- *Applications are not working correctly.*
- *There are some printing problems or fails.*

(www.spamlaws.com)

The web page www.spamlaws.com publishes a simple and understandable distribution of the computer viruses.

- **Boot sector viruses** are the viruses which are infecting diskettes and hard drives. The virus infects the boot or Master Boot Record of the disk and when the user re-boots the computer and the infected diskette is inside, the virus is spread into the computer. At this time, the computer infects all the diskettes that will be put into the computer.
- **Program viruses** *"becomes active when the program file (usually with extensions .BIN, .COM, .EXE, .OVL, .DRV) carrying the virus is opened. Once active, the virus will make copies of itself and will infect other programs on the computer."* (www.spamlaws.com)

- **Multipartite viruses** are connected boot and program viruses. It begins with the infection of a file. If the infected program is running, it changes the boot record. When the computer is turned on again, the virus infects the disk and other programs.
- **Stealth viruses** are able to change themselves in case they are not recognised by the antivirus software. They can change their size or hide themselves in the memory.
- **Polymorphic viruses** are changing themselves all the time, usually by changing their binary patterns. So it is very difficult have them detected by the antivirus software.
- **Macro viruses** are programmed like a macro in documents. Especially Microsoft Word or Excel support macros in their documents. Once the computer is infected, all documents produced on this computer are including infected macros.

A lot of viruses are spread via emails. They are usually included in the attachment. These emails can be normal emails from a user's contacts, who have their computers infected or it can be an email received as spam. Viruses infecting computers via emails are usually hidden inside Microsoft Word files and when they are opened they can infect the computer.

3.1.2 Trojan horses

According to Hák (2005), this type of malware is not able to reproduce itself or to infect files. The only action of a Trojan horse is that it is acting like an .exe program and that it deletes the concerned file. There are some types of Trojan horses such as **Password-stealing Trojan horse** (recording pressing the keywords and sending this information to the creator), **destructive Trojan horses** (deleting the files or formatting the disc), **TrojanDownloaders** (trying to download other damaging software from the Internet) or **Proxy Trojan horses** (using the computer to send spam).

Trojan horses are downloaded as software, games; they can also be hidden in website links, pop-up or banner ads. The biggest threat inflicted by a Trojan horse is that it can take over a user's computer, change a computer or its system settings and turn off the anti-virus or anti-spyware programs. Once the Trojan horse is executed, it creates a so called backdoor. It allows a hacker to access a user's computer and put some additional viruses or software like spyware onto the user's computer. There are some special Trojan horses which are able to turn on the webcam at the user's computer so the hacker can watch the user or his room to get some personal information. All the Internet threats are very well connected. The usual way how the computer gets infected is the downloading of some spyware or program from the Internet. The most usual signs that the computer is infected by Trojan horse are:

- some emails can be sent to user's contact without user's note
- some pop-up windows start to appear
- the anti-virus acts like it is turned off
- wallpaper and taskbar changes
- sometimes the mouse is not working as usual.

(www.spamlaws.com)

3.1.3 Worms

Mr. Hák (2005) describes computer worms as network packets working on a lower layer than the classic viruses. These packets are redirected from the infected file to other systems on the Internet. The functionality of the worm is based on security "holes" in the operating system through which the worms are able to reproduce themselves. The result can be total a congestion of the network.

A computer worm is very similar to a computer virus. The computer worm does not need some host file to reproduce itself. The worm can find the weakness of the computer systems and thanks to the security "hole" propagate itself. The worm does not need the

user to execute him (by double-clicking the file or opening the email attachments), he releases the document with the macro included inside and that gives him the opportunity to "travel" from computer to computer. The worms are usually infecting the whole network using many infections. (www.spamlaws.com)

Some of the worms are also used to collect data from the computers. Similar to spyware, they focus on usernames, passwords, bank account number or the login dates of the PayPal service. These worms *"operate by automatically filtering network traffic for patterns of data that typically come before the transmission of a passing username and password."* (www.spamlaws.com)

3.1.4 Spyware and Adware

Internet threats other than computer viruses are spywares and adware. This software is running in the computer and even the user knows about it. *"Generically, adware (spelled all lower case) is any software application in which advertising banners are displayed while the program is running. The authors of these applications include additional code that delivers the ads, which can be viewed through pop-up windows or through a bar that appears on a computer screen. The justification for adware is that it helps recover programming development cost and helps to hold down the cost for the user."* (http://searchcio-midmarket.techtarget.com)

In some cases, the user knows about the installation of the adware. This is usually the case when he wants to download some free application. Usually, the terms specify that he can download it for free but that advertising adware will be included. So the adware is the price for downloading for free. One of the disadvantages is that the adware is running in the computer's background, so it obviously causes a slowdown of the computer. Adware is using the system memory, CPU cycles and Internet bandwidth. (Feistein 2004)

"Adware is the software that periodically pops up advertisements on a user's computer. It displays ads targeted to the individual user based on key words entered in search engines and the types of Web sites the user visits. The marketing data are collected periodically and sent in the background to the adware Web server. Adware is known as "contextual marketing."" (www.pcmag.com) Adware is able to track what a customer is searching for on the Internet. This information is sent to the server which processes these facts about the user's behaviour and "sends back" the advertisement related to pages or keywords the user was searching for. Nowadays, there is a "thin line" between adware and a little more dangerous spyware. The software usually connects these two threats together. *"Adware, or advertising-supported software, is any software package which automatically plays, displays, or downloads advertisements to a computer. These advertisements can be in the form of a pop-up. The object of the Adware is to generate revenue for its author. Adware, by itself, is harmless; however, some adware may come with integrated spyware and other privacy-invasive software."* (www.hky.com) Who receives all this personal data of the user has a good possibility to target related advertisement on him. The more advertisement the advertiser is sending the bigger is the chance to find some serious customer. There is also the possibility to just sell this information to other advertisers or third parties.

"Spyware is a type of malware that can be installed on computers, and which collects small pieces of information about users without their knowledge." (www.computershack.info) The most common ways how spyware can be installed onto a user's computer is downloading some suspicious email attachment, downloading inside another downloaded program or downloading after the user clicks on some link in pop-up advertising windows. There are special sites where the user has to download some application in order to see some video or to listen to some music...it is spyware instead. Sometimes, spyware can be wanted. For example parents use it to control what their children are doing on the computer or employers can control their employees and see what they are doing during their working hours. But this data is not sent anywhere; it is usually saved onto the computer to be gone through it later. (www.spamlaws.com)

"The presence of spyware is typically hidden from the user, and can be difficult to detect. Typically, spyware is secretly installed on the user's personal computer. Sometimes, however, spywares such as keyloggers are installed by the owner of a shared, corporate, or public computer on purpose in order to secretly monitor other users."
(www.computershack.info)

So the danger of spyware is that the information about the user is collected and sent to someone else, usually via the Internet. Spyware can collect private information like passwords, bank account numbers, information about chat programs, web pages visited, cookies and any other information about your hard drive. Usually, the creators do not want to steal a user's money directly. They use this information to create some interesting offers or advertisements exactly suitable for each user. Spywares collect for example the information about the web pages the user has visited or the keywords he was looking for.

"Attempting or gaining access to someone's computer without their consent or knowledge is criminally illegal according to computer crime laws, such as the United States Computer Fraud and Abuse Act and the United Kingdom's Computer Misuse Act."
(www.spamlaws.com)

It does not mean that it is illegal. The spyware creators are protecting themselves very well. The most important for them is that they are mentioning their spyware in the licence agreement of the product where the spyware is hidden. The user has to click "Agree" or "Ok" and at this moment he is in the fact agreeing with the spyware installation. Even if the user is reading the licence carefully, it is very difficult to understand its implications.

But of course, there are some forms of spyware that are for sure illegal. If the creator does not include the information about the spyware in the licence agreement or does not mention it before the installation at all, it is considered an illegal act.

"Several bills have been voted on in the United States Congress including the Spy Block Act (Software Principles Yielding Better Levels of Consumer Knowledge) and the SPY Act (Securely Protect Yourself Against Cyber Trespass), both passed in 2004, and the I-SPY Act (Internet Spyware Prevention) passed in 2005 and reintroduced in 2007." (www.spamlaws.com) These bills give the user the possibility to protect himself legally. If somebody is controlling his computer, collecting his data or sending him pop-ups which are not possible to close, he can use these bills against him.

All the Internet threats mentioned are closely connected. *"eBlaster is a well known Trojan application developed by SpectorSoft, a company that has established a reputation for developing various spyware programs. This infection is frequently distributed via email and targets individuals and businesses using a Windows operating system and Hotmail or Yahoo mail accounts. Once installed, eBlaster has the ability capture email contents, instant messages, chat sessions, any web sites the user visited and keystrokes entered on their computer. This data is then automatically forwarded to the email address of the Trojan creator."*

The official SpectorSoft pages write about the eBlaster: *"Install eBlaster on the computer you wish to monitor and walk away. It's that simple! eBlaster will immediately go to work by automatically recording EVERYTHING your children and employees do online, including:*

- *Emails sent and received*
- *Both sides of chats and instant messages*
- *Keystrokes typed*
- *Files uploaded/downloaded*
- *User activity*
- *Web sites visited*
- *Online searches*
- *Facebook and MySpace activity*
- *Program activity"*

So eBlaster software is very complex in the way of spying a user's activity at all. The official version is that it is software created to help parents and employers to get an overview of their children or employees. It could be easily misused to exploit the users. SpectorSoft is protecting itself by warning that any usage of their software without letting the user know is a violation of their terms.

Another threat connected to spyware is a browser hijacker. This program is changing the browser settings without the user knowing about it. Usually, the program changes homepage, browser settings or it can redirect the user to a site containing malware. It can also decrease the security settings of a user's Internet browser so that the computer is more likely to be infected by malware. Nowadays, one of the most dangerous types of spyware is the keystroke logger or keylogger. This program tracks the user's activity online and thus can collect private data, usernames, passwords and so on.

It is not easy to find the exact line between adware and spyware. These two "members" of malware are very often connected together. Adware can be a small part of spyware and so the unwanted advertisements shown to the user are often carefully chosen after the unsuspected research lead by the spyware part.

3.1.5 Pop-up windows

In general, pop-ups are the windows that appear on a user's computer screen without him wanting them there. They can appear after the user clicks on some link, after visiting some web page or they can also appear just randomly. These random windows are usually full of advertisements, interesting offers and sometimes full of inappropriate content like pornography. The pop-ups are not only annoying but they can also be a sign that the user's computer is infected by some spyware or adware or the worst – a Trojan horse. (www.spamlaws.com)

According to Feinstein (2004) *"pop-up can be defined as a separate browser window, usually displaying the advertisement, that open automatically when you visit or leave a web page"* Instead of traditional banners, the pop-up window "grabs" the user's attention. The strategy of pop-ups is to make the user look at the advertisement and click on an interesting offer.

One type of pop-up is the so-called **pop-under**. These windows are not shown at the front of the web page but they stay hidden behind the web browser window. When the user closes all browser windows, the pop-under stays on the screen. It can be used like a "second chance" of the webpage to get the user's attention. The advertisers also work with the possibility that the user finds this pop-under later and he is not able to connect it with some specific web page so he is not able to avoid visiting this page. The pop-up and pop-under windows are also able to reproduce themselves. These windows are able to create another pop-up every minute. When the user closes one window, another one appears in a few seconds. There is usually one initial window hidden on the screen. Another type of pop-up appears at the moment the user closes the specific web page, visits another one or closes the whole browser. And these pop-ups can also be generating other ones when they are closed. Not all the pop-ups show as a separate window. There are so-called **floating pop-ups** which appear directly on the web site, floating around the page the user is visiting. This type can also be created in Macromedia's Flash software so the advertisements are "accompanied" by sounds and animations. (Feinstein, 2004)

Not all pop-ups present some offer or advertisement. Some of them are more sophisticated. They look like an official system error message or like a window from instant messaging asking the user to install some application or software. These pop-ups are usually connected with spam messages. *"This is actually very common in versions 2000 and XP of the Windows operating system. These messages are typically distributed by intruders taking advantage of vulnerable entry points such as TCP ports 135, 137, 139 and 145 or UDP ports 135, 137 and 138."* (www.spamlaws.com)

The worst types of pop-ups are those which are offering some sexual services, pornography or other offensive content. These pop-ups are 99% likely to be connected to some type of adware or spyware.

The largest part of pop-ups is created in JavaScript. On the Internet, there are many pages with advice and codes on how to create pop-ups on a web page. The designers of a pop-up can set the size of the pop-up window, the type (pop-up, pop-under), position of the pop-up and so on. The window can be created so big that there is no obvious chance to close it or it can fit the browser perfectly that it is not obvious that it is a pop-up window. In connection with Flash or DHTML (Dynamic Hypertext Markup Language), floating pop-ups can be created. (Feinstein, 2004)

3.1.6 Rootkits

"A rootkit is malware that is installed on a computer by an intruder for the purpose of gaining control of the computer while avoiding detection. Unlike other malware, rootkits are capable of avoiding the operating system scan and other related antivirus/anti-spyware programs by hiding files and concealing running processes from the computer's operating system." (www.spamlaws.com) A rootkit acts like a Trojan horse in cooperation with other malware. Rootkits which are able to work with files, folders or drivers are called **User mode rootkits**. Their job is to copy files onto the hard disks and they are activated every time the user turns the computer on. They are not so dangerous because they can easily be tracked and removed. In contrast to the User mode rootkits, the **Kernel mode rootkits** are working on the PC operating system layer and so they are able to influence the operating system of the computer. It leads to unexpected events or system crashes. The most resistant type of rootkits is **Firmware rootkits**. Every time the computer is turned off, they are creating wrong code inside the firmware. (www.spamlaws.com)

"The main purpose of a rootkit is to make unauthorized modifications to the software in your PC." (www.hyders.com) It can cooperate with spyware, because rootkits

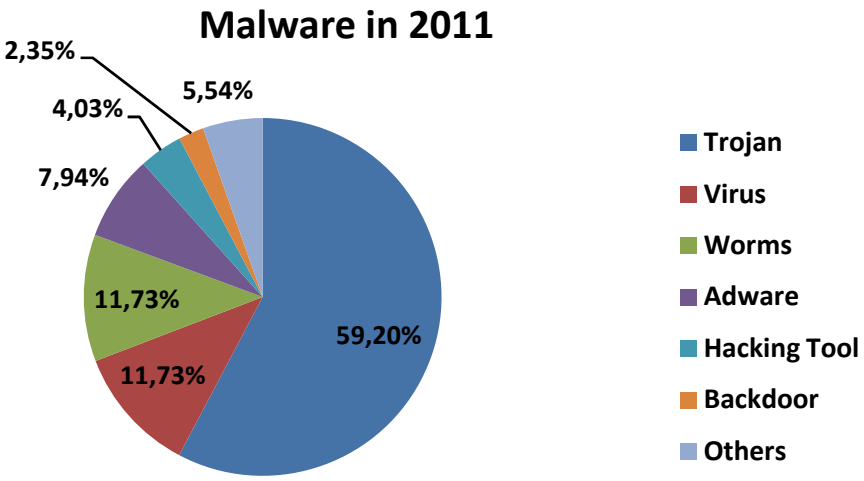
can make it harder to find it. Rootkits are also able to create backdoors; it *"is a modification that is built into a software program in your computer that is not part of the original design of the program. It creates a hidden feature in the software program that acts like a signature so the intruder can use the software for malicious purposes without being detected."* (www.hyders.com) They can also modify the bytes and give them specific orders which disable the software protection. Another of their abilities is to rewrite the source codes of the PC's software.

3.1.7 Malware nowadays

On the web page pandasecurity.com, there is a study about the most usual types of malware. The results of this study are based on the Panda ActivScan, a free online scanner. The study says that 34% of the malware that has ever existed was created in 2010. *"The speed with which the number of new threats is growing has dropped with respect to 2009: since 2003, new threats grew by at least 100 percent every year. However, so far this year they have increased by around 50 percent."*

"Besides offering information about the main security holes in Windows and Mac, the 2010 Annual Security Report also covers the most important security incidents affecting the most popular social networking sites. Facebook and Twitter have been most affected, but there have also been attacks on other sites like LinkedIn or Fotolog, for example." (www.pandasecurity.com)

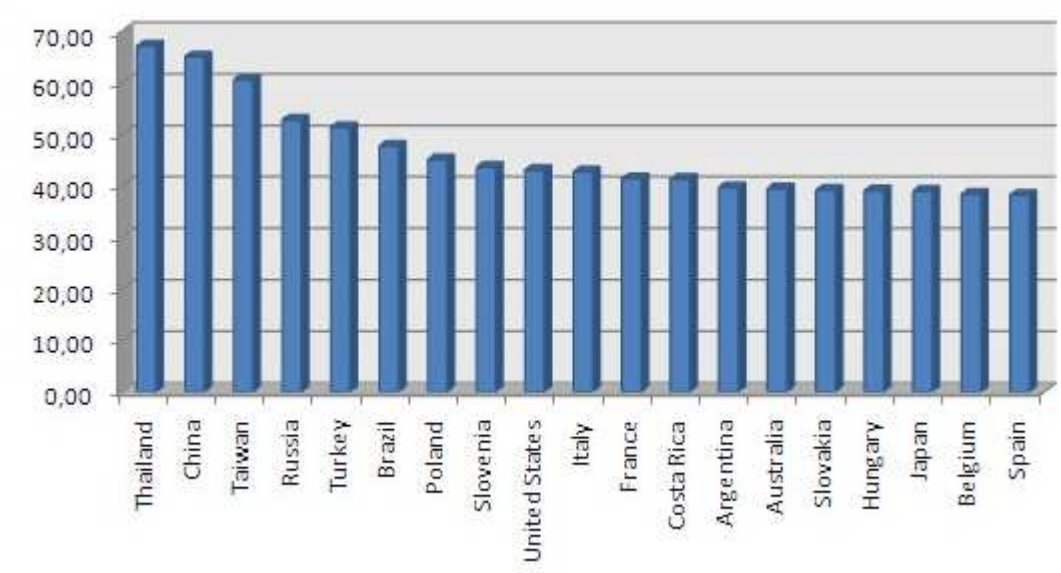
The study says that in January 2011, 50% of the computers scanned all over the world were infected with some kind of malware. The most widespread malware is Trojan horse, followed by viruses and worms.



Pic.1: Malware around the world in January 2011

(Source: http://press.pandasecurity.com/wp-content/uploads/2011/02/Graph1.jpg)

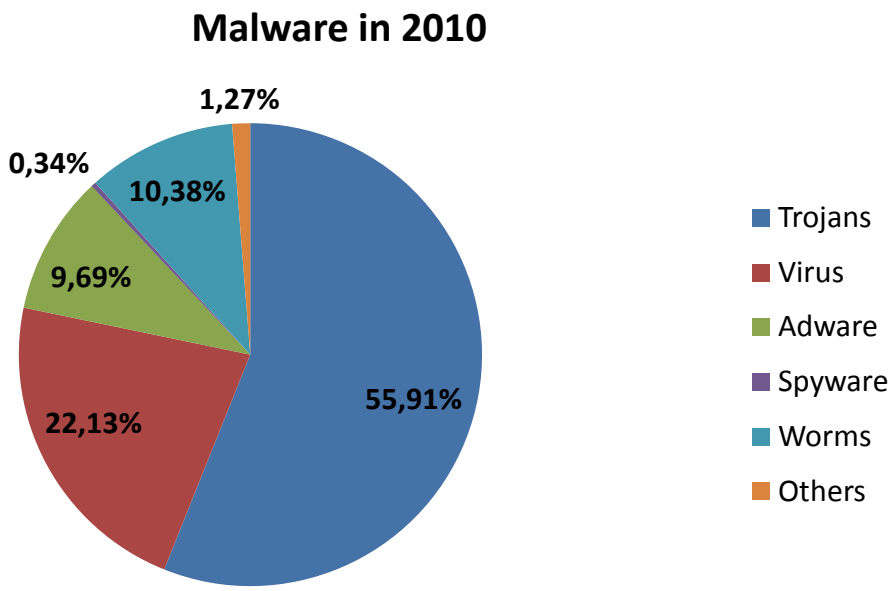
The same web pages also published the list of the countries with the largest amount of infections. The first place is taken by Thailand, the second by China, and the third place by Taiwan followed by Russia, Turkey, Brazil and Poland.



Pic.2: The list of the countries with the largest amount of computer infections

(Source: http://press.pandasecurity.com/wp-content/uploads/2011/02/Graph3.jpg)

The Annual Security Report 2010, created also by Panda Security, shows that the trend is almost the same as in the January 2011.



Pic.3: Malware around the world in 2010

(Source: http://press.pandasecurity.com/wp-content/uploads/2011/01/MALWARE-FAMILIES.jpg)

Also the list of the countries with the highest percentage of computer infections is not so different. Thailand is leading, followed by China on the second place and Taiwan on the third place. These countries are followed by Ukraine, Latvia and Saudi Arabia. Russia ranks 7th, Turkey 11th and Brazil 14th.

3.2 Spam

Without a doubt, spam is one of the most widespread Internet threats all around the world. Sometimes, spam is considered to be a part of malware. This definition is wrong because by itself, spam cannot infect a user's computer, it cannot replicate itself and there is no possibility that spam installs some backdoors into the system. But it does not mean that spam is not dangerous.

3.2.1 Basic characteristics

There are many spam definitions. Some are very simple, some are very complex. Here are some of the most exact and interesting ones. According to Ken Feinstein (2004) *"spam is unsolicited email sent in bulk from an organisation (or person) with which you have no affiliation."* The server www.spamlaws.com presents the definition of spam as *"the term that refers to submitting the same message to a large group of individuals in an effort to force the message onto people who would otherwise choose not to receive this message."* According to Adámek (2009), spam is characterised as messages automatically sent to many receivers who did not order these emails and they are not able to cancel receiving them. These emails are usually sent from foreign countries, written in foreign languages (especially in English) and they are offering some products and services. These offers are sent frequently and usually contain the same or very similar offers. Not so aggressive forms of spam are messages sent by national companies with the possibility to cancel receiving these offers.

Spam emails have some features which are common to all types of spam. These features are anonymity of the sender, sending the offer to a large amount of email addresses and the fact that the receiver of spam was not asking for it. The statistics say that almost 99% of the email account owners received some kind of spam message at least once. Just 14% of the receivers read these offers and the percentage of the users who really used the offers to buy some product is even lower – just 4%. (www.spamlaws.com)

3.2.2 Content and categories of the spam

Many users think that spam is usually just about some “interesting” offers or advertisements which are annoying and just fill up email boxes. But in fact there are different strategies and different approaches via which spammers (persons creating the spam) are trying to “get” money from users. So it is very good to know about these categories and learn how to recognise how each spam is working.

One of the most usual types of spam is so called **financial spam**. The page www.spamlaws.com shows two main types of spam related to financial risk. One of them is **the Nigerian sting**. *“This particular spam targets middle class, middle age, business and professional men who would never be as easily deceived by a lottery scam. Estimates put the losses from these "Nigerian Advance Fee" operations at over $1 million "every single day" in the U.S. alone.”* (www.crimes-of-persuasion.com)

So these emails include the message about some desperate and needy person who needs the user’s help. They say that they have a lot of money but they are just in a bad situation and they need to send their money to the user’s account. *“In this fraud mail is sent claiming that the sender has millions of dollars in an account that must be shipped out of his country. The recipient is asked to cover the expenses (a mere few thousands of dollars), and it return the millions will be transferred into the recipient’s bank account. The recipient will be asked to cover growing expenses before the millions are transported and eventually may even be conned into flying to that country”* (Costales, Flynt, 2005) The only thing they want is to find out information about the user’s bank account to completely withdraw the funds from it.

Another type of financial spam is called **phishing**. It is the spam that looks like an email from some reputable business or a company the user knows well. The content of these emails is usually asking for an update of a user’s personal information and identification dates like password, user name and so on. The message can also inform the

user that his account was hacked and they need this information to protect the account. The creators ask the user to send this information back or to click the link which will redirect the user to his account where he can change the settings. This information can be used to identity theft and to abuse it in many ways that are dangerous for the user's privacy. The phishing email looks almost the same as an official email the user is used to receive – the same address, graphic, footer. If the email includes some link to the company's web site, the site looks exactly the same as the official one with the only difference – the personal data is sent to the creator of the phishing email. (Adámek, 2009)

Another type of spam is the so-called **IT spam**. This kind of spam is widespread - especially at universities. The spammers are sending emails impersonating as a teacher, administrator, dean of the university or as a schoolmate. The content is usually asking for clicking the link to verify some information and so on. When the spammers reach this information, they are able to send more emails to the contacts from a user's email account. (www.spamlaws.com)

Nowadays, there are many possibilities for spammers to use social networks and community forums to spread spam. The spammers create their own accounts and then they are able to send all the advertisements to several forums and chat rooms. *"Advanced spammers often hack into their user profiles and go on numerous comment posting sprees filled spam messages. They also fill up the bulletin boards and even directly email other users with spam advertisements."* (www.spamlaws.com)

"We hate spam and phishing just as much as you do, and we work hard to keep it off Facebook. We do this in a number of ways. We employ technical systems to automatically detect and block spammy behaviour, provide easy report links across the site, and develop unique and innovative tools to help you protect your account and information from unauthorized access... According to our complaints, the defendants, among other things, represented that in order to qualify for certain fake or deceptive offers, people had to spam their friends, sign up for automatic mobile phone subscription services, or provide other information... In fact, Facebook holds the record for the two

largest judgments in the history of the CAN-SPAM Act - an $873 million judgment against Adam Guerbuez and Atlantis Blue Capital and a $711 million judgment against Sanford Wallace... Be wary of suspicious-looking posts or messages, and report spam when you see it using the links located throughout the site." (www.facebook.com, Updates in Facebook's Fight against Spam and Spammers, January 27th, 2011)

"Twitter strives to protect its users from spam and abuse. Technical abuse and user abuse is not tolerated on Twitter.com, and will result in permanent suspension. Any accounts engaging in the activities specified below are subject to permanent suspension..." (support.twitter.com)

The most common content of the spam is without any doubt advertising. Almost every Internet user who has an email account has already received at least one email offering some special product or service at a special price. Some of these emails are legitimate but some of them are based on deceptive content. The page www.spamlaws.com mentions the biggest problems caused by the spam. The first one is that the user simply wastes his time by filtering out all the emails he receives. Another one is that spam filters are using the bandwidth on the IP server and these filters can also prevent the user from receiving emails which is not a spam. Last but not least, another problem is that the spam is really annoying.

Of course, not all the advertisements are "innocent". One of the most common spam content is pornography. If the user receives this kind of spam, it does not mean necessarily that he has visited some pornographic web site, even though this is very usual. These emails are very well recognisable. Even if a user's email client or email account itself does not display images, the text has usually special graphics, there are many different links and the content itself is very "scandalous".

3.2.3 Spam and the law

Spam is considered the sending of a huge amount of emails, which the user does not ask for. It is not necessary just unsolicited advertising, but also emails with pornographic, insulting or racist content that could be regarded illegal.

In the Czech Republic, the primary rules of spamming through the Universal Law on Advertising are connected to the Act No 138/2002 Coll. from June 1st, 2002. The rules for emails to be considered spam were:

- It must be an advertisement.
- The advertisement must be unwanted.
- The advertisement must lead to the user's expenses or be annoying.

On September 7th, 2004 Act No 480/2004 Coll. became part of Czech laws, which is about some information companies' services and introduced some changes in some acts. The main change was that if the recipient has not agreed to the spread of business communication via electronic services, the delivering of such commercial emails is illegal even the recipient does not perceive them as annoying. (Mareš, 2010)

"The Directive 2002/58/EC was partly transposed in 2004 by Act No 480/2004 Coll. on certain information society services, embodying particular provisions on unsolicited commercial communications, and including new strong competence for OPDP in combating this "commercial spam". The Directive was essentially subsequently implemented in 2005 by Act No 127/2005 Coll. On electronic communications which simultaneously implements a number of other directives pertaining to the "telecommunications package"." (Twelfth Annual Report of the Article 29 Working Party on Data Protection, 2009)

3.2.4 Different roles in the spamming process

The process of sending a spam email is not only involving the spammer, advertiser and the recipient. There are many other parties involved. Of course, there are spammers, who are responsible for the majority of the spam emails on the Internet. *"They are sending massive amount of emails to the people who don't want it, using techniques such as Form: and Subject: message lines and relaying their tonnage through email systems owned by spam-friendly providers around the world or hijacked from unsuspecting computer owners"* (Goodman, 2004) Another important role "play" companies, which offer products or services via the Internet and hire spammers to do this instead of them. Spam usually includes the link to the websites of these companies. In the spamming process, there are also the companies providing their email lists to companies they want to offer their products. *"To reach as wide an audience as possible as cheaply as possible, frequently employ one or more Email Marketer characters to supply addresses and do the mailing."* (Goodman, 2004) All of the roles mentioned above are "originators" of the spamming process. Also scammers must be included here (broadcasting a large amount of messages, usually from some very distant place in order to get some personal data to do identity theft or card fraud) and moreover virus or worm writers using spam emails to spread their malware.

But spammers, advertisers and other originators could not be so successful without the software and facilities enabling them to send spam emails. There are spam software suppliers who offer their software to find and collect email addresses from the Internet, *"locate the computers that can be used as unsuspecting relays or proxies for untraceable spamming, and pump high volumes of mail through another character, the Bulletproof ISP."* (Goodman, 2004) There are also some Internet Service Providers who ignore spam reports or complaints. Sometimes also the infrastructure which cares about the transmission of the data around the world can be responsible for helping the spammers even if they do not know about it. The large amount of users is also helping the spammers. These users have already clicked on some spam link or the spyware was installed on their computers so they are sending a lot of spam emails every day. They do not know about it

and because they are usually online all day long, they are really helping the spammers to find new contacts. (Goodman, 2004)

On the other side, in the spamming process, like in all other processes, people take the role of trying to stop spam from spreading wide. There are still many Internet Service Providers (ISP) who take care about their web and try to fight spammers. The same could be said about the email administrators, especially those who are working for some university or a big company. They should block spam and set really save filters. Of course, there are antivirus and anti-spam creators and activists included in this group. In general, also the legislators, lawyers, courts, prosecutors are very important in the fight against spammers. (Goodman, 2004)

All these roles are very tightly connected. The people, who are trying to help, could be at the same time assisting the spammers without their knowledge. Also the spammers and cooperating companies could be victims of spammers. So the roles in the spamming process are not so obvious.

3.2.5 Collecting of email addresses

If spammers want their emails to be as effective as usual, they need to find a large number of email addresses. One of most often used methods is using the user's registration. Especially if the user registers himself on pages with non-verified content, pornographic or gambling web sites, there is an almost 100% chance that he will receive spam emails. But also the registration on creditable web pages could be abused if the administrator is not careful enough or he is cooperating with spammers or creates spam by himself.

According to Adámek (2009), another possibility for collecting email addresses can be some promotional actions, questionnaires or competitions. The trading with these email

lists is very usual in marketing, sometimes the lists are sorted by the themes, a customer is potentially interested in. These lists also include the names, addresses, age, sex and other information which can be used by more sophisticated spammers. Also phishing is working with this information when the official email with the name and personal dates is created.

Many spammers or marketing companies are using **spambots**. These software robots are going through the web pages and collect the email addresses made public on forums and web pages. The most common source is the "mailto:" commands in hyperlinks. The usual places where to find the email addresses are:

- *"chat rooms*
- *message boards*
- *Usenet newsgroups*
- *discussion mailing lists*
- *online service member directories*
- *web pages."*

  (Levine, Young, Everett-Church, 2004)

"A sequence of several characters with an @ sign and a dot or two that, once exposed, can be traded, rented, and sold for real money without your permission or knowledge; a character sequence that allows strangers to use up your bandwidth, mail server disk space, PC disk space, and time without your permission; a sequence that, no matter how much sanctity you ascribe to it, will be desecrated by spammers who wouldn't care less." (Goodman, 2004)

There is the problem that some free email providers can profit from allowing spammers to send messages to their users. It's also possible to collect the email addresses on web sites via some application like "send this paragraph to your friend" and other tricky forms. Another way is just to guess the email address. This method is called dictionary method. So the spammer has no idea if the address exists but he tries it. The first step is to collect the list of most used domains. Then, he can think about the names. *"Perhaps not*

surprisingly, the dictionary method works well with the last names too – and sports team names, cities and states, names of movie stars, names of book characters, and ever random words from the dictionary!" (Levine, Young, Everett-Church, 2004)

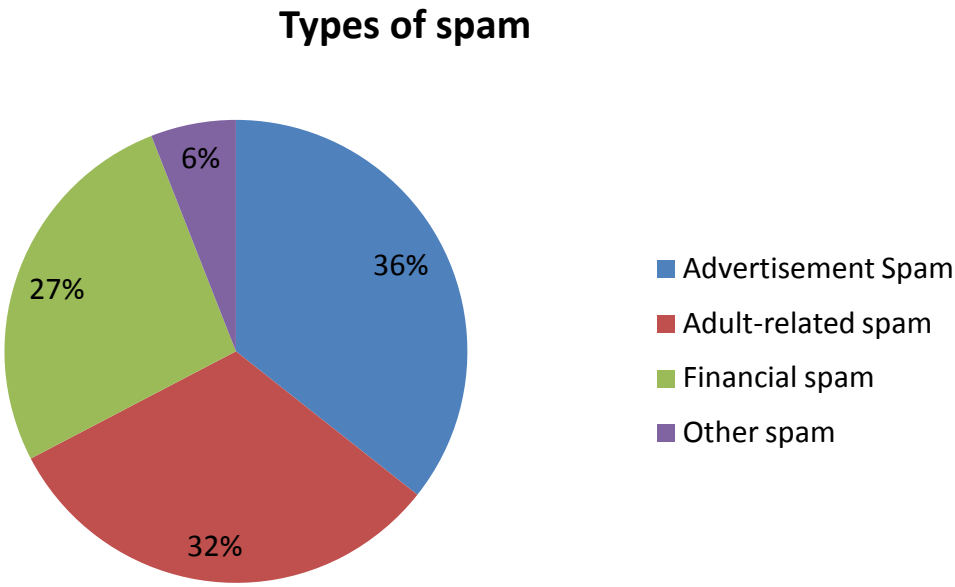
3.2.6 Why to send spam

The people who are sending spam emails have their reasons why to do it. According to Goodman (2004), they have different reasons. One is that there is a low percentage of success that the user will buy some product offered by the email in general. So the more email the company sends the higher the chance for profit. Another reason is the trade with the email addresses. *"A few advertisers actually spend $0.30or more per address from highly targeted lists from magazine subscribers or other sources that rent their customer lists using the direct mail model. Rather than place these highly valuable lists into the advertiser's hands, the list owners tend to require that the actual mailing process flow through their own trusted service bureaus."* (Goodman, 2004) This fact creates a big competition between spammers and the price of email lists gets really high.

Another reason is that for the spammer, the cost of sending spam emails is really low. *„A product seller without his or her own address database will have to pay for addresses, ranging from fractions of a penny to quite a few cents per address. But the disguised senders have their own lists, so there is no incremental cost per address for those spammers"* (Goodman, 2004) So for the marketers it is profitable to ask the spammer for help. They just create the offer they want to send and the spammer, using some special software, is able to send the offer to millions of email adreesses in a very short time.

3.2.7 Spam in numbers

In general according to the page www.spamlaws.com 14, 5 billions of spam messages are sent daily. It amounts to 43% of all the email traffic, other companies say 73%. The most usual content of spam emails are advertisements, the second place is taken by adult-related offers and on the third place rank emails related to financial matters.



Pic.4: Main content of spam (approximately)

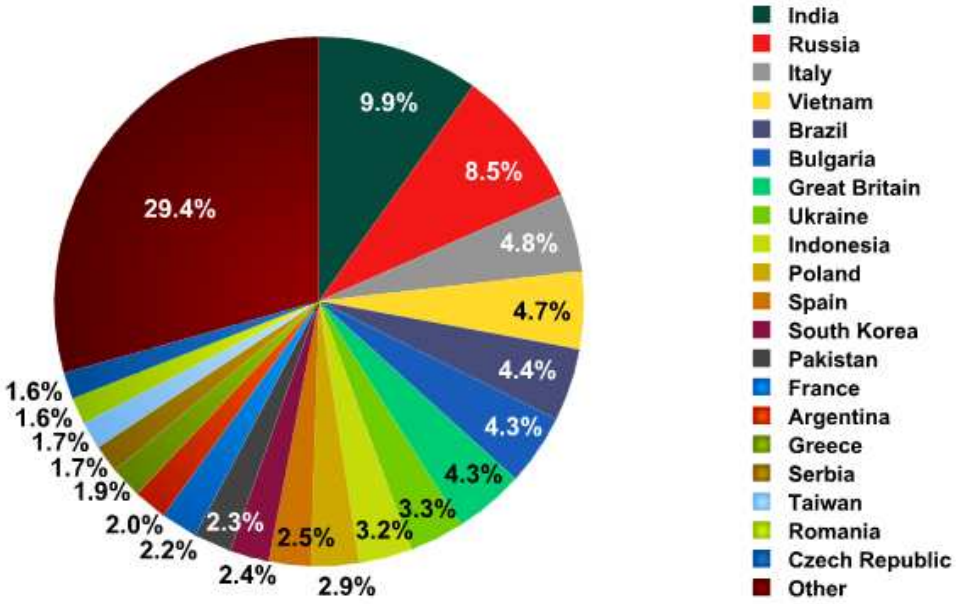
(Source: http://www.spamlaws.com/spam-stats.html)

According to analyses from Kaspersky Lab ZAO, in December 2010, spam in email traffic averaged 77, 1%. Spam with malicious attachments accounted for 4, 6% of all mail traffic. Graphical attachments were found in 1, 75% of all spam emails. The most malware was detected in emails sent from India, Russia and Vietnam. The most often sent malicious content was the Trojan horse called Trojan-Spy.HTMLFraud.gen which is trying to convince users to give away confidential or financial data. *"This particular Trojan uses spoofing technology and appears in the form of an HTML page resembling the site of a*

well-known bank or e-pay system where the user is asked to enter a login and a password." (www.securelist.com) Other malicious contents were especially email worms.

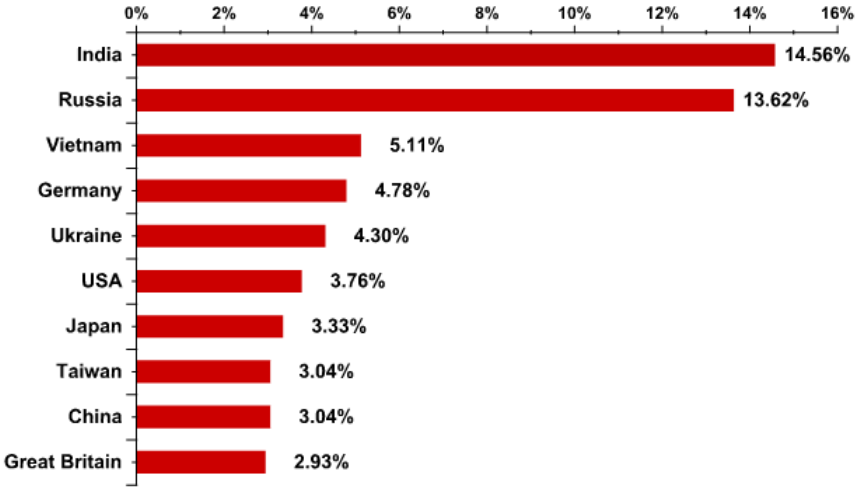
The main source country of spam in December 2010 was India, followed by Russia, Italy and Vietnam. The Czech Republic was on the 20th place with approximately 1, 6% of all the spam sent. (Picture 5)

Compared to these numbers, malicious content in spam emails was sent especially from India and Russia. The "first" European country is Germany on the 4th place. In general, the largest amount of malicious spam comes from Asia. (Picture 6)



Pic.5: Source countries of spam in December 2010

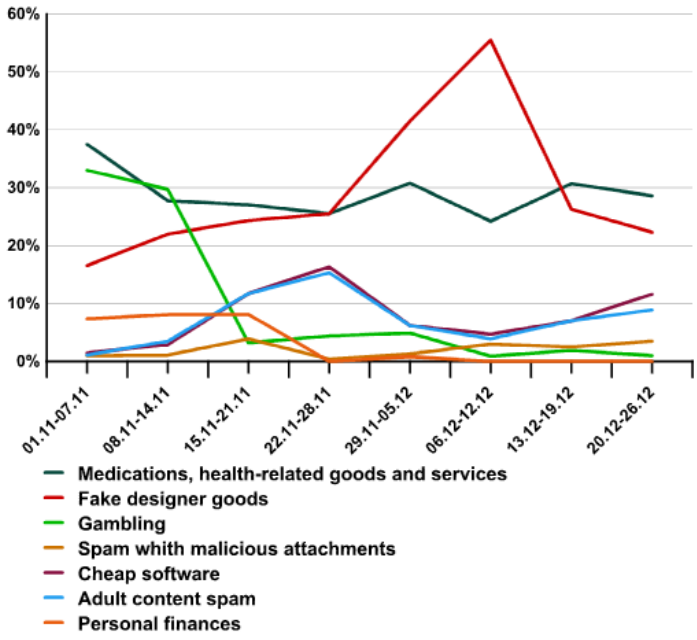
(http://www.securelist.com/en/images/vlill/spamreport_december2010_pic04.png)



Pic.6: Source countries of malicious content

(http://www.securelist.com/en/images/vlill/spamreport_december2010_pic05.png)

Depending on the content of spam, the most usual are advertisements about many types of medications (usually Viagra or medications requiring prescription), medical services (plastic surgery), fake products (replicas of watches), gambling (announcing to the user that he has won a lottery or he has some free money he can use to play in a casino) and offers for cheap software or adult sexual services.



Pic.6: Content of spam emails November-December 2010

*(*http://www.securelist.com/en/images/vlill/spamreport_december2010_pic07.png)

The page www.m86security.com made the statistics about spam from the last year ending in the first half of February, 2011 public. Surprisingly the amount of spam received by the sample of users the page is monitoring is decreasing. On the graph it is visible that the amount has decreased to approximately half the amount.

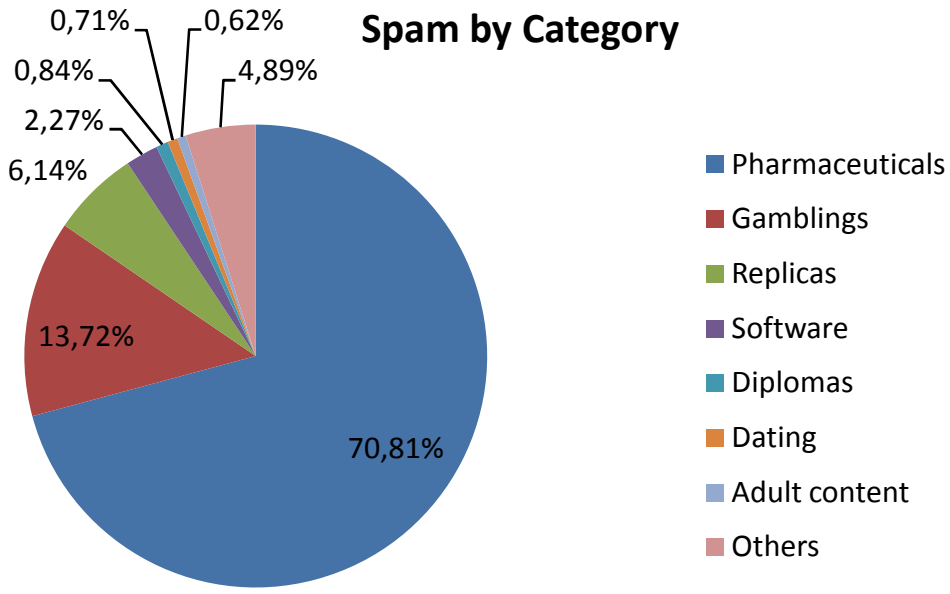
Pic.7: Amount of spam received from April 11th 2010 – February 20th 2011

(Source: http://www.m86security.com/images/trace/276/276-16-SVI_time.gif)

In the second week in February 2011, the origin of spam by countries was almost the same as in December 2010. This time the first place belonged to Russia, followed by India and Brazil. The European country placed first is Romania on the 8th place and Poland on the 10th place. But viewed by continent (continents are ranked in order of the volume of spam they contribute), Europe is first, followed by Asia, South America, North America, Africa and the lowest percentage is held by Oceania.

According to research of M86 Security Labs, the total percentage of spam received in email accounts is 85, 2% and the average size of spam email is 3Kb – 8Kb.

Unsurprisingly data about the content of the spam is collected. In the first place we still find medication and pharmaceutical offers, followed by gambling content and offers of different replica products.



Pic.8: Percentage of spam content

(Source: http://www.m86security.com/images/trace/276/276-5-spam_type.png)

3.2.8 Anti-spam techniques

Because spam is such a big problem nowadays, there must be a complex solution how the email box should be protected. There are many specific functions to avoid spam in anti-spam software. Because the number of these techniques is so large, only techniques used later in the company analysis are described.

One of the most usual problems with spam filtering is that the filters are not able to deal with all the spam and non-spam messages with a100% success. There are still false negative or false positive emails. According to the page www.pcguide.com, *"A false*

positive occurs when the scanning reports finding a virus when there is in fact no virus present. The chances of this occurring depend on the type of virus checking being done, and also on the general quality of the software... A false negative occurs when the scanning software does not find a virus that in fact exists on the system." The problem is to set the software settings well. If an administrator wants to avoid false positives, the setting should not be so strict so there are more false negative messages in the incoming mails. Very often the keywords search is used. The anti-spam software goes through the subject or content and searches for keys, which were defined as "possible spam keywords". This way is not always the best one, because some "innocent" words include some keyword, for example word "specialist" includes word "Cialis" or also the word "analysis" can be confusing for the software.

Spam filters are usually cooperating with blacklists. According to Mgr. Jan Čihák, these lists include IP addresses of well known spammers, proxy servers, open-relays, and sometimes the domains or URL links are included in the content of an email. The administrator can choose which blacklists to use, many of them are online and updated very often. Some software is connected with some kind of local databases of these IP addresses or domains. The company can also create their own whitelists. *"Anti-spam software must have an efficient way to automatically build extensive Whitelists. Whitelists should identify all valid business partners, so that their mail is never flagged as spam. Good anti-spam software should include the facility to automatically create and update these whitelists."* (www.gfi.com) So the company can choose the domains or the IP addresses which are trusted and which will never be delivered to the spam folder.

Another useful technique is using greylisting. *"Each time a given mailbox receives an email from an unknown contact (IP), that mail is rejected with a "try again later"-message (This happens at the SMTP layer and is transparent to the end user)."* (www.greylisting.org) This causes a delay of the emails but on the other hand, the spammers are usually using software, which is not resending emails and not trying again. It is very important to set this feature at the first level of spam filtering; otherwise it is useless.

"Bayesian filtering is based on the principle that most events are dependent and that the probability of an event occurring in the future can be inferred from the previous occurrences of that event... This same technique can be used to classify spam. If some piece of text occurs often in spam but not in legitimate mail, then it would be reasonable to assume that this email is probably spam." (www.gfi.com) This technique could be very useful but the administrator has to create databases at first. The first database includes the words and information (such as IP addresses or domains) from the ham emails (non-spam emails) and the second one takes the same information and words but from the spam emails. Then, each word gets its probability value (how often the word occurs in spam email as opposed to ham). *"When a new mail arrives, it is broken down into words and the most relevant words – i.e., those that are most significant in identifying whether the mail is spam or not – are singled out. From these words, the Bayesian filter calculates the probability of the new message being spam or not."* (www.gfi.com)

4. Spam filtering in practice – spam filtering in different types of companies

4.1 Introduction of the chosen companies

To find out the different approaches of spam filtering in the companies, the author of the thesis at hand arranged meetings with four persons who are responsible for spam filtering in their companies or they are able to explain their way of filtering. These four companies were chosen because of the expected different knowledge of computers and email security and also because of the different number of the administered mailboxes.

The first company is in the media group MAFRA, which is representing many of the smaller but well known companies. There are about 2000 mailboxes administered in the Czech Republic and the administrator also takes care of 1000 mailboxes in Slovakia. The whole IT department is working with the settings of the email addresses and one main administrator takes care of the settings of the anti-spam filter. The users have different training in the spam and Internet issue. The company is administrating many different domains, all controlled by one administrator.

The second company chosen for this thesis is the free email provider seznam.cz. The number of the active users using the services of this company is around 8 million. There is a team controlling the mailboxes and settings in the company. There is also one administrator responsible for anti-spam settings. The users of the email addresses are a sample of all citizens using the email services. The administrators take care of six domains.

The third chosen company is the company creating software, more precisely antivirus software. The number of the administered mailboxes is around 200 within the company. The company has its own domain and there is an administrator looking after the

settings and the security of their mailboxes. The employees are supposed to be well educated on the problem of spam and Internet threats.

In this part, the fourth chosen company is the secondary grammar school in Prague 7, Gymnázium Nad Štolou. The number of mailboxes within the school domain is 70. The company's email was created by a member of the institution and the spam filters are configured by the same person. The knowledge about the spam problem is not so profound in the company, especially because a lot of the employees are older people without ample education about nowadays problems and the threats on the Internet.

4.2 The analysis of spam filtering in MAFRA media group

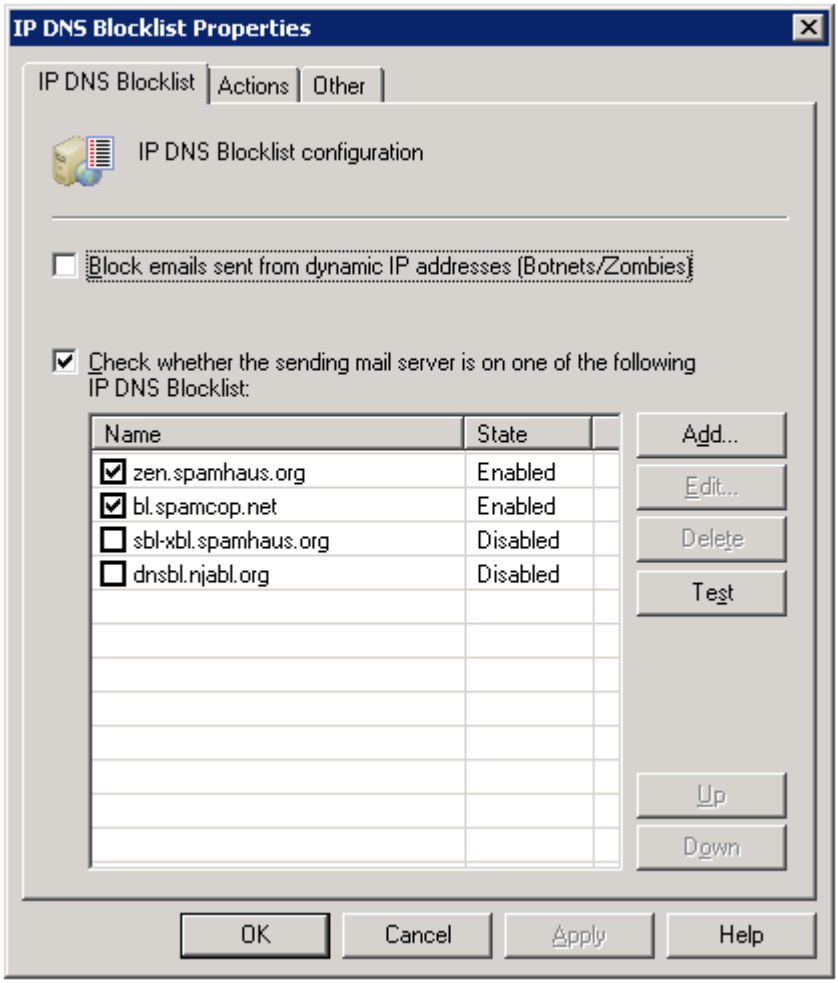
All the information used in this part (about the MAFRA media group) was advised by the ICT Infrastructure Manager of the MAFRA media group, Mgr. Jan Čihák. The meeting was arranged for February 2011 in Prague.

The MAFRA media group is the group that is representing almost 20 independent companies including newspapers, TV stations, radio stations or web servers providing online photo albums. Each of these companies has its own email domain, sometimes more than one. The majority of the companies are located in the Czech Republic, but some of them are located in Slovakia. The whole media group (concerning all the domains used) is receiving approximately one million emails per day. In the past, 98% of these emails were spam emails. Nowadays, thanks to the improved skills in anti-spam filtering setting, the number decreased to 93%. The company is proud that they can say that the number of false positive emails is almost 0.

The company is using two levels of spam filtering. On the first level, the Symantec Brightmail Gateway is used. According to the official website www.symantec.com, this spam filter offers inbound and outbound messaging security, real-time anti-spam and antivirus protection, data loss protection and is able to catch 99% of spam with less than one in a million false positives. Because filter is used as the first level filtering, there is especially hardware communication. There are a lot of different definitions at this level, which can specify the accuracy of the filter. This first level does not allow 98% of spam even to be delivered to the mail box, so the final user has no idea about this huge amount of spam. These spam emails are deleted immediately so there are no statistics about them. The default setting is used at this level.

At the second level, GFI MailEssentials software is used. This software is using two anti-spam engines. This software is used to block phishing emails, to set the blacklists the company wants to use or to tag the spam emails, which came through the first level. In

contrast to the first level, there are a lot of “personal” settings at level two. The company is using some real time blacklists and also chooses which of them to use. From the IP DNS blacklists the **zem.spamhouse.org** and **bl.spamcop.net** are used. The URI DNS blacklists used in MAFRA media group are **multi.surbl.org** and **black.uribl.com**. The blacklists are used not just to block some IP address but there are also lists of the URL links included in the content of the email. So the spam email can be sent from any IP address or email domain but if it includes the URL link listed in the blacklist, the spam is evaluated as spam.



Pic.9: IP DNS Blocklist in MAFRA media group

(Source: print screen from MAFRA media group IT office)

There are also whitelists used. They are usually automated. If the user receives the false positive email and reports it to the IT office, they check it and if it is an IP address without threats, it can be added to the whitelist. The GFI Email Essentials also offers the phishing definitions which can detect almost all the phishing pages.

The keyword detection is disabled in the settings. This is done in this way because with the large amount of emails received every day, there is a high possibility that there will be an increasing number of false positive emails. But in the software, definitions are working with the new types of spam. If some new, dangerous spam appears, the definitions are created exactly for this spam. Usually, the keywords are checked in the subject field, but the software can also go through the content of the email. The strings of the keywords are used very often nowadays.

If the email gets through the first level, it is tested by the second one. If the second level is not sure about the existence of spam, it adds the tag “[spam]” before the subject of the email. Then the user is alerted that there is a possibility the email is spam and it depends on him, if he opens the email or not. Except going through the sender and the subject and content of the email, Bayesian filtering is used. This filtering is in fact “teaching itself” with the help of the user or the administrator. The user can mark the email as spam or non-spam and the software remembers these definitions. So it is possible to say that the user is able to create his whitelist. It is a good solution in a company with many users, because some users want to receive some offers but others mark it as spam. The GFI software is not just controlling the received emails but also the out-going emails. If an email is sent from some of the users to non-existing email addresses, the administrator gets to know about this email and he can think about the possibility that the user is sending spam.

The GFI Email Essentials offers also the usage of greylisting. It means that the message is rejected as Error 400 (Bad Request). The emails are usually resent after 5 minutes, then after 10 minutes and so on. The idea is that the spammer is not sending the email again because the amount of emails sent by him is so large that he has no interest to

resend it to these email accounts. But the greylisting also causes delivery delays. The reason why the MAFRA media group is not using the greylists at the second level is obvious. It could be used just at level one, because otherwise the emails will be lost somewhere between the levels.

The costs for running these spam filters are connected to the number of email boxes. In MAFRA case, it is 1000 CZK per mail box and year. These costs are customized and they depend on the contract between the company and the software company. Every year, there is a new selection procedure.

The email boxes are of course accessible via webmail. Thanks to the Microsoft Exchange and the Outlook web applications, they are also accessible from any other computer. Also the access via cell phones is possible, using the support of ActiveSync. The members of the IT office have also access via their service notebooks thanks to the imap4 protocol. And finally, just Mgr. Jan Čihák has access to all the servers, data and emails via Outlook Anywhere, Microsoft Exchange 2003 SP2.

Mgr. Jan Čihák provided also the statistics including the domains from which the largest amount of emails come from. The first place belongs to „unknown“ domains, but the first concrete domain is **mmnc.net** (2,113,553 processed, 98% detected, 0% of spam), followed by **gmail.com** (7,137 processed, 7% detected, 2% of spam), **seznam.cz** (10,157, less then 1% detected, less then 1% of spam) and **mafra.cz** (1,383 processed, 7% detected, 7% of spam). The domains sending just 100% of spam are for example **job.com**, **computrabajo.com**, **freelot.nl**, **netscape.net**, **luckymail.com**, **dieneme.com**, **humanic.com** or **yahoo.co.jp**.

4.3 The analysis of spam filtering in Seznam.cz

All the information used for the part about the company Seznam.cz was provided by Mr. Miroslav Chocholouš. The meeting was arranged for February 2011 in Prague.

Seznam.cz is one of the largest email providers in Czech Republic. The providing of email boxes is for free, that is also the reason why Seznam.cz has 8 millions of active mail boxes (active mail box is the mailbox which the user has visited at least once a month in at least 3 contiguous months). Seznam.cz offers 6 possible domains, the domains usually belongs to some servers belonging to Seznam.cz (for example firmy.cz or spoluzaci.cz). Seznam.cz receives an unbelievable number of emails per day. 98, 5% of the received emails are spam. Unfortunately, it is impossible to find out the percentage of the false positive emails. The number of users is so large, that there is no possibility to measure it somehow. Especially because a lot of users are not well educated on the spam problem so they are probably not able to go through the spam and find false positive emails. Mr. Miroslav Chocholouš said that there are just approximately 8 emails per week where the user is complaining or alerting that he had false positive message in his spam folder. These emails are solved by the main spam administrator, who also takes care of all the settings.

Seznam.cz is also using two levels of spam filtering. The first level was created by the IT office in Seznam.cz. They created their own spam filter to be able to optimize all the settings and features they need. It is also almost impossible to pay for some software if they are providing such a big number of email boxes. This software is written in Perl programming language. Some competitors say about Seznam.cz that they are just “rewriting what was already written”. The first level is able to catch 97% of spam and reject it. They are using blacklists of IP addresses (13 blacklists) like uribl.com or the SORBS (Spam and Open Relay Blocking System) which offers up-to-date databases and is possible to go through all the servers from which email were received to find out if some proxy server or open-relay server was used. Seznam.cz is also using whitelists, used for the

Czech Republic and Slovakia. At the first level also the size of an email is controlled, because the average size of spam email is usually not larger than 5 Kb.

Seznam.cz uses greylisting at the first level. There are 6 greylists running in the company, 3 at each server room.

At the second level, SpamAssassin is used at Seznam.cz. This anti-spam is open-source software. It can run a variety of local and network tests in order to identify spam. At this level, the IP addresses and envelope sender addresses are checked at first. If the email in not marked as a spam, the receivers addresses are checked. This is done because the spammers can try to create a fake email address on the existing email domain. So if at the same time Seznam.cz receives a lot of emails for the nonexistent addresses, it shows that these emails are probably spammed. SpamAssassin also goes through the domains, URL links in the email content and also works with key words and regular expressions for better searching and differentiating spam and non-spam subjects and content. The second level is able to reject 97% of emails which go through the first level. So 7% of these emails are analyzed and one third of these emails the user receives are ham (non-spam emails). Daily Seznam.cz delivers 20 million of ham emails. The rest are marked as spam and are shifted to the folder called “Spam a viry” from where it can be removed.

The email box is, of course, reachable from any computer. If the user wants to connect via cell phone, it is also possible with the usage of encryption.

4.4 Analysis of spam filtering in Avast!

All the information used in the part about the Avast! spam filtering are from Mr. Petr Chocholouš, the meeting was arranged for February 2011, in Prague.

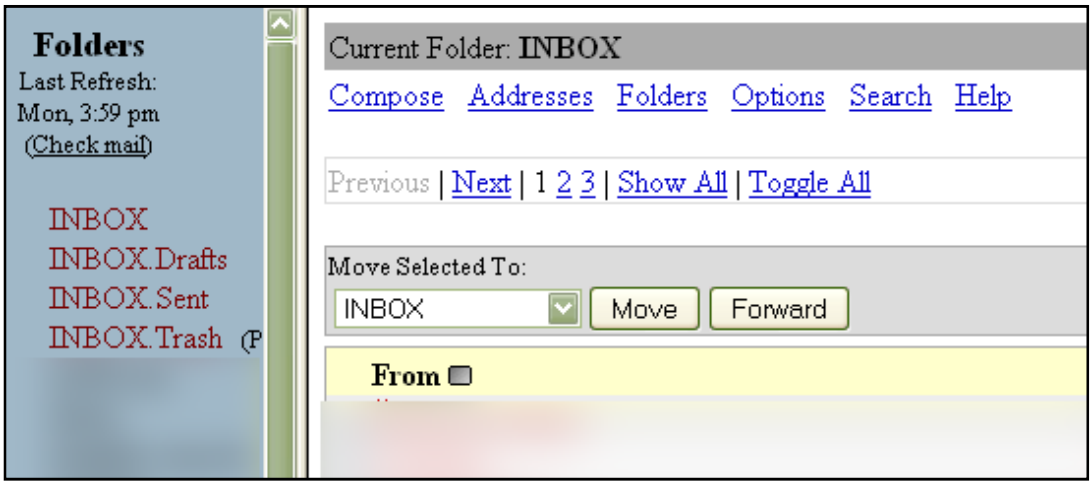
Avast! is the company which provides free antivirus solutions. The company has its own domain and there are about 200 email boxes in the company. Like in all the mentioned companies, the percentage of spam emails is about 98%. Daily Avast! receives more than one million email messages. The false positive messages are hard to measure because a lot of companies or users are writing just once so it is impossible to go through all the spam emails and check for the false positives.

The company has also two levels of spam filtering. In contrast to Seznam.cz, they are using SpamAssassin at the first level. The spam administrators are not using the default settings but they work with their own. At the first level, they are using the blacklists (almost all the possible blacklists, which can be a little contra productive). The SpamAssassin is also going through the spam score of the email and evaluates the possibility of spam email. It is controlling the header and the body of all the emails. Another way is that the unsubscribed domains are directly deleted. The greylisting is also disabled in the Avast!. The emails marked as spam are saved into the hidden folder so the users (employees) have no possibility to go through these emails. That is also another reason why it is impossible to find false positive messages. The only way is that the user is really sure he should receive some important email and this email has not been delivered yet. In this case, there is the possibility that the administrator of the spam filter can try to find this lost email in the hidden folder. But this folder is automatically deleted (every month).

At the second level of spam filtering, the Avast! company is using their own software. Because the company is providing especially the antivirus software with the anti spam included, the emails are also very carefully scanned in case they include some virus,

worm or Trojan horse. At the second level, the strategy of the keywords is used. The setting is, obviously, created by the administrators. The “automatic learning” is also used at this level; the principle is the same as the Bayesian filtering.

The users of the company can use web mail to access their email boxes, but in the whole company almost only other email clients are used. The messages are also accessible via cell phones.



Pic.10: Avast! web mail – no special folder for spam emails

(Source: print screen from Avast! company)

4.5 Analysis of the spam filtering in a secondary grammar school

All the information used in the part about the Secondary grammar school Nad Štolou' spam filtering are from Mr. David Fridrich, the meeting was arranged for March 2011 in Prague.

Secondary grammar school Nad Štolou is the institution giving education to more than 800 students aged 12 to 19 years old. About 70 email boxes have been created for the employees and are administered by one email and spam administrator in this institution. All the employees use one domain. The percentage of rejected spam emails that are never delivered to the mail boxes is approximately 99%, the occurrence of false positives is rare. Daily, the secondary grammar school receives around 700 spam emails. There is a surprisingly high percentage of spam emails. One of the reasons is that the users are not very well educated in IT security and they use their company emails for the registrations on many different web pages. The second and probably the main reason is that the web pages of the secondary grammar school Nad Štolou used to be "written" in a very unsecured way so it was really simple for the spambots to find out all the email addresses.

The institution uses a lot of software and programs to secure its email. One of the programs is the Postfix, which works like the well known Sendmail program. Another one is the amavistd-new interface, which helps with checking the content, scanning for viruses or cooperating with the SpamAssassin – also used in the institution. All this software is free.

At the first level, SpamAssassin is used. The administrator does not use IP blacklists or the blacklists of URL links often found in content. It is also not necessary to use greylisting in this institution. Only the keywords are checked. The administrator has chosen these keywords thanks to the approximately 40 spam emails. So it is the approach similar to the Bayesian filtering.

It is possible to say that the non-spam emails received by the employees of the school are usually of personal nature. Only a few emails are received from companies or other institutions. So the whitelists are created only if the employee asks for it. For example if the user wants to send some materials from his personal email to the company email, there is the problem that if he does not fill out the subject, the message should be considered as a spam. So he can ask for adding this to the whitelist.

The administrator of the email also takes care of the false negatives. But the user has to report some expected email that was not delivered. All the spam emails that go through the spam filters are sent to the special email address spam@gymstola.cz. Only the administrator has access to this email box and the spam emails are erased automatically every week.

The company email is accessible from other computers via webmail and also via cell phones. The administrator also has access to all email boxes and the server via VPN (Virtual Private Network).

The last report shows, that from 700 spam emails received daily none of them was marked wrong (false negative or false positive). Only two spam emails went through the spam filters into the received emails.

4.6 Comparison of the companies

The research shows some interesting facts about the spam filtering in the chosen companies. The first thing is that the companies do not obviously use just one level of spam filtering nowadays. The possibility of more levels allows the administrators to combine many approaches and choose the setting which fits the company security needs the best.

Each company uses different spam filtering methods and different ways how to store the spam emails (table 1). But in the end, the result of the received spam emails, the number of false positive emails and the percentage of rejected spam emails is almost the same in all the companies. So it is obvious that every company just needs a really good administrator who is able to create spam filtering settings suited to the company needs. It is not necessary to use all the spam filtering methods to protect the email boxes.

	MAFRA	**Seznam.cz**	**Avast!**	**Nad Štolou**
Number of email boxes	2000 + 1000	8 000 000	200	70
Filtering levels	2	2	2	3
Software used	Symantec + GFI	Own SW + SpamAssassin	SpamAssassin + own SW	Postfix + amavistd-new+ SpamAssassin
Paid SW	✓	✗	✗	✗
Blacklists	✓	✓	✓	✗
Keywords	✗	✓	✗	✓
Greylisting	✗	✓	✓	✗
Bayesian filtering	✓	✗	✓	✗
Spam storage	tagging	“Spam a viry”	hidden folder	email address

Table 1: Comparison of the spam filtering in the chosen companies

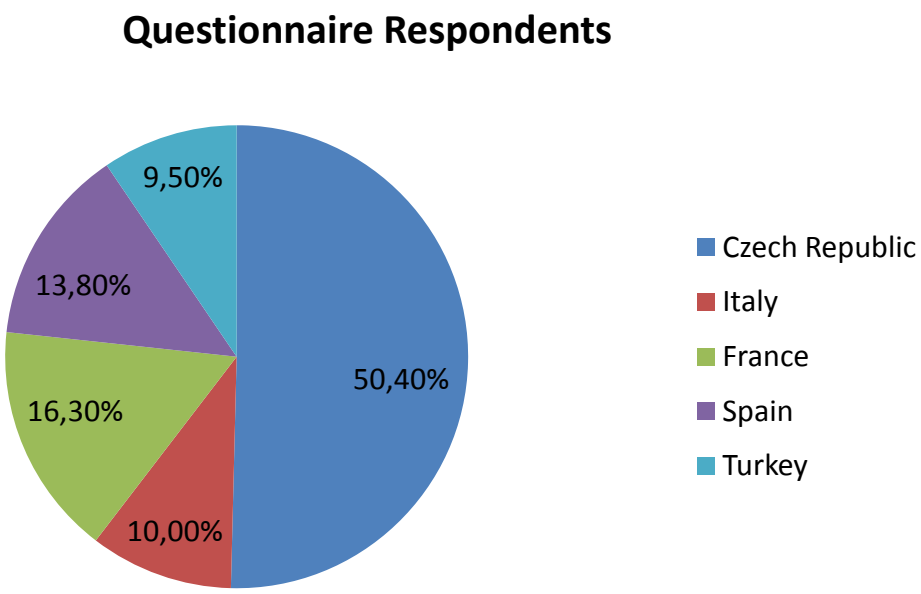
(Source: own research)

It seems that there is no need to pay the expensive licences or software programs. There are many free software programs and spam filters which are used by the companies that really need high security of their email boxes. So it looks like there is no big difference between the qualities of free software and paid software. However, MAFRA media group is paying a lot of money for the licences. The reason is that the large commercial companies (especially media companies) cannot afford the large number of false positive emails. They also need some feedback from the creators because of possible mistakes or wrong filter settings. MAFRA media group has tested a lot of paid and free software and the administrator decided to use the paid software because of the results of testing the software directly in the company.

5. Spam content and amount in chosen countries

It is possible to find different statistics about spam in general. The whole amount of spam all over the world, the most usual content of spam, the percentage of spam emails in all the emails received or sent. The questions is, are there any differences between the spam content depending on countries, private or company emails?

The questionnaire (Supplement 5) was sent to the citizens of five different countries: Czech Republic, Italy, Spain, France and Turkey. From 232 answered questionnaires, 103 respondents are women, 129 are men. 50, 4 % of all the respondents are from the Czech Republic, the rest are residents of other countries.



Pic.11: Respondents divided by the countries of origin

(Source: own research)

Less than 10 spam emails per week reach 57% of the Czech users, 45% of French users, 52% of Italian users, 34% of Spanish users and 32% of Turkish users. On the other

hand, more than 20 spam emails per week day are received by 21% of Czech users, 21% of French users, 13% of Italian users, 22% of Spanish users and 41% of Turkish users. It looks like the Turkish and Italian users have high numbers of spam emails received.

Depending on the email account providers, the result is shown in the supplement 1. According to the questionnaires, the first place in the smallest amount of spam emails belongs to **seznam.cz**, followed by **gmail.com** and **yahoo.com**.

The number of received false negative and false positive emails also depends on the email account provider. The amount of these emails broken down to the provider is shown in supplement 2 and supplement 3.

The questionnaires' answers from the respondents show that if the user has company email, usually the amount of the received spam email is lower than when using a public one. In some countries, the amount is approximately the same; in some countries it is never higher (supplement 4).

The content of spam emails in these five countries seems to be interesting. In general, one of the most usual content is offers on Viagra and other medication and pills. These medications should improve the sex ability or are available only with prescription. Another "popular" content are the offers from online casinos and emails which tell the user he has won some great amount of money in an online casino he has never visited.

There are also some contents occurring just in some countries. For example emails from the companies that offer the enlargement of some body parts are not received by residents of Italy and Spain. Another example is that only Czech users receive the spam emails concerning charity for Africa. The rest of the examples are shown on the table 2. These results are deduced from the limited number of answers, so in general, there is a possibility that they can be misleading.

CONTENT	Czech Republic	France	Italy	Spain	Turkey
Viagra	**50,5%**	**23,5%**	17%	18%	23%
Drugs	**18%**	10, 5%	**26%**	19%	**32%**
Body parts' enlargements	11%	2,5%	/	/	18%
Dating	13%	13%	8,5%	/	23%
Casino winnings	17%	**29%**	**22%**	**28%**	13,5%
Job offers	14,5%	2,5%	/	12,5%	/
Phishing	9,5%	/	/	6%	/
Product's offers	**17%**	**34%**	**52%**	**40,5%**	**32%**
Replica watches	14,5%	13%	/	/	/
Travelling	2,5%	13%	/	19%	4,5%
Charity (Africa)	3%	/	/	/	/
Sexual content	12%	16%	/	3%	**41%**
Loan offers	4%	/	8%	/	/
Investments	4%	/	8,5%	3%	4%
Insurance	/	8%	4%	3%	/

Table 2: The content of spam email received in chosen countries

(Source: own questionnaires)

6. Conclusions

Spam is everywhere. In all the chosen companies, most of the emails trying to get to the user's email box are spam emails. It does not matter, if the user is well educated in IT security or if it is the usual Internet user. The spam is sent even if the user does not use his email for the registration or does not visit some special web pages.

It also does not matter how many spam emails are sent to the company. It seems that all the spam filters can catch around 98, 5% of spam emails regardless of the number of email boxes in the company domain.

Each of the companies uses different spam filtering tactics. First of all, they are using different software filters from different providers, free or paid. Even if the same software is used, it does not mean that it is used on the same level of filtering. Another thing is that the spam filters offer many possibilities in their settings. So each company is able to choose which techniques it wants to use. Companies also differ in their approaches to spam filtering. They have different ways how to deal with spam messages received in the incoming emails. But in the end, all the companies are successful in protecting email boxes against spam. So the results are almost the same even if the companies use different strategies. It is caused by the setting that are possible to be adjusted exactly for company needs.

The results of the questionnaires support the idea that it really does not matter where the user is from – the problem of spam is everywhere. There are some basic facts confirmed by the questionnaire. Italy is ranked 3rd in terms of being the origin of spam emails. Also the amount of spam received in this country is the highest from the five chosen countries. The content of spam is approximately the same in all the chosen countries. There are some differences, though. Respondents from some countries do not receive some content at all and some countries seem to receive specific offers. But there is the chance it is caused only by the low variability of the respondents.

There is also a high percentage of false positive and false negative emails received by users using the well known email account providers. The reason is that these providers work with millions of email addresses so it is impossible to create the filtering so accurate that it would eliminate these messages. Each user receives many spam emails, some user requires the offers from some companies but for another user it is marked as a spam. The large providers of email accounts also use the keywords for their filter's setting. The users using these private emails are usually not well educated on the spam problem so they are sending the emails without the subject filled out, or with the subject that can be easily commute to the spam email. It is also impossible to create white lists for each user, so the percentage of false negative emails is higher than in the company email, where the whitelists are more simply to set.

Anti spam software is developing all the time. Many of the programmers and experts work hard to protect the users. But also the spammers' tactics and technologies are improving. So it seems that the future of spam filtering will not be so different from the current situation. Maybe, there is a chance that there will be some perfect spam filter covering all the possibilities of "spam attack" but there is still the certainty that the spammers have the same knowledge and possibilities to go through these filters. So it is possible that the "war" between the spam filters' providers and the spammers will never end.

7. Bibliography and sources

ADÁMEK, Martin. Spam:jak nepřivolávat, nepřijímat a nerozesílat nevyžádanou poštu. 1st ed. Praha: Grada Publishing, a.s., 2009. 168 pages. ISBN 978-80-247-2638-0

COSTALES, Brian, FLYNT, Marcia. Sendmail Milters: A Guide for Fighting Spam. 1st ed. Massachusetts: Pearson Education, Inc., 2005. 329 pages. ISBN 0-321-21333-5

FEINSTEIN, Ken. How to Do Everything to Fight Spam, Viruses, Pop-Ups, and Spyware. Emeryville: McGraw-Hill Companies, 2004. 234 pages. ISBN 0-07-225655-9

GOODMAN, Danny. Spam Wars. 1st ed. New York: SelectBooks, Inc., 2004. 330 pages. ISBN 1-59079-063-4

LEVINE, John R., LEVINE YOUNG, Margaret, EVERETT-CHURCH, Ray. Fighting Spam for Dummies. Indianapolis: Wiley Publishing, Inc., 2004. 224 pages. ISBN 0-7645-5965-6

CORRONS, Luis. *Http://pandalabs.pandasecurity.com* [online]. 2011 [cit. 2011-01-26]. PandaLabs Annual Report 2010. Accessible from WWW: <http://pandalabs.pandasecurity.com/pandalabs-annual-report-2010/>.

HÁK, Igor. *Moderní počítačové viry* [online]. 3rd edition. Czech Republic, 2005 [cit. 2011-03-15]. Accessible from WWW: <http://viry.cz/viry.cz/kniha/kniha.pdf>.

MAREŠ, David. *Http://www.epravo.cz* [online]. 2010 [cit. 2011-02-10]. Internerová reklama a nevyžádaná obchodní sdělení šířená elektronickými prostředky. Accessible from WWW: <http://www.epravo.cz/top/clanky/internerova-reklama-a-nevyzadana-obchodni-sdeleni-sirena-elektronickymi-prostredky-63234.html>.

SORKIN, David E. *Spamlaws.com* [online]. 2009 [cit. 2011-01-17 – 2011-02-15]. Accessible from WWW: < http://www.spamlaws.com/spam.html>.

SORKIN, David E. *Spamlaws.com* [online]. 2009 [cit. 2011-01-17 – 2011-02-15]. Accessible from WWW: < http://www.spamlaws.com/antivirus.html>.

SORKIN, David E. *Spamlaws.com* [online]. 2009 [cit. 2011-01-17 – 2011-02-15]. Accessible from WWW: < http://www.spamlaws.com/spyware.html>.

SORKIN, David E. *Spamlaws.com* [online]. 2009 [cit. 2011-01-17 – 2011-02-15]. Accessible from WWW: < http://www.spamlaws.com/adware.html>.

SORKIN, David E. *Spamlaws.com* [online]. 2009 [cit. 2011-01-17 – 2011-02-15]. Accessible from WWW: < http://www.spamlaws.com/computer-virus.html>.

Http://ec.europa.eu [online]. 2009 [cit. 2011-02-05]. Twelfth Annual Report of the Article 29 Working Party on Data Protection. Accessible from WWW: <http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2009/12th_annual_report_en.pdf >.

Http://hyders.com [online]. 2010 [cit. 2011-01-21]. RootKit Stealth Attack. Accessible from WWW: <http://hyders.com/?p=464>.

Http://searchcio-midmarket.techtarget.com [online]. 2001 [cit. 2011-01-19]. Adware. Accessible from WWW: <http://searchcio-midmarket.techtarget.com/definition/adware>.

Http://support.gfi.com [online]. 2011 [cit. 2011-02-10]. GFI MailEssentials `Bayesian' anti-spam filter. Accessible from WWW: <http://support.gfi.com/manuals/en/me9/me9manual-1-05.html>.

Http://support.twitter.com : Spam and Abuse [online]. 2011 [cit. 2011-01-28]. The Twitter Rules. Accessible from WWW: <http://support.twitter.com/entries/18311>.

Http://www.computershack.info [online]. 2010 [cit. 2011-01-19]. What Is Malware. Accessible from WWW: <http://www.computershack.info/WHAT-IS-MALWARE-SPYWARE.html>.

Http://www.crimes-of-persuasion.com [online]. 2010 [cit. 2011-01-28]. Nigerian Email Scams. Accessible from WWW: <http://www.crimes-of-persuasion.com/Crimes/Business/nigerian.htm>.

Http://www.facebook.com [online]. 2011 [cit. 2011-01-28]. Updates in Facebook’s Fight Against Spam and Spammers. Accessible from WWW: <http://www.facebook.com/note.php?note_id=442722120765>.

Http://www.greylisting.org [online]. 2011 [cit. 2011-02-10]. Greylisting. Accessible from WWW: <http://www.greylisting.org>.

Http://www.hky.com [online]. 2010 [cit. 2011-01-19]. Viruses and other malware (malicious software). Accessible from WWW: <http://www.hky.com/virus_prevention.php>.

Http://www.m86security.com [online]. 2011 [cit. 2011-02-05]. Spam Statistics. Accessible from WWW: <http://www.m86security.com/labs/spam_statistics.asp>.

Http://www.pcguide.com [online]. 2001 [cit. 2011-02-05]. False Positives and False Negatives. Accessible from WWW: <http://www.pcguide.com/care/data/virus/scanFalse-c.html>.

Http://www.pcmag.com [online]. 2011 [cit. 2011-01-19]. Adware definition.

Accessible from WWW:

<http://www.pcmag.com/encyclopedia_term/0,2542,t=adware&i=37577,00.asp>.

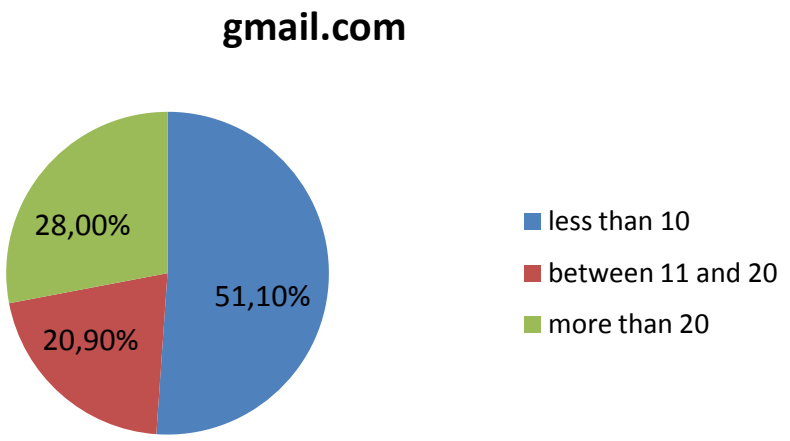
Http://www.securelist.com [online]. 2011 [cit. 2011-02-04]. Spam report: January 2011. Accessible from WWW: <http://www.securelist.com>.

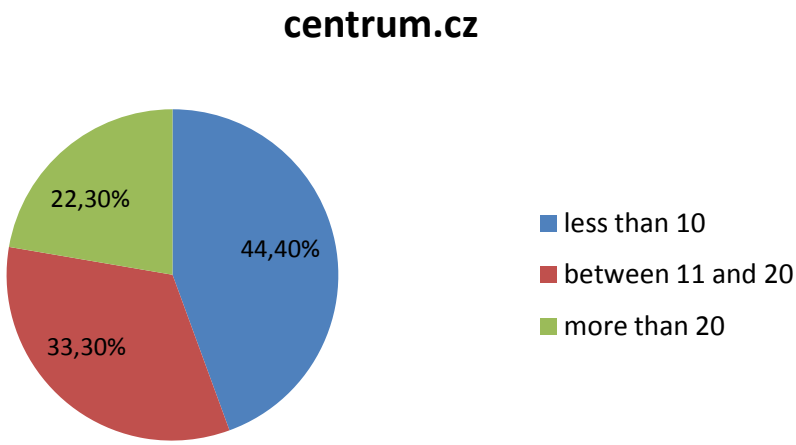
Http://www.spectorsoft.com [online]. 2011 [cit. 2011-01-21]. EBlaster. Accessible from WWW: <http://www.spectorsoft.com/products/eblaster_windows/index.asp?affil=6>.

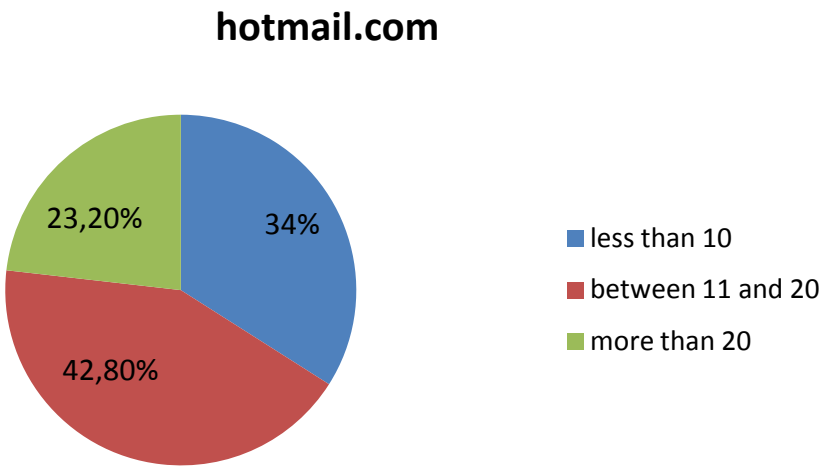
Http://www.uoou.cz [online]. 2004 [cit. 2011-02-03]. Zákon č. 480/2004 Sb., o některých službách informační společnosti. Accessible from WWW: <http://www.uoou.cz/uoou.aspx?menu=23&submenu=25>.

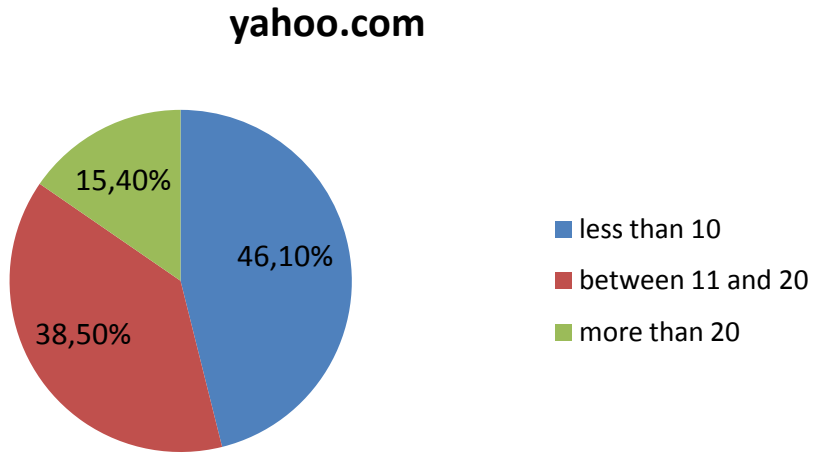
8. Supplements

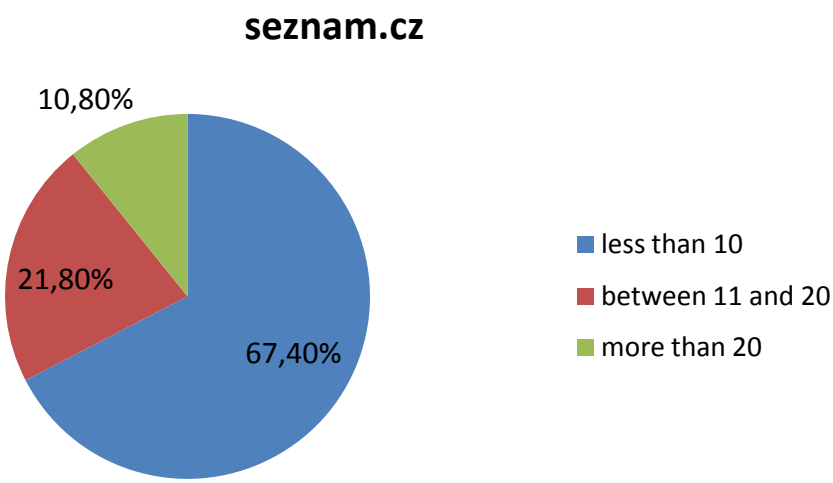
8.1 Amount of the spam emails received divided by the email providers

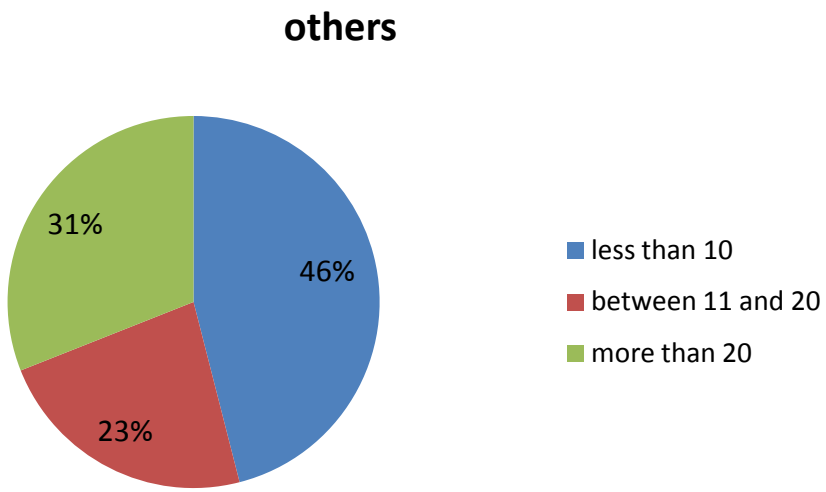





hotmail.com
23,20%
34%
42,80%
less than 10
between 11 and 20
more than 20


yahoo.com
15,40%
46,10%
38,50%
less than 10
between 11 and 20
more than 20


seznam.cz
10,80%
21,80%
67,40%
less than 10
between 11 and 20
more than 20



Source: own questionnaires

8.2 False negative emails depending on the email provider

FALSE NEGATIVES	never	1-5times/year	more often
centrum.cz	39%	28%	33%
gmail.com	28%	59%	13%
hotmail.com	19,5%	62, 5%	18%
seznam.cz	28%	52%	20%
yahoo.com	15%	46%	39%
others	38,5%	38,5%	23%

Source: own questionnaires

8.3 False positive emails depending on the email provider

FALSE POSITIVES	never	1-10times/year	more often
centrum.cz	11%	11%	78%
gmail.com	30%	44%	26%
hotmail.com	22%	36%	42%
seznam.cz	21%	58%	21%
yahoo.com	23%	54%	23%
others	8%	54%	38%

Source: own questionnaires

8.4 Comparison of the amount of spam received in company email and private email (divided by the countries)

COMPANY EMAILS	lower	same	higher
Czech Republic	62%	17%	21%
France	85%	5%	10%
Italy	43%	57%	0%
Spain	78%	11%	11%
Turkey	100%	0%	0%

Source: own questionnaires

8.5 Questionnaire

Sex: *

- [ ] Male
- [ ] Female

Country: *

- [ ] Czech Republic
- [ ] France
- [ ] Italy
- [ ] Spain
- [ ] Turkey

Which email provider do you use the most? (for your private email) *

- [ ] gmail
- [ ] hotmail
- [ ] yahoo
- [ ] seznam
- [ ] centrum
- [ ] Jiné:

How many spam emails do you recieve weekly? (approximately) *

- [ ] less than 10
- [ ] 11-25
- [ ] more than 25

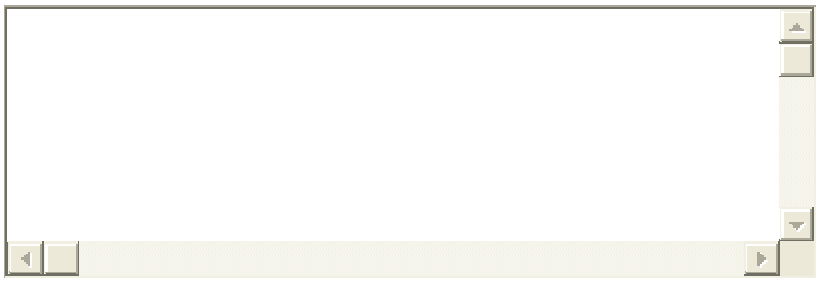
How many times was the official (non-spam) email marked as a spam? *

- [ ] never
- [ ] 1 - 5 times / year
- [ ] more often

How many times was the spam email marked like a non-spam and appears in your Income messages? *

- [ ] never
- [ ] 1 - 10 times / year
- [ ] more often

What is the usual content of the spam emails you recieve?Please, write 3 most usual. *



Do you have some official company email? *

- yes
- no

If yes, the amount of spam in your comapny email is

- lower than in the private email
- approximately same as in the private email
- higher than in the private email