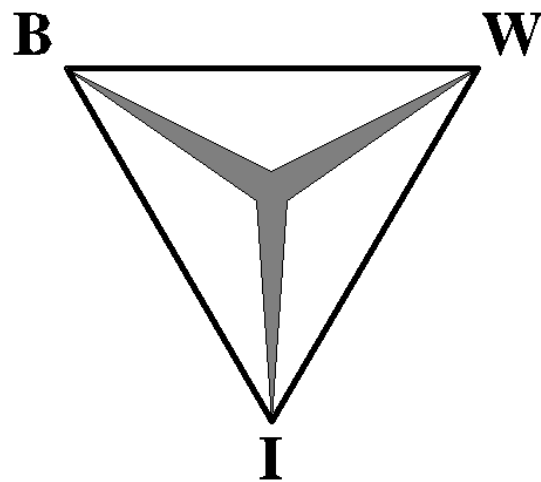


SPAM FILTERS

An overview



Spamfilters, an overview
Werkstuk Bedrijfskunde & Informatica
Frank Kuiper
Student ID 0549789
vrije Universiteit *amsterdam*
Supervisor: Dr. Wojtek Kowalczyk
June 2004

Preface

During the final years of the study Business Mathematics and Computer Science (Bedrijfskunde en Informatica, BWI or BMI in english), it is required to write a so-called *werkstuk*. In this paper, the student must individually asses a specific problem, in which he integrates the three aspects of BMI: Business, Mathematics and Computer Science.

Initially, I was going to write a paper about automatic e-mail response, focusing on using text mining to identify the subject of the e-mails. After talking to Wojtek Kowalczyk about the scope and relevance of this subject, I changed this to what could be considered an opposite but very much related topic: e-mail spam. I decided to write the paper on this subject because of the relevance - nowadays many organizations and private persons deal with spam every day, at great cost in time, money, and annoyance. Also, there have been some new developments in the last few months, such as a “spam law” in the Netherlands and initiatives like Sender ID.

In this paper I will give an overview of the problem of spam and some often used solutions and methods used for recognizing and dealing with spam. As it is intended for a broad audience, methods and algorithms will only be presented briefly.

I want to thank the people that were kind enough to help me while writing this paper: Eric, Hans, Kees, Raymond, Robbert, Roel and Wojtek.

Executive summary

Internet spam has become a large problem for most e-mail users. Ostermann research stated that in 2003, more than 80% of the IT decision makers consider spam to be a “serious” or “very serious” problem in their organizations and that spam costs \$1400 per year per e-mail user in lost productivity when no spam filtering would be used. The American ISP AOL is reported to delete more than 2 billion messages per day and the organization of Dutch Internet Service Providers, NLIP, estimated that in 2002 a third of all e-mail in the Netherlands was spam. These numbers are still rising. Clearly, stopping or filtering spam would benefit organizations and private persons. This would result in a reduction of IT costs and/or time wasted by the employees. These days, fighting spam is a necessity. This fighting can and should be done using different filtering techniques and by increasing security on servers and clients.

In this paper I will give an overview of the problems with spam in organizations and review some of the options available to combat spam. The central question here is:

What are the options for organizations and private persons to minimize the costs, incurred because of spam?

Many applications exist to filter spam, these filters can be used at the central e-mail servers or the desktop client. Filtering could also be outsourced to third parties such as an Internet Service Provider. Applications use different methods to filter spam, based on the origin of the e-mails or its contents. As most methods can be used together, the best results are obtained by combining the results of several methods and by combining server side and client side filters.

Contents

Preface	iii
Executive summary	iv
1 Introduction	1
1.1 What is spam?	2
2 Organizational aspects	5
2.1 Reasons for spamming	5
2.1.1 Spoofing of SMTP headers	6
2.1.2 Open relays	7
2.1.3 Open proxies	8
2.1.4 Trojan horses	8
2.2 Opt-in and opt-out	9
2.3 Why is spam a problem?	10
2.4 Legislation	12
2.4.1 Habeas Haiku: copyright against spam?	13
2.5 Practice in Dutch organizations	13
2.5.1 Vrije Universiteit	14
2.5.2 Multikabel	14
2.5.3 Tweakers.net	15
2.5.4 Argitek	16

3	Fighting spam	17
3.1	Network	18
3.2	Applications	20
3.2.1	Server side filters	21
3.2.2	Clientside filters	22
3.3	Hardware or outsourcing	23
4	Filtering methods	24
4.1	Origin	24
4.1.1	Blacklist	24
4.1.2	Vipul's Razor	25
4.1.3	Whitelist	25
4.1.4	Challenge/Response	26
4.2	Contents	27
4.2.1	Preprocessing	27
4.2.2	URI Blocklist	28
4.2.3	Phrase matching	28
4.2.4	Text mining	28
4.3	Traffic Volume	31
4.4	e-Mail minefields	31
4.5	Scoring	31
5	Conclusion and recommendations	33
A	Telecommunicatiewet	36

Chapter 1

Introduction

Internet spam has been around for 10 or more than 25 years, depending on the definition. On May 3, 1978¹, the computer company Digital Equipment Corporation sent all Arpanet users on the US west coast a message about an open day where the new range of machines would be shown. On April 12, 1994, Laurence Canter used a script that flooded online message boards with an advertisement for the legal services of his law firm, Canter & Siegel. “Send coconuts and cans of Spam to Cantor & Co.,” one Usenet reader wrote, thinking of Monty Python’s “Spam Sketch” where the other items on a menu were buried beneath spam; spam, spam, spam, egg and spam; spam, spam, spam, spam, spam, baked beans, spam, spam, spam etc. Thus the term spam was born, much to the chagrin of the makers of this lunch meat², the Hormel company. The term Spam (Spiced Ham) has become synonymous to Unsolicited Bulk E-mail or Unsolicited Commercial E-mail. This is also the reason solicited / wanted e-mails are referred to as (ordinary) Ham.

Spam has become a large problem for all users of e-mail; businesses, non-profit organizations and private persons alike. A survey organized by the European Commission states that in 2002, 55% of the Dutch internet users “had a problem with spam”, an increase of 11 percent-points compared to 2001. Ostermann research stated that in 2003, more than 80% of the IT decision makers consider spam to be a “serious” or “very serious” problem in their organizations. The American ISP AOL is reported to delete more than 2 billion messages per day and the organization of Dutch Internet Service Providers, NLIP, estimated that in 2002 a third of all e-mail in the Netherlands was spam and the numbers are still rising, as shown in figure 5.1.

¹<http://www.templetons.com/brad/spam/spam25.html>

²<http://www.spam.com/es.htm>

In this paper, I will give an overview of the problems with spam in organizations and review some of the options available to combat spam.

The central question here is:

What are the options for organizations and private persons to minimize the costs, incurred because of spam?

To answer this, I will search for the answers to some more specific questions:

- Why is spam a problem in organizations?
- What is the practice in some Dutch organizations?
- Which methods for prevention of spam are available?

I will do a literature study and speak to people that are responsible for the e-mail servers in a few organizations with which I have personal contacts. Obviously, this is not meant to be a representation of the way spam is dealt with throughout the Netherlands as the number of interviewed people is too low. It used as a way to get an idea of the way spam is dealt with within different kinds of organizations.

1.1 What is spam?

First of all, it is necessary to define what is considered to be spam. Mail-abuse.org uses the following definition:

An electronic message is “spam” IF: (1) the recipient’s personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; AND (2) the recipient has not verifiable granted deliberate, explicit, and still-revocable permission for it to be sent; AND (3) the transmission and reception of the message appears to the recipient to give a disproportionate benefit to the sender.

This definition is rather broad. It does not specify the methods used or the contents of the message. There are various ways for classifying spam. Delivery method, sender, contents and size of the mailing are such classifiers.

U.S. based Pew Internet & American Life Project did a survey among 624 Americans (648 for items 6-13), where they asked among other thing what people considered spam, see table 1.1.

Although the number of surveyed people is not high, it shows that the definition of spam is not very clear. 92% of the interviewed people considered Unsolicited

CHAPTER 1. INTRODUCTION

What US e-mail users consider spam

Sender or subject matter	percentage
Unsolicited commercial email (UCE) from a sender you don't know	92%
UCE from a political or advocacy group	74%
UCE from a non-profit or charity	65%
UCE from a sender with whom you've done business	32%
UCE from a sender you have given permission to contact you	11%
UCE containing Adult content	92%
UCE with investment deals, financial offers, moneymaking proposals	89%
UCE with product or service offers	81%
UCE with software offers	78%
UCE with health, beauty, or medical offers	78%
Unsolicited email with political messages	76%
Unsolicited email with religious information	76%
A personal or professional message from one you don't know	74%

Table 1.1: E-mailers' definition of spam depends on the sender and the subject matter of the message. Source: Pew Internet & American Life Project June 2003 Survey. Margin of error is $\pm 4.2\%$

Commercial E-mail (UCE) from a sender they didn't know spam, only 65% thought this when the mail was sent by a non-profit or charity organization and 11% considers commercial mail from a sender they have given permission to contact them as spam.

Different forms of spam can be sent via e-mail, usenet groups³, Short Message Service (SMS), Instant Messenger networks such as ICQ or MSN Messenger, etc. Others might also consider pop ups on websites to be spam. In this paper, I will only look at e-mail spam because it is the most-seen form for most persons and organizations and because the other methods use very different techniques and / or have very different requirements, such as real-time scanning.

Spam can also be classified according to the sender: business partners, friends and acquaintances or e-mail lists one has subscribed to are usually regarded as legitimate senders, the boundaries are more vague when receiving messages from strangers or organizations one has had no dealings with before.

Another way to consider a mailing to be spam is according to the contents. Common types are commercial spam ("get access to porn sites", "buy Viagra!"), scams ("work from home", "give me \$1000 to get 1 million out of Zaire") and political, religious or ideological messages. Others are chain mails forwarded by acquaintances. Recently, arguably racist e-mails have been sent to many Dutch e-mail addresses with subjects such as "Bankrott des Gesundheitswesens durch Auslaender!" and "Das kann unmoglich sein"

³Usenet, also known as News, is a large amount of text-based discussion groups, usually accessed through the internet

So-called “phishing” and viruses are somewhat nastier. Password phishing, is getting sensitive information such as passwords and other personal information from a victim by masquerading as a trustworthy organization. An example is when one receives an e-mail from a credit card company, stating that something went wrong in the database and the victim needs to go to the (fake) website of the credit card company to give the codes⁴.

In this paper I will consider messages from all of these sources spam, as long as there are no previous dealings with the sender. Furthermore, I consider spam to be spam only when it is sent in (unpersonalized) bulk.

It is important to note that a message could be unwanted (bills) or unsolicited commercial e-mail (job applications, sales inquiries etc.) or bulk e-mail (subscriber newsletters, discussion lists, information lists, etc.) and not be spam.

I therefore rather use the often-used definition that Dutch anti-spam organization Spamvrij.nl also uses: spam is any “Unsolicited, Bulk e-mail” (UBE), i.e. any e-mail that was sent in large quantities (bulk) and where the recipient did not gave prior, explicit and provable consent for the sending of the type of e-mail that was send.

I will not consider viruses. Viruses are arguably easier to find when the virus scanner is up to date, also the amount of viruses is reported to be much lower than the number of spam mails, Messagelabs names 1 virus mail in every 200 e-mails versus 1 spam in every 3 e-mails although others give very different numbers (such as Tweakers.net, whose servers receives some 39% spam and 9% virusses, see also table 2.1).

⁴The program SpoofStick (<http://www.realtimecredentials.com/spoofstick>) is an easy to use program that helps against these attacks by showing the real address.

Chapter 2

Organizational aspects

2.1 Reasons for spamming

The business case of spam is very simple: very low costs, higher returns. Sending e-mail is very low-cost. From any PC, one can send many e-mails in a short time. Because of this, only a very small percentage (much lower than 1%) of the recipients of spam have to respond to it for it to be profitable for the spammer. With a \$1 per unit profit margin, and only a 0.1% response rate, a spammer could make \$10,000 by sending 10 million email messages. Given the very low costs of sending e-mails and the relatively low costs for setting up a website with a so-called bullet-proof hoster¹, the profits are very high. And although spammers have “complained” about the extra costs of getting around spam filters for years², it obviously hasn’t stopped them yet.

Therefore, a way to stop spammers is to make the equation *income – costs* low enough to not warrant the costs and risks involved.

Finding the addresses to target can be rather easy: one could ask a company to send their mailing list for you. These companies have many e-mail addresses to send e-mails to and have the (cheap) means to send large amounts of e-mails - using the same technologies as regular e-mailings but also using so-called Open Relays or PC’s of innocent internet users that have been taken over by trojan horses. These terms will be explained in the next section.

It is also very easy to buy CD’s full of e-mail addresses. The first few hits I had on internet search machine Google gave me millions of e-mail addresses for a few tens of dollars. The magazine PC-Active has, as a test, purchased such a CD and offers a

¹Bulletproof hoster: a (web) hoster that does not care that it is being used for spamming, usually located in countries without spam laws

²<http://www.detnews.com/2002/technology/0208/06/technology-553815.htm>

2.1. REASONS FOR SPAMMING

service³ to see if your own address is on this specific CD - although there are many more CD's out there. There are even spam mails that ask if you want to buy these CD's. . . Recently, an AOL employee was arrested for stealing over 90 million e-mail addresses from 30 million customers, including the zip codes and credit card types (not numbers). He sold these addresses to a spammer for \$52,000.

Another way to get many addresses is to “guess” addresses of specific domains with the assistance of programs, or harvest addresses from web pages or usenet. These harvesters will look for any strings that resemble an e-mail address. A few minutes of searching on the internet gave me a several harvesters, as shown in figure 2.1. Another way are viruses or other malware that infect the PC's of home users and use their address book to send out spam and more viruses. A virus could also be used as a harvester by sending the contents of the address book (which would contain working addresses) to the maker of the virus.

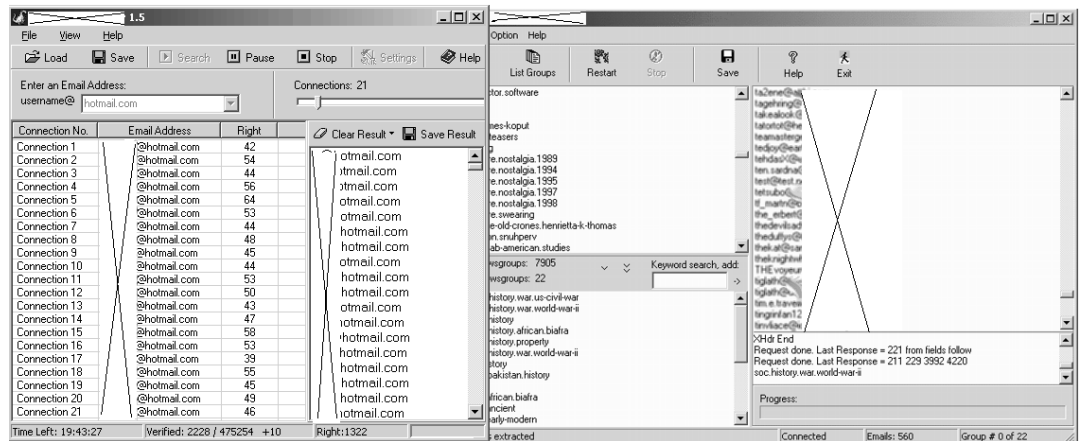


Figure 2.1: e-Mail harvesters

Furthermore, spammers can use several techniques to hide their identity, which makes it harder to block their mails and to prosecute them. These techniques are spoofing of SMTP headers, open relays, open proxies and zombies making use of Trojan Horses, terms that will be explained in the next sections.

2.1.1 Spoofing of SMTP headers

The protocol that is used when sending e-mails, SMTP⁴, was created in a time when security was not as much an issue as there was only a relatively small group of users. Designers probably assumed that all computers and people connected to the network would be trustable. Basically, there were and are no checks to see if the sender is

³<http://www.aktu.nl/pc-active/spam>

⁴Simple Mail Transfer Protocol, see <http://www.ietf.org/rfc/rfc821.txt>

CHAPTER 2. ORGANIZATIONAL ASPECTS

who he says he is. This means it is very easy to spoof the characteristics of an e-mail. The originating IP address, the sender e-mail address and any other characteristic could be forged, which makes it harder to identify the true sender of an e-mail.

A forged sender address could be used to make it look like the mail is from a trusted source. For example, e-mail lists are used to send legitimate e-mail to a group of people with a shared interest. Many people have configured their spam-filters in such a way that they will always let through mails from lists they have opted into⁵. This way, mail will be passed through most spam filters. Also, any error messages will be sent not to the real sender but to the random user, or in this case the real operators of the mailing list.

2.1.2 Open relays

To send and receive e-mail, an e-mail server is required. In “the old days” of the internet, relaying to any server was useful and relaying is built into the SMTP protocol. For example, connections usually were made by dial-up telephone lines and to avoid long distance connection costs, on the ARPANET⁶ a relaying architecture was developed where a computer passed on e-mails to the next computer in a chain of computers until the destination was reached. Many e-mail servers are after a default installation still configured for relaying to and from any servers: securing and configuring the server is required, something that not everyone knows how to do properly, or knows even that it is necessary.

A mail server should only receive mails that are addressed to users with specific domain names (such as @vu.nl for servers of the Vrije Universiteit) and it should only send e-mail from known users. However, some servers⁷ are configured in such a way that they would send e-mails to and from everyone, e-mails for users outside the domain of the server are relayed to the right server (just as e-mails from authorized users are). In this way, spammers can send mass amounts of spam without using their own ISP or bandwidth. The e-mail will look like it is sent from this relaying server.

An advanced (and perhaps intended) form of this is reverse NDR (non-delivery report). These reports were already used to fool recipients in opening spam by attaching the real spam message to what seems to be a non-delivery report. Now, the normal non-delivery reports are abused to send spam from other servers: spam is created with the address of the intended recipient in the sender field and with a random, non-existing recipient at the domain of the targeted server. The mail server cannot deliver the message and sends a NDR e-mail back to what seems to be the sender of

⁵Opt-in: e-mail that one has given prior consent to receive this kind of e-mail, see also section 2.2

⁶ARPANET: a predecessor of the current internet

⁷You could test your own server at <http://www.abuse.net/relay.html>

the original message, the spam recipient. Now the spam victim receives the NDR report and, if the mail server hasn't stripped the contents of the mail, also the original spam.

Normal open relays are easily avoided by configuring the mail server, this second form is harder to stop. A simple "catch-all" address or simply removing mail to unknown users would prevent any NDR mails but this would mean that legitimate senders never know their mail didn't arrive and it would mean not following the standards.

2.1.3 Open proxies

Another way is for the spammer to use "open proxies". Most organizations have many computers on their networks, but have a small number of servers that are directly connected to the Internet. These servers direct the traffic between the internet and the local network, sometimes using proxies. These proxies provide more efficient and secure web browsing for the local users and provide security from the outside to the internal network (which is also done using technologies such as Network Address Translations and firewalls). When the proxy is not configured correctly, it lets unauthorized (outside) computers connect through the proxy (which is then considered "open"). A spammer now can use this proxy to make it look like e-mail is sent from this proxy server, which makes it harder to track the spammer.

Another way to use poorly configured servers, is to find mailscrips on websites used for newsletters and such. When not configured correctly, these could be abused by sending mails by others than the owner of the website. Many web servers have errors such as "script not found or unable to stat: /www/cgi-bin/mailler" in their error logs because of (probable) spammers that are searching for these scrips.

2.1.4 Trojan horses

The German magazine c'T has, together with a student at the Hamburg university and Scotland Yard, recently discovered that spammers pay virus writers to get access to infected machines: they gained access to 52.479 machines for only \$300. The spammers use these infected systems to send spam without the knowledge of the owners of the PC's, that are infected with viruses such as Randex or Deadhat. This way, the spammers have access to thousands of machines to use without any extra costs while being almost untraceable. This way of sending spams has grown considerably, according to network management firm Sandvine these trojan horses now account for four-fifth of all spam mails⁸.

⁸http://www.theregister.co.uk/2004/06/04/trojan_spam_study

2.2 Opt-in and opt-out

In the Netherlands, people can use a sticker to indicate that they don't want to receive unaddressed advertisements in their "real" mail box. In practice, these stickers are widely respected. The digital variant of this is e-mailing the sender that one doesn't want to receive any more e-mails. This is called opt-out. Providing a working opt-out e-mail address is required by the Dutch Telecommunicatiewet (see also section 2.4). Opt-in e-mail is e-mail that the recipients have previously requested. The Dutch Telecommunicatiewet states that e-mail can only be sent when the recipients has previously opted into receiving these mails. As shown in figure 2.2, this is not only good for the recipients but also the senders, as the results of the e-mailing are higher.

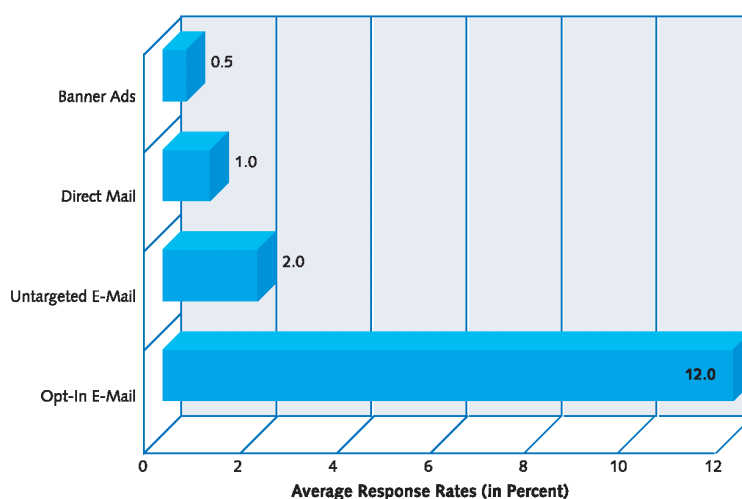


Figure 2.2: Response rates using different marketing methods. Source: the Yankee group, 2000

When one sends a opt-out request to the sender of an e-mail, the sender would know for sure the address is correct and mail sent to it is really read. Spammers sometimes take advantage of this by sending more spam, because the e-mail address has now been verified. For this reason it is often said never to opt-out from spam lists. The US Center for Democracy & Technology tried to find out if this is true by creating e-mail test accounts. After receiving enough spam, they sent opt-out requests to the senders of the e-mails. Many companies responded to this by removing them from the list. As Ari Schwartz, associate director for the CDT said: "Knowing who to opt out from is key. Opting out of legitimate companies drops you off their lists, but when you do that with 'real' spammers, the results are unclear."⁹

⁹<http://www.pcworld.com/reviews/article/0,aid,116572,pg,3,00.asp>

2.3 Why is spam a problem?

For most mail users both at home and in organizations, the largest problem with spam will probably be time lost because of spam. When 30% to 80% of the e-mail is spam, it takes time to receive, read, delete and prevent spam. Too many spam mails, and the habit of some spammers of trying to mimic “real” e-mails (in this context often called *ham*, see also the introduction to chapter 3) makes it more probable that wanted e-mails are missed because of a too strict spam filter or a user that just misses the mail. These e-mails that are falsely labeled spam are called false positives (and spam that is not identified as such is called a false negative).

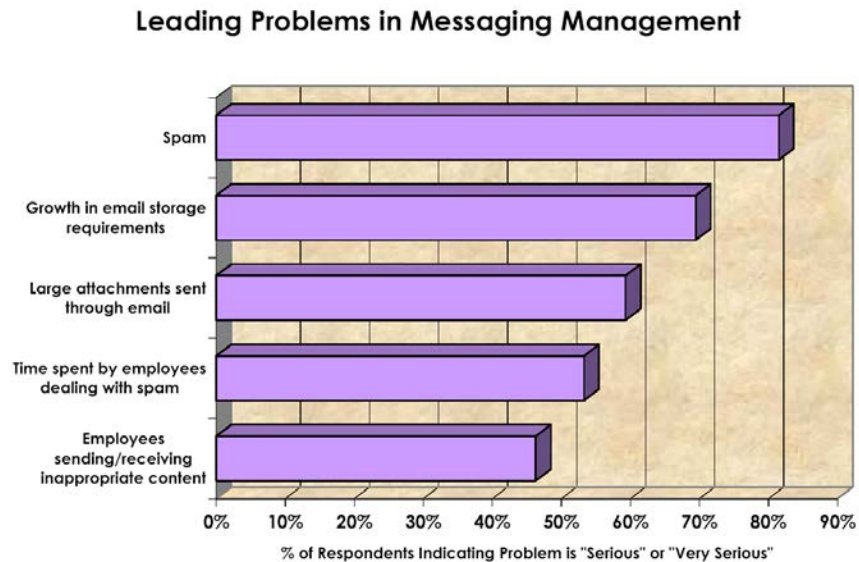


Figure 2.3: Spam is identified as the leading messaging management problem more than 80% of IT decision makers believe that spam is a “serious” or “very serious” problem in their organizations. Source: Osterman Research, 2003

Already in 2001, costs were estimated at some 10 billion US dollars by the European Commission[Dro01] only for the bandwidth used. Although the bandwidth arguably started to be less of a problem as more and more people and organizations are getting faster broadband connections, with the emergence of wireless internet bandwidth is again starting to be a problem, especially when people have to pay their ISP for each received Megabyte or minute spent online. Also, not everyone has access to broadband connections. The same could be said for the amount of storage the spam takes up, even free e-mail accounts such as Gmail or Hotmail have quite large storage space but many home users still have a strict quatum for the amount of mail one could receive. The problem of bandwidth, and server load is in absolute numbers worse

CHAPTER 2. ORGANIZATIONAL ASPECTS

for ISP's and large companies. Spam and viruses (these are not always separated in the numbers I have seen) can amount for as much as 80% of the capacity of the mail servers.

The estimates for costs incurred because of spam for a given organization differ significantly between different research companies. In 2003 Osterman Research calculated \$1400 per year per e-mail user in lost productivity when no spam filtering would be used and \$15 in IT-related costs for filtering plus \$560 in lost productivity per year per e-mail user. In the same year, Nucleus Research estimated these costs at \$874 per e-mail user per year. Ferris research states \$10 billion per year for the US as a whole, or a little over \$140 per e-mail user per year.

Also, companies can be put on blacklists because one of their clients is spamming willingly or is abused by a spammer using an open relays or a trojan horse. When one is on these blacklists, e-mail will be marked as spam or not even accepted by the mail server it is send to. This means their clients can not send any e-mails. In February 2004, the ISP TDS Telecom was put on blacklists for a few days - none of their clients could send e-mails to AOL addresses. This could happen to any ISP.

Spammers usually use fake addresses to send mail from. Sometimes these are non-existing addresses but it also happens that the address of an innocent person is misused as the sender. The real owners of this address will usually get large amounts of bounces when a percentage of the targeted addresses no longer exist. This could even be used to attack the mail servers or the good name of others, such has happened to Spamvrij.nl¹⁰. This spoofing of the address with the intent to attack the victim is called a "Joe job", named after the 'attack' on Joe's CyberPost (joes.com) that received many such bounce mails. But usually, spammers use random addresses.

Another problem is security related; spammers pay programmers of trojan horses to infect PC's. These infected computers can now be abused to relay their spam mails. Not only does this make spamming easier, it also funds virus writers and infects many machines that can be abused for other reasons, such as Distributed Denial of Service attacks on servers on the internet.

The specific content of spam is another problem for some, especially the spam for adult services, products and websites. This could affect small children at home, but it can also be a problem in organizations ("we are all adults but still..."). Dutch researcher Lodewijk Asscher states in [Ass04] that an employer could be sued for providing a hostile work environment by not blocking sexually explicit e-mails such as is seen in some spam mails.

¹⁰<http://www.spamvrij.nl/stichting/vervalsingen.php>

2.4 Legislation

Two years after the start of the case, the Supreme Court of The Netherlands has decided that internet provider XS4ALL could refuse e-mails that Dutch organization AbFab send to users on the XS4ALL servers:

Anyone who without authorisation makes use of property to which another party has an exclusive right, and who thereby infringes that exclusive right, is acting unlawfully vis-à-vis the beneficiary of the right, unless there is justification. The right to freedom of speech does not constitute such justification. This fundamental right cannot serve in principle to justify transgressive use of property to which another party has exclusive rights.

Before the decision was given, AbFab was already declared bankrupt but it at least provides jurisprudence.

Directive 2002/58/EC[Ped02] from the European Union states that for commercial communication (advertising, e-mails, fax, sms, etc.) the opt-in principle must hold - the recipient must give an approval in advance. The deadline for implementing this directive was November 1, 2003, but it took until May 19, 2004 for the Dutch law to be passed. Article 11.7 from the Dutch Telecommunicatiewet (see Appendix A) now states that unsolicited e-mails from companies one has no dealing with is forbidden. Also the option to opt-out of any further mailings is required.

As a result of this law the Dutch Onafhankelijke Post en Telecommunicatie Autoriteit (OPTA) has opened a website¹¹ where private persons can register a complaint when receiving an electronic (e-mail, SMS, fax, etc.) spam message. Unfortunately, this only applies to spam sent to the address of private persons (and not organizations) and where the sender or the person that ordered the spam is in the Netherlands. Its use is therefore rather limited. After registering the complaint, the OPTA can give a warning or a fine; not only the sender of the spam is accountable, but also those who use their services. This makes it easier for both opting out of receiving any more e-mails and finding out who really send the e-mails.

The US also has a “spam-law”: The Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003 (CAN-SPAM). It defines rules for sending e-mails, such as providing an opt-out address and providing correct headers. Also, e-mails with sexual content must be clearly marked as such. Just like the Dutch Telecommunicatiewet, the CAN-SPAM act also applies to the ones that ordered the spam.

Although the people I talked to, state the law might not have any results when looking at the number of spam mails still reaching their mail servers, it might have some results in the US. Spammer Howard Carmack, who sent an estimates 850 million

¹¹<https://www.spamklacht.nl>

CHAPTER 2. ORGANIZATIONAL ASPECTS

spam e-mails, has been found guilty of crimes such as fraud and identity theft. He has been sentenced to seven years in jail and was earlier already fined 16,4 million USD. The US ISP AOL blocks the websites of spammers and sues them for spamming. This seems to be working, February 20 AOL received 2,6 billion spam mails, March 17 it had dropped to 1,9 billion and the number of customer complaints dropped from 12,7 million to 6,8 million¹².

Unfortunately, any national law stops at the border. The international nature of the internet and especially e-mail, makes effective regulations much harder. Regulations could stop some spam, but it is questionable if it is able to stop spammers completely or even noticeably over longer periods of time, as they can easily move to or through another country that has no rules for spam. At the moment, applications such as SpamPal by default block all mail coming from countries such as China. Because of this, for most people all mail coming through China will be marked as spam when these blacklists are used.

2.4.1 Habeas Haiku: copyright against spam?

The Habeas Haiku¹³ is a short poem, copyrighted by Habeas Inc. The idea is that a licensed set of e-mail headers is added to each e-mail. Private persons and companies can now ask Habeas for a license to use this haiku in their e-mail, part of this license are certain conditions on the e-mails that are being sent. So ideally, recipients can automatically pass (“whitelist”) e-mails containing the haiku.

When a company sends out spam containing this haiku, Habeas can sue the senders using the (US) copyright laws instead of the less-clear spam laws. After a recent lawsuit, Habeas was awarded over \$100.000 from one spammer, William Carson. Unfortunately, the sheer amount of spammers mis-using the haiku compared to the number of legitimate users might prevent making automatically passing e-mails containing the haiku an option.

It must be said that copyright works both ways: Computer Associates has recently claimed it owns the patents on several spamfilter techniques such as the often used Bayes filters. If this claim holds, many freeware and opensource spamfilters might become unusable.

2.5 Practice in Dutch organizations

To get an insight to the practical side of spam filtering, I contacted a few organizations and spoke with the people responsible for their spam filters. Obviously, no conclusions could be made from these few interviews about the way spam is dealt with throughout

¹²<http://www.washingtonpost.com/ac2/wp-dyn/A9449-2004Mar19>

¹³<http://www.habeas.com/servicesHowSWEWorks.html>

the Netherlands. It was meant to get a general idea of the way spam is dealt with in practice and find out what their main problems are.

2.5.1 Vrije Universiteit

The Amsterdam based Vrije Universiteit Amsterdam (VU) has several departments that handle e-mail completely independent from each other. I have talked to Hans Leidekker from the IT Facilities department. This department administers the e-mail servers of the centralized services and some faculties at the VU. Other faculties, such as the Business and Exact Science departments have their own mail servers and spamfilters. Their servers block about 15000 spam mails and some 12500 real e-mails are sent through the servers per day.

Handcrafted whitelists and blacklists and public blacklists are used. Because of the many open relay servers and trojan horses all IP addresses of the Dutch ISP's have been blocked, except for the mail servers of these ISP addresses and some IP addresses of trusted users (such as trusted employees and students with their own mail servers). Any and all e-mail connections from IP addresses on the private and public DNS blacklists, are blocked from these IP addresses. Since this has been implemented, the server load has decreased tens of percents. Lists such as relays.ordb.org, bl.spamcop.net, sbl.spamhaus.org and cbl.abuseat.org are used. When a message is blocked, the sender will get a message explaining why and by which list it is blocked, so that legitimate senders can contact the administrators of the lists and remedy the situation.

E-mail that is not blocked is sent to SpamAssasin, which marks the e-mail according to the score of the handcrafted static rules, the Bayes algorithm, Razor and public Blacklists. Furthermore, there are many e-mail addresses that are used as a trap. These addresses are put on the website, but invisible for regular users. These addresses have, just to be sure, a clear text that states that no e-mail should be sent to it, just in case anyone would use the addresses. Now, any mail that is sent to these addresses is counted as spam and the sender and e-mails can be used to block this spam in the future.

2.5.2 Multikabel

Multikabel¹⁴ is a Service Provider in the Dutch province of Noord-Holland that provides several communication services such as Radio and Television, Internet and telephony using cable and DSL.

As an ISP, the spam services have to be tailored to the different groups of customers. Employees, private customers en business customers all use different settings. As

¹⁴<http://www.multikabel.nl>

CHAPTER 2. ORGANIZATIONAL ASPECTS

of yet e-mail for business customers are not filtered, clients with a business account could run their own filters on their internal servers. The e-mail for the employees is scanned more active than those for the private customers - which is possible because the numbers of e-mails is very much lower than that for the customers, the servers would not be able to handle checking all customer e-mails in that detail. The sheer amount of e-mails sent to customers restricts the available options for filtering.

Filtering e-mail for private customers has a few drawbacks. Not all customers are very well accustomed to the internet and computers in general. This means using the filters has to be very user friendly, otherwise the call center of Multikabel would be swamped with calls to the helpdesk. Therefore, spamfiltering is not active by default but it can easily be turned on and adjusted by the customers themselves, using an easy-to-use website. The customers are informed of the options by Multikabel's newsletters and information the Multikabel website.

Until recently, there was only the option to use specific or all blacklists from a given list such as DBSL, Spamcop and Spamhaus. Each of these blocklists blocks certain kinds of addresses mail is sent from. Filtering using each list can be turned on or off according to the wishes of the customer. This unfortunately means that false positives were likely, because the blacklists usually contain a few errors. These cost were minimized by adding two levels of spam: e-mails that were not flagged by any list are received without change, e-mails that are flagged by between 1 and 4 lists are tagged with Spam? in the subject and e-mails that are listed in more than 5 lists are tagged Spam!. This way, deleting mail that is tagged Spam! was still rather safe.

Another option that recently has been implemented is to use a combination of rules, heuristics and blacklists (using SpamAssassin). As the chance of false positives is much lower than using only blacklists, Multikabel offers the option to delete mail on the server when using the heuristics; but only when the customer explicitly chooses this and only when the confidence the message is spam is high or very high. Otherwise, misclassified e-mails could be deleted without the knowledge of the customer. Deleting mail from the server both lessens the server load and the amount of mail the customer has to receive.

2.5.3 Tweakers.net

Tweakers.net is a Dutch website¹⁵ offering IT news and one of the largest forums world wide, targeted at an audience that is "above average" interested in computer technology. It's mail servers send and receives e-mail for a few hundred employees, volunteers and customers/sponsors - a total of some 4200 POP3 boxes and e-mail aliases.

Because of the target group of the website, most users are more than average acquainted with computer and internet related issues. Several employees and volunteers

¹⁵<http://www.tweakers.net>

2.5. PRACTICE IN DUTCH ORGANIZATIONS

Date	Clean mail	Spam	Virus mail	Total
June 8	7798 (52.01%)	5880 (39.22%)	1316 (8.78%)	14994
June 9	8149 (52.20%)	5981 (38.31%)	1481 (9.49%)	15611
June 10	8562 (53.03%)	6245 (38.68%)	1338 (8.29%)	16145
June 11	7971 (52.19%)	5893 (38.58%)	1410 (9.23%)	15274
June 12	7608 (50.87%)	6083 (40.67%)	1265 (8.46%)	14956
June 13	7408 (50.47%)	5937 (40.45%)	1333 (9.08%)	14678
June 14	7647 (53.02%)	5486 (38.03%)	1291 (8.95%)	14424

Table 2.1: Tweakers.net e-mails per day, june 2004

receive large amounts of spam, probably because for a period of time, their addresses were put on the website in machine-readable plain text format. This made it easy for harvesters to collect the addresses. For a long time now, addresses are only shown as machine-unreadable pictures, which lessens this risk considerably. Clearly, other means of gathering addresses are still used.

Viruses are deleted, spam is tagged. No spam is deleted from their servers so the the load on their server is not reduced (and even increased because of the filtering) as the mails still have to be stored. However, compared to the load of the web- and database servers and the web traffic, this would probably not make a difference.

Tweakers.net also uses SpamAssasin, using adjusted standard settings and using blacklists, keywords and several other SpamAssasin options.

2.5.4 Argitek

Argitek¹⁶ is a relatively small (5 employees) firm that advises on e-business application architecture subjects. For filtering spam (and viruses), the filters of the Internet Service Provider are used. This flags most spam and stops most viruses. Mail is received into each employees' own POP mailbox, where the persons that receive most spam have a private spamfilter on the desktop clients, SpamPal. This way, most spam is stopped by the filters on the mail servers and the rest is caught by the Bayes filter in SpamPal. No spam is deleted, only tagged and moved to a different folder in the Outlook mail client so no mail is easily missed.

¹⁶<http://www.argitek.nl>

Chapter 3

Fighting spam

Important concepts when talking about (the performance of) spamfilters are false positives and false negatives. As said earlier, “good” e-mails are sometimes referred to as ham, the opposite of spam when talking about spamfilters, just as Hormel’s Spam (Spiced Ham) versus the “real” eatable ham. False positives are ham that are misidentified as spam and false negatives are spam that are misidentified as ham.

	Spam	Ham
Identified as spam	OK	False positive
Identified as ham	False negative	OK

Table 3.1: False negatives and false positives

When running a spam filter on n e-mails, let n_S denote the number of spam and n_H the number of legitimate e-mails, $n_H = n - n_S$. Now $n_{S \rightarrow S}$ is the the number of correctly classified spam, $n_{S \rightarrow H}$ the number of spam messages incorrectly identified as ham (false negative), $n_{H \rightarrow S}$ false negative and $n_{H \rightarrow H}$ correctly identified ham. Now the false acceptance rate (FAR) and the false rejection rate (FRR) are defined as $FAR = \frac{n_{S \rightarrow H}}{n_S}$ and $FRR = \frac{n_{H \rightarrow S}}{n_H}$. Ideally, these rates are both 0%. But as even humans sometimes misidentify spam and ham, this is as of yet not possible with (near-)realtime automated filters.

In most situations, the cost of false positives is much higher than that of false negatives, especially when identified spams are deleted without the intended recipient noticing this: not receiving an important e-mail from a customer is significantly more costly than receiving a single spam mail. The problem for spamfilters is to minimize FAR and FRR, or rather minimize FAR given a maximum FRR.

Because the cost of a false positive is so much higher, spam is often not labeled as such unless the confidence of “spamliness” is very high, for example a message could be marked as “possibly spam?” when the confidence is 98% and “spam!” when it

is 99.99%. Now it would be arguably safer to delete “spam!” mails and using good filters that would lessen the amount of (visible) spam considerably.

3.1 Network

There are some options that would provide a technical solution to stop spam by changing the network structures. An obvious change would be to use new or existing authentication and encryption methods to make all e-mail identifiable and traceable to its source. Another method is to demand that users have to be positively identified on a mail server before accepting an e-mail. However, both methods would mean a radical change of the way things are done and that is something that is quite impossible to do on the short term on the world wide internet containing many kinds of servers and clients. The same goes for the suggestion to let an e-mail client work out a non-trivial (mathematical) problem so sending more than a few e-mails would make sending it very slow.

Another suggestion is to charge a very small amount of money for each mail. Trojan horses would be a huge problem, as the victims would have to pay for each spam mail. Also, the costs for making the micro-payments at the moment still outweigh any proceeds so implementing this could as of yet be too costly for senders and / or implementors.

Sender ID

May 2004, Microsoft and SPF's author Meng Weng Wong stated that they would merge their technologies into one technique with the name Sender ID. However, specifics of this implementation have not yet been made public.

Caller ID¹ is a mechanism developed by Microsoft to stop spoofing. The administrator of a specific domain can, in the so-called DNS² records, specify the IP addresses of the machines that are allowed to send e-mail from their domain. The recipient of e-mail can now reject e-mails from their users when the mails are not sent from one of the specified IP addresses.

When receiving e-mail, the receiving server examine the mail to determine the domain where the mail was sent from by examining the e-mail envelope. The receiving mail server now queries the DNS for the list of outbound mail server IP addresses of the specified domain. If the sending IP address is not on the list, the e-mail is probably spoofed and therefore possibly spam or a virus.

¹http://www.microsoft.com/mscorp/twc/privacy/spam_callerid.mspx

²Domain Name System, the system that is also used to translate domain names to IP addresses, smtp.cs.vu.nl = 192.31.231.66 etc.

CHAPTER 3. FIGHTING SPAM

Sender Policy Framework (SPF³) basically does the same, it also specifies the domain names of accepted mail servers that send e-mail in the DNS records. e-Mails that are not originating from the correct servers can be rejected at the time the mail is received or after receiving the body of the mail, depending on the way it is implemented.

If no SPF record is available, the mail is accepted and other ways to detect spam must be used. Also, receiving servers that do not check the SPF records use the normal ways of detecting spam. So it only affects e-mails where both sides use SPF. SPF/Caller ID would mean that their addresses could no longer be misused as spam senders as long as their servers are correctly secured, “Joe jobs” would become impossible without access to the right mail servers.

Because the addresses of the servers used to send mail from have to be known, these schemes are only practical for organizations where these addresses are known and fixed. Many employees of organizations and most customers of ISP’s also send mails from other locations than their work or home PC. This means most ISP’s and other organizations won’t be able to use SPC/Caller ID without making their mail servers available to all internet users.

An option for these organizations is using a STMP server that is accessible from all over the world, using existing technologies such as VPN, Secure SMTP, pop-before-SMTP, certificates, etc. to make sure only authorized people use the servers. For most companies with external employees and centralized administration of remote computers this would be a viable option. For most customers of ISP’s and many free e-mail services however, using these authentication techniques are still too hard to use.

DomainKeys

Yahoo is developing a similar scheme to prove the identity of the sender, DomainKeys. A domain administrator creates a public/private key pair which is used to sign all outgoing e-mails. The public key is published in DNS and the private key is made known to all outbound e-mail servers of the organization. When an e-mail is sent by an known, authorized user within the domain, the server automatically uses the private key to sign the e-mail by adding a line in the e-mail headers. Receiving servers now check the e-mail headers with the public key to verify the sender is really mailing from the domain he says he is.

AOL, British Telecom, Comcast, EarthLink, Microsoft and Yahoo have formed ASTA (Anti-Spam Technical Alliance) to try and form standards and create industry guidelines to address the spam problem.

³<http://www.linuxjournal.com/article.php?sid=7327>

Tar pits

When one would delay (all or only suspect) mails for a few seconds or minutes and keep the sender waiting, sending mail would become somewhat costlier. This costs the receiving server more processing power than the sender. But if enough servers use this tar pits, spamming becomes too expensive. The results would be less when spam is sent through zombies but even here it would work, as more zombies are required to sent the same amount of mail in the same time frame.

3.2 Applications

Spam filtering can be done at the server, on each client, or both. Applications that remove spam from servers save significant amounts of bandwidth and disk space, also they are more likely to be managed by a competent staff. On the other hand, server applications have to deal with the specific wishes and usage patterns of the sometimes very different users. Also the costs of false negatives when deleting mails at the server are higher, because the recipient doesn't know he has missed e-mails. Serverside solutions can be software or hardware, also e-mail filtering could be outsourced to third parties, where all e-mail is passed through their servers.

Users can train their local clients for their individual e-mail patterns, also they can keep personal white- and blacklists. Patterns at the group level are probably less adjusted to specific users - the IT department receives other kinds of mail than the marketing department, a 40 year old father reads other e-mail than his 15 year old daughter. Furthermore, the diversity of individual spam filters makes finding a way around these filters harder for spammers. It is worth the time and effort to tailor a spam to the AOL filters, but not to pass my personal spam filters.

Client side solutions can use proxies, where each e-mail first passes a mini-"server" application that inspects the e-mails, or can use a plugin that attaches itself to a specific application. Plugins are by their nature only usable by specific mail applications, often Microsoft Outlook. This can be a problem when upgrading software. Proxies are much easier to setup but giving feedback when mail is misclassified is harder because there is no "this e-mail should be spam" button that retrains the software.

I will now give a short description of a few kinds of spam filters applications and tools.

3.2.1 Server side filters

procmail

Procmail is an e-mail processing program for Unix (and Linux), using regular expressions⁴. It is used by for example mailing lists but can and is also used as a spam filter. Not only is it easy to build own filters based on the different properties of the e-mail, there are also many scripts and programs available that are used by procmail for filtering spam. Procmail is a powerful tool that can be used to build spam filters, among other things. With some simple rules it is possible to remove significant amounts of spam, but it is not a full spam solution. It can however, easily call other programs to scan mail. For example, procmail can be used to call upon SpamAssassin.

SpamAssassin

Probably because of the combination of power and price, SpamAssassin⁵ is used in all organizations I contacted when writing this paper.

SpamAssassin is an opensource, free mail filter to identify spam on client or server. Integration with many mail servers is possible and commercial plugins for Windows servers and clients such as Microsoft Exchange or Outlook are available. It is a flexible and powerful set of Perl scripts, that uses the combined score from various types of checks to determine if a message is spam. It looks for forged headers, checks the headers and body of e-mails for keywords, matches using regular expressions, the percentage of specific keywords or for example HTML tags, etc. It can use Bayesian filtering, many white/blacklists, collaborative spam identification databases such as Razor and DCC and checks character sets and locales. Also, it allows for many third-party plugins that check for example the body of messages.

The strength is the combination of the different methods, each method gives a message points for the amount of “spamminess”. The mail is only regarded as spam when the total spamminess exceeds a certain threshold set by the administrator or even the recipient. No single rule can mark a mail as spam, each rule gives the mail a score of “spamminess”. Therefore when spammers adjust e-mails to defeat one method, the other methods would still recognize the spam.

As many implementations for Microsoft Exchange Server are not free, the free OR-Filter⁶ from Dutchman Martijn Jongen might be a good alternative for companies with a tight budget and an Exchange server. Although SpamAssassin is free for Unix-like servers, this does not mean administering the server is free. It requires a skilled

⁴Regular expressions are templates defining patterns to be matched to a given text string such as an e-mail address.

⁵<http://www.spamassassin.org>

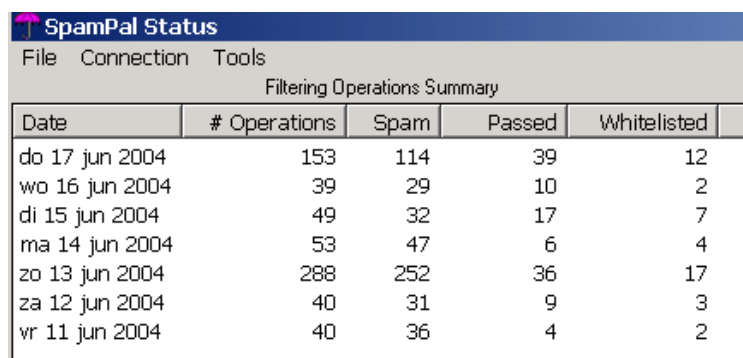
⁶<http://www.martijnjongen.com/eng/orfilter>

administrator as shown in for example [See04]: the default settings of SpamAssassin give result that are not nearly as good as tweaking the settings and weights.

3.2.2 Clientside filters

On the client, there are many different options. Easy to use spamfilters such as Robin Keir’s K9⁷, POPFile⁸ or SpamPal⁹ can be very well used on the client to identify spam, based on the personal e-mail patterns. Installation is very easy and depending on the usage and e-mail patterns, very good recognition rates are possible. Both K9 and POPFile use the Bayes theorem, SpamPal uses blacklists and using many plug-ins such as Bayes, regular expressions and basic whitelists of “good” or “bad” keywords.

Many e-mail clients filter spam directly or through a plug-in, but many other e-mail programs can only move messages to specific mailboxes using basic rules. For example, all mails containing the word “cheap!” and all messages with the subject “*spam!*” are moved to a spam mailbox, these subjects are set by spamfilters at the e-mail server or by a local client side spamfilter that is configured to act as an intermediary between the e-mail client and the server. Programs such as K9 and SpamPal work as such a proxy without integration with the e-mail client program.



SpamPal Status					
File Connection Tools					
Filtering Operations Summary					
Date	# Operations	Spam	Passed	Whitelisted	
do 17 jun 2004	153	114	39	12	
wo 16 jun 2004	39	29	10	2	
di 15 jun 2004	49	32	17	7	
ma 14 jun 2004	53	47	6	4	
zo 13 jun 2004	288	252	36	17	
za 12 jun 2004	40	31	9	3	
vr 11 jun 2004	40	36	4	2	

Figure 3.1: Results using SpamPal on one of my personal accounts

Often-used Microsoft Outlook has a set of standard rules and heuristics that are supplied by Microsoft and updates through it’s Office Update website. Mozilla Thunderbird has a build-in bayes algorithm that trains itself using the build-in “this is spam” button.

SpamCop is a Peer to Peer (P2P) solution, users can tell SpamCop that a specific message is spam. If enough others do the same, other clients receiving the same mail

⁷<http://www.keir.net/k9.html>

⁸<http://popfile.sourceforge.net>

⁹<http://www.spampal.org>

will flag it as spam. False positives are almost assured, because the definition of spam is very different for some people, as show earlier.

3.3 Hardware or outsourcing

Most (smaller) organizations cannot be expected to employ administrators that know the ins and outs of spam filtering and all other aspects of their jobs. There are specific hardware network e-mail filter appliances, such as from CipherTrust, that are managed by it's creators. Besides the easier management, the dedicated hardware can also take the load of the mail servers so that these can perform their main function better; routing e-mail.

Third parties are also an option, either by employing the services of specialists to administer ones own e-mail servers or by routing all e-mail through the spamfilters of third parties (like it is done by many private persons and their ISP servers). The spamfilters of ISP's could be enough for many private persons or small organizations, larger organizations could employ the services of specific e-mail service providers or let others administer the e-mail servers.

Chapter 4

Filtering methods

Anti-spam filters use various ways to identify spam and ham. Checks can be done based on the origin of the e-mails, or their contents. Also some other methods are used, such as looking at the amount of mail from a specific server in a given time frame or by setting up dummy addresses. When looking at the contents of e-mails, techniques such as word matching, phrase matching, heuristics or statistics are used. In this chapter I will briefly present some typical approaches to spam filtering.

4.1 Origin

When receiving e-mails, the receiving server learns a few things about the origin and the route of the e-mail. Relevant information here are mainly IP address, domain and e-mail addresses. The origin of an e-mail is usually checked with existing black- and whitelists. Unfortunately, spammers also know the providers of these blocklists and several distributed Denial of Service (dDoS) attacks have happened to their servers with the intent to block the services of these lists. Even when the attacks are not enough to make the servers unavailable, it increases the costs of running these servers and especially free services hurt because of this - some victims of a dDoS have stopped providing the service¹.

4.1.1 Blacklist

(The originating addresses of) e-mails can be checked with so-called blacklists (also called blocklists). These can be administered by third parties or by the own administrator of the mail server.

¹<http://msnbc.msn.com/id/3088113>

CHAPTER 4. FILTERING METHODS

There are a few different kinds of blacklists: an *open proxy list* blocks e-mails from known open proxies (usually zombies). See also CBL in section 4.4. *Open relay* lists block e-mail servers that are configured incorrectly and as such abused to route spam through.

Other lists are more subjective. The *Spamhaus*² team maintains a free list of spam sources, *SpamCop*³ is a (\$30/year) list that is maintained by end-users: when enough end-users say a message is spam, it is added to the list.

Many blacklists list large sections of internet address space. Peer to Peer applications such as SpamCop provide blocklists that sometimes also block legitimate e-mails from bona fide mailing lists: what is considered spam by one, is a mailing list for another. Also some lists block all e-mails from customers of specific providers, or even block whole countries. Companies could even be blacklisted by some lists because of their policy or the way they present themselves to the world, even if the proprietors of the site have never sent spam. For example, this is the case with the rfc-ignorant blacklist. This means choosing the right blacklists is very important in minimizing the number of false positives.

4.1.2 Vipul's Razor

Vipul's Razor⁴ is a distributed, collaborative ("P2P"), spam detection and filtering network. Razor supplies a distributed and constantly updating list of spam e-mails that is used by e-mail clients to filter out known spam. It looks at (the checksum) of messages, not e-mail addresses or host names. When an e-mail arrives, the checksum or signature is compared to the list on the servers. When a spam is discovered that was not recognized by the server, the user is supposed to report the signature to the server, so that when enough people do this, the signature is added to the blocked list. Distributed Checksum Clearinghouse⁵ is a similar list.

4.1.3 Whitelist

The opposite of a blacklist is a whitelist. These whitelists are organization- or person-specific and lists addresses, domain names and / or servers that are always passed through the filters. For clientside filters this usually contains the user's address book. On e-mail servers it could be a handcrafted list of mail servers, domains and addresses that might be blocked by lists. This way, whitelists are a way to make sure mail from specific known senders or servers reaches you.

²<http://www.spamhaus.org>

³<http://www.spamcop.net>

⁴<http://razor.sourceforge.net>

⁵<http://www.rhyolite.com/anti-spam/dcc>

4.1.4 Challenge/Response

If a message arrives from someone who is not on the whitelist, the C/R software automatically sends a reply and the message is put on hold. This reply contains the request to respond in a specific way. This request is formed in such a way that it is easy for a person but hard for a computer to reply to it. For example, the reply might ask the person to send another e-mail with a specific word in the subject line, or in the body of the mail. Another option is to open a specific URL on the recipient's website or ask the sender to fill in a web form (such as shown in figure 4.1). Only when this reply is received, the sender is put on the whitelist and the e-mail becomes available to the recipient. This way, only humans that take the time to respond will get through.

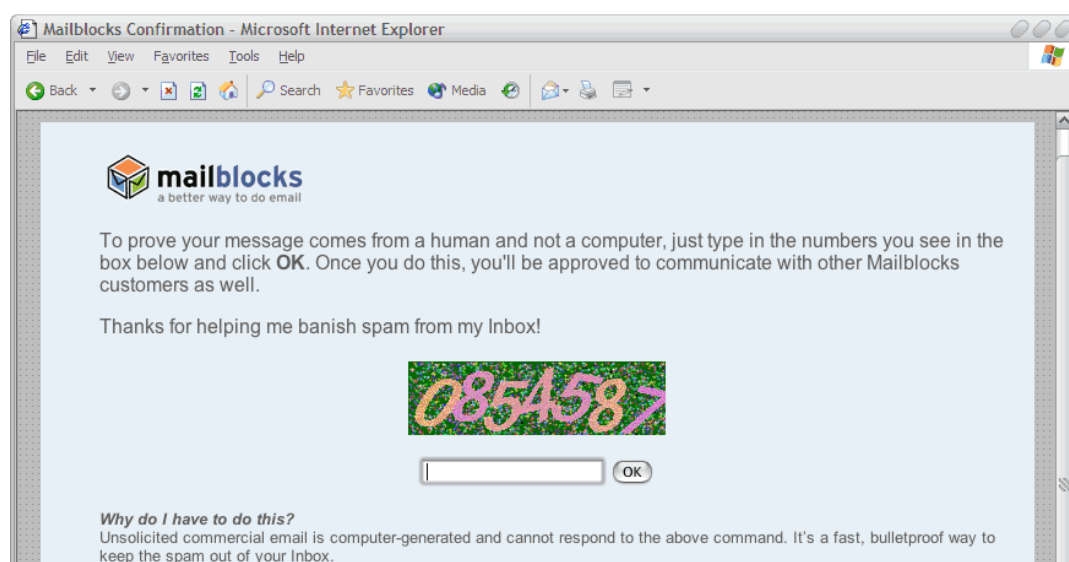


Figure 4.1: A Challenge/Response. Source: Mailblocks

This technique has several problems. When a spammer uses existing fake return addresses, you are (a small) part of overloading the “senders” e-mailbox. Also, receiving e-mails from (machine-sent) mailing lists becomes impossible without first placing the sender on a whitelist and many people would not want to respond to the challenge as it sends the signal “my time is more valuable than yours”. Also, spammers could direct someone to their website by sending out fake challenges pointing to the spammers site.

4.2 Contents

Blocking e-mails “at the gate” by checking if the sender is on a blocklist catches much “low-hanging fruit” in a way that costs little processing power or bandwidth. Filters that also look at the contents take much more processing time per e-mail but can be more precise with less false positives. The contents not only consists of the human-readable contents, but also the e-mail headers⁶.

Basically, there are two kinds of content analysis: “superficial” analysis, based on string matching, some syntactic checks, etc and “non-superficial” analysis based on text mining techniques.

4.2.1 Preprocessing

Before using these techniques to identify spam, the contents of the messages could be preprocessed to circumvent some techniques that spammer use to fool spamfilters.

The best known way to fool spamfilters is by substituting the letters in such a way, that it is readable by humans but not picked up by the spam filter. V!agr@ is readable for humans, but if the spam filter looks for Viagra, it will miss it. The same happens with words such as v*i*a*g*r*a, substituting a letter with a picture of the letter or inserting HTML characters that are ignored by a mail client.

On the other hand, this kind of substitution has become a good sign for spam, no human would write Viagra as it’s HTML equivalent “V i - ag r a”

A “dictionary salad” spam is where large sections of the text are randomly chosen dictionary words - hoping to bury the contents spam in the random dictionary noise. “Long story” spams add a few paragraphs from a normal e-mail or a news story to a spam mail. This way, the filter might be fooled when the right “good” words were added.

HTML de-commenting and partial rendering of messages makes it more useful because tricks to fool spamfilters are ignored. For example, a message could contain invisible HTML tags between words or even letters. Also, using replacement functions to substitute ! with i etc increases the chances of identifying spam. Words or phrases can be matched by Regular Expressions. Disregarding the impact it would have on server loads, using known tools of search engines might increase this further. For example, “soundex” uses tables to relate “Viagra” and (in Dutch) “Viachra” and deliberate spelling errors can be corrected just as easily as the kind of errors normal search engine users make when searching with Google.

⁶Reading e-mail headers: <http://www.stopspam.org/email/headers.html>

4.2.2 URI Blocklist

Most spam messages link to a website where the victim (or customer) can purchase the products or services of the spammer. SURBL⁷ (Spam URI Realtime Blocklist) is very different from other blacklists. It blocks e-mails based on the domain names in message body Uniform Resource Identifiers (URI's, usually the addresses or URL's of websites). SURBL is not used to block spam servers like most other RBL's, it allows receiving mail servers to block specific messages based on the domains that occur in the body of the e-mail.

URL filtering check not just the message and the headers but the links within the message. For example when a e-mail links to www.buymyviagra.com and this is on the list, this is an indication a message is spam. This works probably best for spam that only contains a link, as other content-based filters usually cannot be used with these messages. Also, it targets the people that *ordered* the spam, not the spammer itself.

Comparable to injecting dummy words in an e-mail, this method could be fooled by inserting (invisible) links to regular websites such as www.spamassassin.org or www.vu.nl, just as other content-based filters could be fooled by inserting “normal words”, as shown in the next section.

4.2.3 Phrase matching

Spam filters can check incoming mail with sets of keywords or phrases. For example, when an e-mail contains the word “viagra” or the phrase “get free access”, I could just delete the e-mail as I'm not expecting any serious mails using this word. These lists are more effective combined with scoring, see section 4.5.

4.2.4 Text mining

Text mining is used to identify spam e-mails after receiving training on messages that have been manually classified as spam or ham. Various techniques are used to identify spam, with different results. The best results are when the received e-mails have a specific profile. This usually means a personal mailbox or a (homogeneous) department of a company.

Naïve Bayes

Naïve Bayes has since the paper by Sahami et al. ([Sah98]) become a de facto standard for identifying spam with the aid of text mining. And it is having results, judging by the methods spammers use to get past them, like dictionary salads.

⁷<http://surbl.org>

CHAPTER 4. FILTERING METHODS

The well-known Bayes theorem is at the basis of the Bayes method, $P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$. When identifying spam, the filter looks at n words from an e-mail $\vec{x} = (x_1, x_2, x_3, \dots, x_n)$, the values of attributes $\vec{X} = (X_1, X_2, X_3, \dots, X_n)$. Binary values are used for X_i , $X_i = 1$ if a specified word (such as “viagra”) exists in the message, $X_i = 0$ otherwise. The problem is to estimate the probability a specific message is in the spam class $C = c_{spam}$ (or alternatively, the class $C = c_{ham}$).

$$\begin{aligned} P(C = c_{spam} | \vec{X} = \vec{x}) &= \\ P(C = c_{spam} | x_1 \wedge x_2 \wedge \dots \wedge x_n) &= \\ \frac{P(x_1 \wedge x_2 \wedge \dots \wedge x_n | C = c_{spam}) \cdot P(C = c_{spam})}{P(x_1 \wedge x_2 \wedge \dots \wedge x_n)} & \end{aligned} \quad (4.1)$$

The values of $P(x_1 \wedge x_2 \wedge \dots \wedge x_n | C = c_{spam})$ are practically impossible to estimate, because there are too many possible combinations of \vec{X} in the dataset and the known data is limited.

However, by making use of Bayes and the “naïve” assumption that occurrences of words can be viewed as independent of each other,

$$\begin{aligned} P(C = c_{spam} | \vec{X} = \vec{x}) &= \\ \frac{\prod_i P(x_i | C = c_{spam}) \cdot P(C = c_{spam})}{P(x_1 \wedge x_2 \wedge \dots \wedge x_n)} & \end{aligned} \quad (4.2)$$

The probability the message is spam can be compared to the probability it is ham;

$$\frac{P(C = c_{spam} | \vec{X} = \vec{x})}{P(C = c_{ham} | \vec{X} = \vec{x})} = \frac{\prod_i P(x_i | C = c_{spam}) \cdot P(C = c_{spam})}{\prod_i P(x_i | C = c_{ham}) \cdot P(C = c_{ham})} \quad (4.3)$$

Given enough data, this can be calculated by taking the number of spam mails N_{spam} and ham mails N_{ham} :

$$\begin{aligned} &= \frac{N_{spam} \cdot \prod_i \frac{N_{spam,i}}{N_{spam}}}{N_{ham} \cdot \prod_i \frac{N_{ham,i}}{N_{ham}}} \end{aligned} \quad (4.4)$$

where $N_{spam,i}$ is the number of spam that use the word X_i .

To create a filter that identifies spam with low false acceptance and false rejection rates, the words \vec{X} must be chosen. For this, the Mutual Information $MI(X_i, C)$ is calculated ([Tho91]).

$$MI(X_i, C) = \sum_{X_i=x_i, C=c} P(X_i, C) \log \frac{P(X_i, C)}{P(X_i)P(C)} \quad (4.5)$$

Now, a specific amount of features is selected for which this MI is greatest. Sahami et al ([Sah98]) used 500 features but one could experiment with other values. Also, different training strategies have been tried, such as train all messages, train until no errors occur and train on error. Finding which amount of features and which update methods work best for spam filtering is outside the scope of this paper.

Bayes is not for most home users, as it requires more-or-less constant retraining. Most people install a program, train it perhaps for a short while by pointing to the existing spam e-mailbox and then forget about it. Consequence is that the results are not optimal. However when integrated in the e-mail client, it offers an easy way to use spamfilter (just push the “spam” or “ham” button when a message is identified incorrectly). Without a client with an integrated filter or when using serverside filters, auto-learning can be used: all new messages are used to train the filter further - here spam can also be flagged by using other methods, as explained in section 4.5.

A novel way that Bayes is (mis)used as, is to configure spam to pass spamfilters. Feedback to train this filter is, instead of user input with “good” Bayes filters, done with the aid of web bugs and clicks to identify spam that got through, and bounces, error messages etc. to identify spam that were recognized as spam. This feedback can be used to find the right combination of words in a “word salad“ to get past the Bayes filter of the intended victims.

Markovian filter

At its basis, the Markovian filter is equal to the Bayes filter. The features are changed from single words to multiple words. Weighting is introduced so that longer features have more weight. These weights can be chosen in several ways, William S. Yerazunis states in [Yer04] that 2^{2^n} weighting gives the best results where n is the number of words after the first word. That is, 1 word gets the weight $2^0 = 1$ and groups of 2, 3, 4, etc. words have weights $2^{2*1} = 4, 16, 64$, etc.

Given the text “Lorem ipsum dolor sit amet”, the weights than are:

Feature Text	weight
Lorem (something else)	1
Lorem ipsum	4
Lorem (something) dolor	4
Lorem ipsum dolor	16
Lorem (something) (something) amet	4
Lorem ipsum (something) sit	16
Lorem (something) dolor sit	16
Lorem ipsum dolor sit	64

Table 4.1: Weighted Markovian Features

This way, a single sentence can overrule many single words and short chains and label

a message spam or ham. Downside to this is that the amount of memory used for the database and the time to scan a single message is increased because of the very much larger amount of data. This means that, for a given memory size, less keywords of sentences can be kept in the memory of the filter software.

4.3 Traffic Volume

The same mail is sometimes sent to many users on the same mail server, some spammers send an e-mail to every known account on a domain in a very small time frame. A mail server could detect this flood and block any further messages from the same server, or it could tag the messages as possible spam.

4.4 e-Mail minefields

Some spam campaigns send an e-mail to every known account on a domain. An e-mail minefield can be formed by adding a large set of dummy e-mail addresses to the website of the organization. These email addresses are intentionally leaked to spammers. An example of this is s23ef34fser@domain.nl. As no human would send email to these addresses, any email to the addresses must be spam and can immediately be added to the blacklist.

Composite Blocking List⁸ is an example of a blacklist using these minefields. Any e-mails reaching these very large spamtraps will be automatically checked to see if the originating IP has characteristics of open proxies, if this is the case the sender's IP is automatically put on the blacklist. More important: it also has a "no questions asked" removal policy, which allows incorrectly listed addresses to be removed in a very short time. This limits false positives.

4.5 Scoring

Any and all of the mentioned methods work with varying results. According to most people I spoke with, better results are possible by combining several methods. Instead of passing or blocking mail based on a single criterium, the filters use penalty bonus points when a specific feature is present in an e-mail. For example:

- if the phrase "BUY NOW" is in the body, add 90 points ,
- if the sender is in my address book, subtract 90 points,
- if the body contains HTML, add 25 points,

⁸<http://cbl.abuseat.org>

- if one of the servers is in the DNS Blocklist, add 50 points,
- if the mail is also sent to someone else in my address book, subtract 20 points,
- if Bayes flags the mail as spam, add 100 points,
- etc.

Now spam is only flagged as such when the number of points exceeds a certain threshold.

The different algorithms and heuristics all look for different items. Each method in itself would be “crackable” by the spammers but using scoring thresholds, all methods combined give a good way to stop spam. As the type of spam and the costs of false negatives and false positives differs per organization and testing a significant amount of current spam using many different methods would be too time-consuming for this werkstuk, I will not call any algorithms “best” but refer to others that have compared methods.

Chapter 5

Conclusion and recommendations

Although the costs of spam are rising for the spammer because of legislation and filtering, spam will continue to exist for the near future. Businesses should take counter-measures when their costs of spam become too high. These counter-measures are a combination of technical, organizational and legislative methods. I have shown some examples of these methods and techniques.

A question then is where to filter e-mails: at the client, the server or both. Filtering at both sides would be preferable for most organizations. Server administrators of most smaller organizations cannot spend too much time training and configuring the mail filters so they could rely on the servers of their ISP or the filters of third parties to filter most “low hanging fruit”, the bulk of the spam. This should be done in such a way that the number of false positives is kept to an absolute minimum. Domain- and user-specific knowledge can be applied at the department or user level using local filters. Client side filters could also be used, but as feedback is needed, it should only be used by knowledgeable users. Home users could use a combination of ISP filters and client applications such as K9 or SpamPal. Larger organizations administering their own e-mail servers have the option to choose between many different applications and methods. Also here, using filters at both the servers and clients would be effective.

There is usually no reason to limit to only a single algorithm or method, combining these in applications give better results than most single methods, although the used methods are limited by the amount of mail being received. Tweaking Bayes filters and Markovian filters and other algorithms such as Support Vector Machines are however worth investigating in detail. Because of the scope and size of this paper, I cannot discuss which algorithms give the best results, for some discussions of alternative algorithms compared to (usually) the Bayes algorithm see [See04], [Sta00], [vN04], [Hid02], [Lon03] and [Sch03].

As has been shown, many options are available to prevent or filter spam. These options are divided between local measures that can be taken by each organization and private person, and measures that have to be done globally through legislation or changes to the protocols used to send and receive e-mail. I will now give some tips & tricks that I came by while reading about spam and talking to the professionals: best practices, if you will.

- Server admins: DO use Non-Delivery Reports on servers. NDR's are an essential part of the way e-mail works. Depending on server loads, stripping the original message and deleting spam and virus mails that is known to be malware is an option but when there is any doubt about the e-mail, the lack of a NDR would be an indication for the legitimate sender that the mail has been sent and received. Also, you yourself could be put on blacklists such as the rfc-ignorant list.
- Trojan horses, open relays and proxies are a cheap and easy way for spammers to send spam. Keeping your own PC's and servers up to date and tightening your security not only keeps your own network and data safe and it helps to keep your internet-costs down, but it also helps the rest of us by your not being mis-used as a "spam zombie".
- Do not make e-mail addresses on your site easily readable by software: use pictures of the addresses or forms on the website, or otherwise obscure the addresses. A correctly configured spamfilter can stop most of the spam but making it easy for the spammers is not necessary. Use honey-pot addresses if you want to receive spam for analysis or placing the senders in your blacklists. However, do not visitors with impaired vision. They cannot always read these obfuscated addresses. Options here are to have an audio version where the address is spoken out loud, or using web forms to send e-mail to (only) an internal address.
- Given the current state of legislation in Europe and the United States, opting out of e-mails originating in these countries *should* work in theory. So do not respond to spam mails unless you think the sender is a trustworthy company. Spammers also receive feedback when images or other contents of an e-mail are downloaded from the internet by the mail reader. Any acknowledgement that you exist and read your mail would make you a more attractive target. Therefore, configure your mail reader in such a way that it cannot download contents of the web. Another option is to only show messages as clear text instead of HTML. This also stops some viruses when using for example Outlook.
- Private persons with an internet domain could use different addresses for different communications. If you have to e-mail an unknown and untrusted person, you could use a new address and abandon this after you are done with it. This way, any further e-mail on this address can be regarded as spam and bounced without processing. Furthermore, you now know that the company you send

CHAPTER 5. CONCLUSION AND RECOMMENDATIONS

the mail to cannot be trusted to respect your privacy. Also, private persons receiving spam from within the Netherlands can report this spam to OPTA.

Depending on the situation, one could use some or all of several filtering methods that are often available for free: use only a few specific generic blacklists because of the risk of false positives and use handmade black- and whitelists. Use handcrafted or publicly available keywords and heuristics, based on standard lists from sources like SpamAssassin or use commercially available software. Server- and client side solutions can use algorithms such as Bayes and other filters and consolidate the different results to label a message as spam or possible spam. But in the end, some spam will get through and some legitimate mails might be blocked by the filters and spammers will try new methods to find ways around your filters. So keeping up to date is advised.

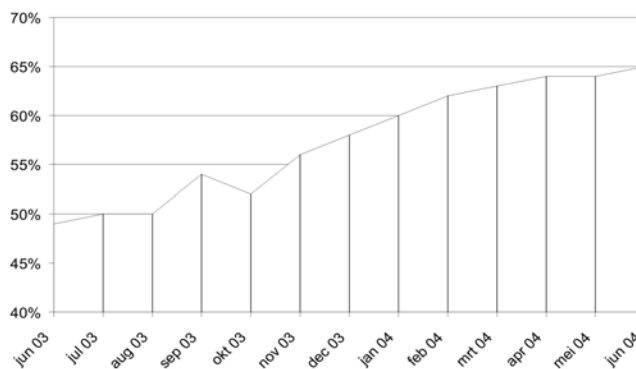


Figure 5.1: Percentage of e-mail identified as spam by Brightmail (from a total of 104 billion checked e-mails)

Appendix A

Article 11.7 from the Dutch Telecommunicatiewet

1. Het gebruik van automatische oproepsystemen zonder menselijke tussenkomst, faxen en elektronische berichten voor het overbrengen van ongevraagde communicatie voor commerciële, ideële of charitatieve doeleinden aan abonnees is uitsluitend toegestaan, mits de verzender kan aantonen dat de desbetreffende abonnee daarvoor voorafgaand toestemming heeft verleend, onverminderd hetgeen is bepaald in het tweede lid.
2. Een ieder die elektronische contactgegevens voor elektronische berichten heeft verkregen in het kader van de verkoop van zijn product of dienst mag deze gegevens gebruiken voor het overbrengen van communicatie voor commerciële, ideële of charitatieve doeleinden met betrekking tot eigen gelijksoortige producten of diensten, mits bij de verkrijging van de contactgegevens aan de klant duidelijk en uitdrukkelijk de gelegenheid is geboden om kosteloos en op gemakkelijke wijze verzet aan te tekenen tegen het gebruik van die elektronische contactgegevens, en, indien de klant hiervan geen gebruik heeft gemaakt, hem bij elke overgebrachte communicatie de mogelijkheid wordt geboden om onder dezelfde voorwaarden verzet aan te tekenen tegen het verder gebruik van zijn elektronische contactgegevens. Artikel 41, tweede lid, van de Wet bescherming persoonsgegevens is van overeenkomstige toepassing.
3. Bij het gebruik van elektronische berichten voor de in het eerste lid genoemde doeleinden dienen te allen tijde de volgende gegevens te worden vermeld:

APPENDIX A. TELECOMMUNICATIEWET

- (a) de werkelijke identiteit van degene namens wie de communicatie wordt overgebracht, en
 - (b) een geldig postadres of nummer waaraan de ontvanger een verzoek tot beëindiging van dergelijke communicatie kan richten.
4. Het gebruik van andere dan de in het eerste lid bedoelde middelen voor het overbrengen van ongevraagde communicatie voor commerciële, ideële of charitatieve doeleinden aan abonnees is toegestaan, tenzij de desbetreffende abonnee te kennen heeft gegeven dat hij communicatie waarbij van deze middelen gebruik wordt gemaakt, niet wenst te ontvangen en indien de abonnee bij elke overgebrachte communicatie de mogelijkheid wordt geboden om verzet aan te tekenen tegen het verder gebruik van zijn elektronische contactgegevens. Aan de abonnee worden in dat geval geen kosten in rekening gebracht van voorzieningen waarmee wordt voorkomen dat hem een ongevraagde communicatie wordt overgebracht.

Bibliography

- [Ass04] Asscher. Regulating spam. directive 2002/58 and beyond, 2004.
- [Dro01] Gauthronet; Drouard. Unsolicited commercial communications and data protection, January 2001.
- [Hid02] Hidalgo. Evaluating cost-sensitive unsolicited bulk email categorization. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 615 – 620, 2002.
- [Lon03] Massey; Thomure; Budrevich; Long. Learning spam: Simple techniques for freely-available software. In *Proceedings of the 2003 Usenix Annual Technical Conference, Freenix Track*, 2003.
- [Ped02] Cox; Pedersen. Directive 2002/58/ec from the european union. *Official Journal of the European Communities*, L201, July 2002.
- [Sah98] Dumais; Heckerman; Horvitz; Sahami. A bayesian approach to filtering junk e-mail. *AAAI Workshop on Learning for Text Categorization*, July 1998.
- [Sch03] Schneider. A comparison of event models for naive bayes anti-spam e-mail filtering. In *Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 307–314, Budapest, Hungary, April 2003.
- [See04] Seewald. Combining bayesian and rule score learning: Automated tuning for spamassassin, 2004.
- [Sta00] Androutopoulos; Paliouras; Karkaletsis; Sakkis; Spyropoulos; Stamatopoulos. Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. In *Proceedings of the workshop “Machine Learning and Textual Information Access”*, pages 1–13, Lyon, France, September 2000.
- [Tho91] Cover; Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

- [vN04] Garcia; Hoepman; van Nieuwenhuizen. Spam filter analysis. In *Proceedings of 19th IFIP International Information Security Conference, WCC2004-SEC*, Toulouse, France, August 2004.
- [Yer04] Yerazunis. The spam filtering plateau at 99.9% accuracy and how to get past it. In *MIT Spam Conference*, 2004.