Roskapostin estäminen

Harri Sundström

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tietojenkäsittelyoppi
Pro gradu -tutkielma
Ohjaaja: Erkki Mäkinen
Tammikuu 2008

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tietojenkäsittelyoppi
Sundström Harri: Roskapostin estäminen
Pro gradu -tutkielma, 58 sivua
Tammikuu 2008

Roskaposti on levinnyt räjähdysmäisesti viime vuosina, ja se tullee aiheuttamaan entistä enemmän ongelmia sähköpostin käyttäjille. Roskapostin estämiseksi on olemassa useita keinoja. Tässä tutkimuksessa on esitelty sisältöpohjaisten, roskapostin tunnistamiseen perustuvien suodatus-menetelmien periaatteita sekä sähköpostinlähettäjän luotettavuuteen ja maineeseen perustuvia estomenetelmiä. Työssä tarkastellaan näiden menetelmien mahdollisuuksia ja heikkouksia. Sisältöön perustuvat suodat-timien menetelmät on jaettu sisältöperustaisiin, tilastollisiin menetelmiin perustuviin, heuristisiin ja tarkistenumeroperustaisiin suodattimiin. Sähkö-postin lähettäjän luotettavuuteen ja maineeseen perustuvat menetelmät ovat musta- ja valkealistaus. Tarkastelu osoittaa, että tämänhetkiset sisältö-perustaiset suodattimet eivät kykene ratkaisemaan roskapostiongelmaa täydellisesti. Lähettäjän maineeseen perustuvat estomenetelmät toimivat huomattavasti sisältöperustaisia paremmin. Tällä hetkellä paras tulos saadaan yhdistämällä useita eri menetelmiä. Tulevaisuudessa tulee kehittää sähköposti-järjestelmä, jossa roskapostin lähetys on tehty mahdottomaksi.

Avainsanat ja -sanonnat: Roskaposti, roskapostin esto, sähköpostin suodatus, spam, sähköposti.

Sisällys

1. Johdanto

Sähköpostin käyttö ihmisten välisessä kommunikaatiossa on yleistynyt nopeasti viimeisinä vuosikymmeninä. Sähköposti on korvannut suurelta osin perinteisen kirjepostin niin yksityisten ihmisten välillä kuin yritysmaailmassa. Sähköpostin yleistyminen sekä sen helppo ja edullinen käyttö on houkutellut erinäisiä tahoja hyödyntämään tätä mediaa saadakseen viestinsä levitettyä nopeasti mahdollisemman monelle ihmiselle. Tätä ei-toivottua sähköpostia nimitetään roskapostiksi. Roskapostien pääasiallinen tarkoitus on hankkia rahaa niiden lähettäjille sekä viime kädessä tuotetta tai palvelua markkinoiville henkilöille. Roskaposti on muodostunut valtavaksi ongelmaksi ja suuri osa internetin välitysresursseista käytetään roskapostien välittämiseen. Tällä hetkellä jokainen sähköpostipalveluntarjoaja on pakotettu käyttämään jotakin roskapostin estomenetelmää. Sähköposti ei ole enää niin vapaa kuin miksi se alun perin suunniteltiin.

Lokakuussa 2006 sähköposti täytti 25 vuotta. Nykyisen kaltainen sähköpostijärjestelmä sai alkunsa kun SMTP:n (Simple Mail Transfer Protocol) kehitys alkoi. ARPANET (Advanced Research Projects Agency Network) oli ensimmäisiä verkkojärjestelmiä, jonka halutuin sovellus oli juuri sähköposti. Sähköpostin ja internetin kehittäjien, alan pioneerien, pääasiallinen tarkoitus on ollut kehittää toimiva internet. Alkuperäistä SMTP:tä suunniteltaessa ei kenellekään tullut mieleen, että joku lähettäisi sähköpostia tuntemattomalle ihmiselle [Robertson, 2006].

Perinteiseen postiin ilmoitus postiluukussa ”ei mainoksia kiitos” saattaa vähentää mainosten määrää, mutta sähköpostijärjestelmään ei tällaista mahdollisuutta ole olemassa. Sähköpostin käyttäjän onkin vaikea, ellei mahdoton, estää ei toivottuja viestejä täyttämästä sähköpostiaan. Käytännössä roskapostin estämisestä huolehtii sähköpostipalvelun tarjoaja.

Yleistä tietoa roskapostin estomenetelmistä on hyvin saatavilla, mutta estopalveluita tai ohjelmia myyviltä yrityksiltä ei ole mahdollista saada tietoa tuotteiden toiminnasta. Vastaavasti yritykset eivät halua kertoa tarkasti, miten heidän roskapostiongelmansa on ratkaistu. Vastaukset auttaisivat myös roskapostin lähettäjiä. Kyseessä on kilpajuoksu uusien roskapostin levittämiskeinojen ja roskapostin estomenetelmien välillä.

Tässä tutkielmassa käyn läpi erilaisia roskapostin estomenetelmiä ja arvioin niiden toteutusten kestävyyttä tulevaisuudessa. Keskityn erityisesti sisältöpohjaisten, roskapostin tunnistamiseen perustuvien suodatusmenetelmiin sekä lähettäjän luotettavuuteen perustuviin, verkko-osoitteiden tai verkkotunnuksiin tarkasteluun perustuviin estomenetelmiin. Päätavoitteena on estää sähköpostin käyttäjää saamasta postia, jota hän ei halua.

2. Sähköpostin ja roskapostin taustaa

Sähköposti on ollut olemassa jo ennen tietoverkkoja ja ARPANET:iä. Sähköpostin varhaisemmat versiot toimivat kuten tiedostokansiot. Postinlähettäjä tallensi viestinsä tiedostona tiettyyn kansioon, josta vastaanottaja huomasi sen tullessaan omalle päätteelleen. Posti toimi siis vastaavasti kuin kollegan pöydälle jätetty muistilappu. Myöhemmin kehitettiin ohjelmia, jotka tekivät saman asian, mutta kuitenkin postia lähetettiin vain saman tietokoneen käyttäjien välillä.

2.1. Internetin ja sähköpostin historiaa

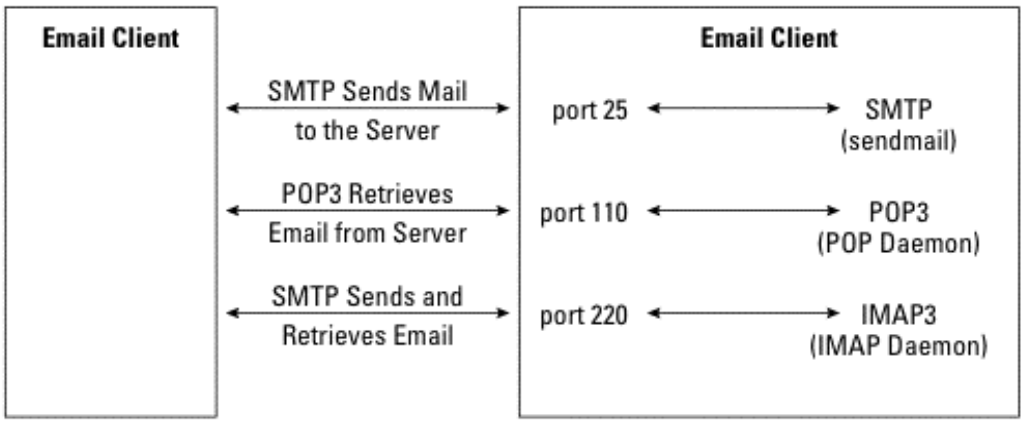
Ensimmäiset tietoverkot olivat pääteverkkoja, joissa keskustietokoneeseen oli kytkettyinä useita päätteitä. Internetin kaltaisten tietoverkkojen katsotaan syntyneen, kun Yhdysvalloissa perustettiin ARPA (Advanced Research Projects Agency) sotilaallista tutkimusta varten vastineena Neuvostoliiton avaruus-ohjelmalle, Sputnikin laukaisun innoittamana. ARPA alkoi kehittää menetelmää ja standardia, jolla voitiin kytkeä useita tietokoneita toisiinsa puhelinlinjoja hyväksikäyttäen. Vuoden 1960 lopulla syntynyt ARPANET levisi ensin vain yliopistojen ja tutkimuslaitosten väliseksi verkoksi, josta nopeasti halventuvan teknologian takia se laajeni kotimaassaan. Pian siihen liittyi myös runsaasti ulkomaisia solmuja mm. Euroopasta. ARPANET suljettiin vuonna 1990, kun internet korvasi sen. ARPANET oli siis internetin varhaisin kehitysmuoto [Hauben, 2007]. Ensimmäisen internetin prototyypin teki sveitsiläinen Tim Berners-Lee, joka kehitti avoimia ja joustavia standardeja informaation jakamiseksi tietokoneverkkoissa. Standardit koskivat niin selaimia kuin tiedon muotoilua internet sivuilla. Nykyään länsimaissa internet kuuluu itsestään selväksi osaksi informaatioyhteiskuntaa.

Sähköposti, sellaisena kun sen nyt tunnemme, keksittiin vuonna 1972. Ray Tomlinson, joka toimi yhtenä ARPANET:in urakoitsijoista valitsi @-symbolin ilmaisemaan tietokoneelta toiselle välitettyä viestiä. Vuonna 1974 sähköposti oli ARPANET:in tärkein ja käytetyin sovellus. Sähköpostin käyttäjien määrä oli tuolloin useita satoja. Tänään sähköpostin käyttäjiä arvioidaan olevan noin 600 miljoonaa [Peter, 2004].

Sähköpostia välitettäessä tietokoneelta toiselle tarvitaan yhteinen käytäntö, jotta koneet pystyvät keskustelemaan toistensa kanssa. Nyt yleisesti käytössä oleva tapa on erittäin yksinkertainen eikä mitenkään varmista, onko sähköpostin lähettäjä se henkilö, joka hän väittää olevansa. Sähköposti-osoitteen väärentäminen oli ja on edelleenkin helppoa. Roskapostin lähettäjät käyttävät juuri näitä standardin heikkouksia hyväkseen hankkiessaan sähköpostiosoitteita ja lähettäessään roskapostia jonkun muun nimellä.

2.2. Sähköpostin yhteyskäytännöt

Sähköpostin välityksessä käytetään useita erilaisia yhteyskäytäntöjä (protocol) (kuva 1). Yhteyskäytäntö on standardi, joka määrittelee laitteiden tai ohjelmien välisen kommunikoinnin. Tyypillisesti SMTP lähettää internetsähköpostin palvelimien välillä. Vastaanottavassa päässä voidaan käyttää useita eri sähköpostipalvelimia, jotka käyttävät joko POP:ia (Post Office Protocol) tai IMAP:ia (Internet Message Access Protocol) [Cole, 2005].



Kuva 1. Eri yhteyskäytäntöjen käyttämät portit sähköpostipalvelimien välisessä liikenteessä [Cole, 2005].

2.2.1. SMTP

Ensimmäinen merkittävä internetsähköpostistandardi oli ja on yhä edelleen SMTP. Muotorakenne (syntax) perustuu edelleen RFC (Request for Comments) 822:n määrittelemään standardiin. RFC:t ovat IETF-organisaation (Internet Engineering Task Force) julkaisemia internetiä koskevia määritelmiä. RFC 822:n mukainen sähköposti koostuu kahdesta osiosta, sähköpostin otsakeosasta (message header) ja sähköpostin runko-osasta (message body), jotka erotetaan toisistaan niin kutsutulla nollarivillä (null line 2 kpl CR, LF). Viestin otsakkeet voivat tulla missä tahansa järjestyksessä. Sähköpostipalvelimet, eli lähettävä ja vastaanottava palvelin, pystyvät näiden otsakkeiden avulla määrittelemään, missä kohtaa viestiä on tietoja sähköpostin lähettäjästä, sähköpostin vastaanottajasta sekä mistä viesti koostuu.

Otsake koostuu seuraavista osista [Hughes, 1998]:

- **Received**: vapaavalintainen osa, jonka järjestelmä luo viestin edetessä.
- **From**: pakollinen osa, jonka muodostaa yleensä lähettävä sähköpostiohjelma. Jos sähköpostinlähettäjän sallitaan vaikuttaa

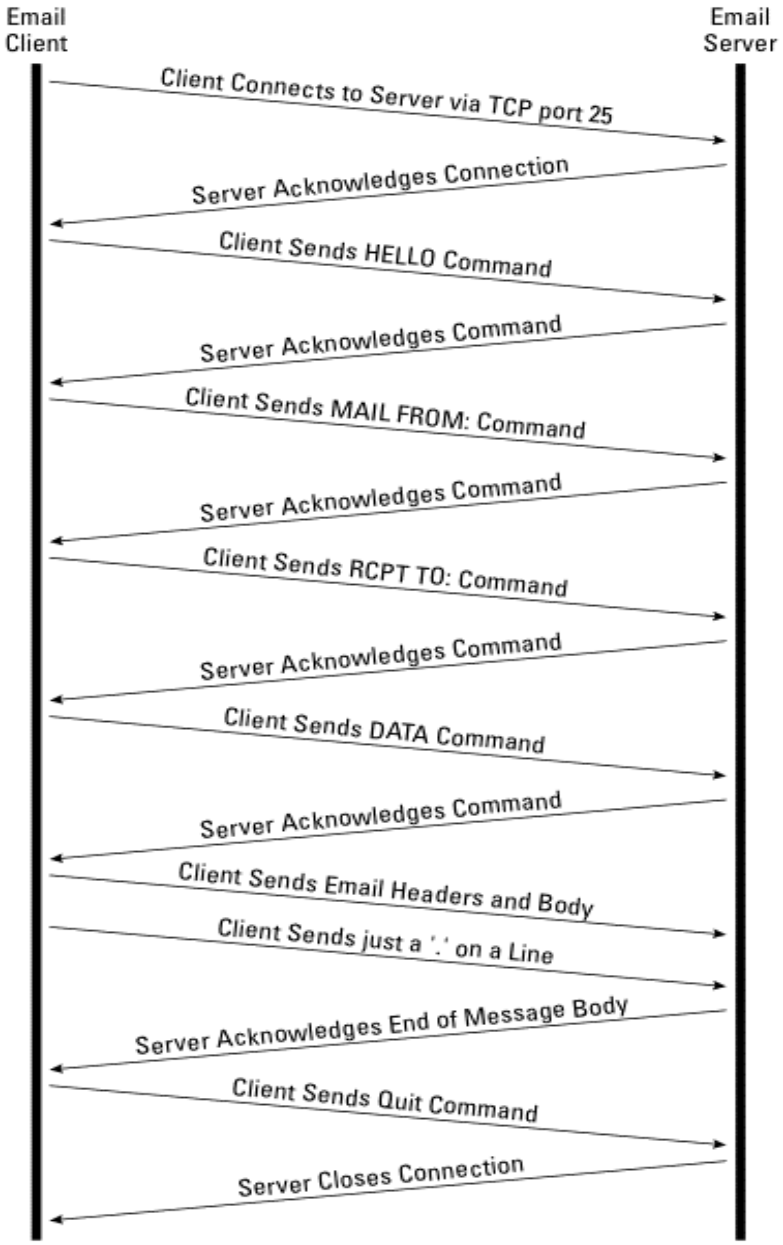
tähän kenttään, on lähettäjän helppo väärentää lähettäjän osoite ja siten välttyä esim. roskapostin lähettäjän vastuilta.

- **Sender**: vapaaehtoinen osa, joka kertoo oikean lähettäjän sähköpostiosoitteen. Esimerkiksi sihteeri voi lähettää sähköpostia esimiehensä nimissä.
- **Reply-To**: vapaaehtoinen osa, joka kertoo mihin osoitteeseen mahdolliset vastaukset tai sähköpostinlukukuittaukset palautetaan.
- **Return-Path**: vapaaehtoinen ja vaihtoehtoinen kenttä edelliselle. Molempia kenttiä ei pidä olla samassa viestissä.
- **Date**: pakollinen kenttä, joka kertoo, koska viesti on laadittu.
- **To**: viestin vastaanottajan tai vastaanottajien sähköpostiosoitteet.
- **CC:** (Carbon Copy) vastaava kuin edellinen, mutta joissakin sähköpostijärjestelmissä vaatii vastauksessa ”kaikille” (reply to all), jotta vastaus toimitetaan myös tässä listassa oleville henkilöille.
- **Bcc:** (Blind Carbon Copy) vastaava kuin edellinen, mutta muut vastaanottajat eivät näe tässä kentässä olevia osoitteita.
- **Message-ID**: pakollinen kenttä, joka luodaan sähköpostin lähettäjä koneen toimesta. Tarkoitettu yksilöimään sähköposti ja vain koneen luettavaksi.
- **Subject**: viestin ydin, kertoo mistä viestissä on kysymys. Vastaukseen sähköpostiohjelmat yleensä kopioivat ”Re:” vastaus tai ”Fw:” tiedoksi, vaikkakaan sitä ei RFC 822:ssa ei ole määritelty.

Muutama muu otsake on myös määritelty RFC 822:ssa, mutta niitä ei juurikaan käytetä. Kaikkia virallisesti määrittelemättömiä otsakkeita tulee edeltää ”X-”-merkintä, esim. ”X-Sender”.

Kuva 2 esittää tyypillistä SMTP-keskustelua palvelimien välillä. Yhteyden-ottopyynnön jälkeen ilmoittaa lähettävä puoli komennolla HELO lähettäjän verkko-osoitteen, jonka jälkeen MAIL FROM kertoo sähköpostin lähettäjän

sähköpostiosoitteen ja RCPT TO vastaavasti vastaanottajan sähköpostiosoitteen. Seuraa DATA-komento, jonka jälkeen edellä määritellyn kaltainen SMTP:n määrittelemä sähköpostin otsakeosa (kuvassa ”Email headers”) ja samassa varsinainen sähköpostiviesti (kuvassa ”body”). Nollarivi päättää datan lähettämisen ja seuraa yhteyden katkaiseminen.



Kuva 2. Kaaviokuva tyypillisestä SMTP-keskustelusta lähettävän ja vastaanottavan sähköpostipalvelimen välillä. [Cole, 2005]

SMTP:n uudelleenlähetys (SMPT relay) oli ensimmäisiä roskapostin lähetysmenetelmiä, jossa todellinen lähettäjä pystyttiin piilottamaan. Menetelmässä käytettiin hyväksi SMTP:n heikkoutta tunnistaa sähköpostin lähetyspyynnön alkuperä. Normaalitilanteessa palveluntarjoaja sallii vain omien asiakkaidensa lähettää viestejä SMTP-palvelimen kautta verkkoon. Ongelma syntyy, kun SMTP-palvelin suostuu lähettämään kenen tahansa toimittaman sähköpostin eteenpäin. SMTP-palvelimesta tulee tällöin avoin sähköpostipalvelin. Ensimmäisissä SMTP-versioissa kaikki hakasulkeissa olevat sähköpostin lähettäjät, MAIL FROM: <sender@domain.com>, hyväksyttiin automaattisesti. Myöhemmissä SMTP-palvelimissa ainoastaan paikallisesta osoitteesta tai erikseen määritellystä sähköpostiosoitteesta voitiin lähettää välityspyyntöjä SMTP-palvelimelle. Tämän jälkeen SMTP:stä on tehty turvallisempia versioita ja myös löydetty uusia heikkouksia. Nykyään tämä menetelmä ei ole suosittu, sillä tällaisten palvelimien IP-osoitteet päivittyvät nopeasti reaaliaikaisille mustille listoille. Mustilla listoilla tarkoitetaan tunnettuja roskapostin lähettäjien IP-osoitteiden listaa, mustat listat on esitelty tarkemmin kohdassa 4.2. [Spammer-X, 2004].

2.2.1.1 MIME

MIME (Multipurpose Internet Mail Extension) täydentää SMTP:ssä olevia puutteita. SMPT:n määrittämä sähköposti hyväksyy ainoastaan ASCII (American Standard Code for Information Interchange) kirjaimia sähköpostissa. Binääritiedostoja ei sinällään ole voinut liittää sähköposteihin. Tätä on pystytty kiertämään binääritiedostoja koodaamalla ASCII-tekstiksi, esim. uuencode on yksi tällainen koodaustapa.

MIME:n täydentämiä sähköpostin otsakkeita ovat mm.:

- **text/plain:** Kertoo kirjainten tulkitsemiseen käytettävän merkkien koodaustavan. Esim. US-ASCII tai ISO-8859-1.

- **Text/enriched**: Viestissä on mukana muotoilukoodeja, joilla määritellään kirjasintyyppejä ja niiden värejä sekä kokoa.
- **Text/html:** Viestissä olevat muotoilukoodit noudattavat internetistä tuttua HTML-standardia (HyperText Markup Language).
- **Audio/<tyyppi>:** Kertoo, miten viestiin liitetty ääni on koodattu.
- **Image/<tyyppi>:** Kertoo kuvan koodaustavan. Esim. JPEG, GIF, TIFF.
- **Video/<tyyppi>:** Viestissä olevan videon formaatti.
- **Application/<tyyppi>:** Ilmoittaa viestin liitteenä olevan tiedoston vaatiman sovelluksen, esim. Word tai Excel.

Näillä MIME:n otsakemäärittelyillä voidaan sähköpostiin sisällyttää kirjaimia, jotka eivät ole ASCII-määritelmän mukaisia, esim. skandinaaviset merkit, erilaisia kirjasintyyppejä, tekstin värejä ja muita tekstin muotoiluja sekä liitetiedostoja. Iso sähköposti voidaan jakaa osiin, jolloin sähköpostipalvelin huolehtii automaattisesti palasten kokoamisesta jälleen yhtenäiseksi, isoksi sähköpostiksi. Osa sähköpostin sisällöstä voidaan koota ulkoisista lähteistä, kuten internetistä. Useita sähköpostin komponentteja voidaan esittää samanaikaisesti, kuten tekstiä, liikkuvaa kuvaa ja ääntä.

MIME:n avulla voidaan edellä mainituista, osasähköposteista muodostaa kokonaisuus useilla tavoilla, mutta tämä ei näy käyttäjälle, sillä viestin kokoamisesta huolehtii sähköpostipalvelin. Tästä syystä näitä otsakemäärittelyjä ei käsitellä tässä tutkielmassa tarkemmin [Hughes, 1998].

2.2.1.2 SendMail

SendMail on yleisin sähköpostin välitysohjelmisto, jonka suosio perustuu osittain sen asemaan Unix-järjestelmien oletusvälitysohjelmistona. SendMailin kehitys alkoi jo ARPANET:in aikana. Ohjelman kehittäjä, Eric Allman, suunnitteli 1979 ARPANET:iin ohjelman ”DeliverMail” joka hoiti sähköpostien välitystä palvelimien välillä. Allman kehitti ohjelmaa edelleen, ja sen

kehittyneempi versio julkaistiin nimellä SendMail vuonna 1983. Sendmail on joustava ohjelma, sillä se tukee useita sähköpostin siirto- ja välitysmuotoja, mukaan lukien SMTP:n ja sen laajennukset [Cole, 2005].

2.2.2. POP

POP:ia (Post Office Protocol) käytetään sähköpostin noutamiseen palvelimelta. Viimeisin versio on POP3, joka on yhteensopiva lähes kaikkien sähköposti-palvelimien kanssa. POP3 ei tue sähköpostin lähettämistä ja on suhteellisen harvinainen, sillä siinä on useita turvallisuusriskejä. POP3 käyttää yksinkertaisia tekstimuotoisia komentoja kommunikoidessaan palvelimen kanssa. Perusversiossa käyttäjätunnukset ja salasanat siirtyvät selväkielisinä, kryptaamatta. Autentikointiin käytetään USER (käyttäjän nimi) ja PASS (salasana) komentoja. Jos palvelin lähettää USER-komentoon hyväksynnän, se tarkoittaa, että on olemassa annetun niminen käyttäjä. Seuraavaksi tarvitaan PASS-komennon hyväksyminen annetulle käyttäjälle. Kun molemmat komennot ovat erikseen annettu ja hyväksytty, niin autentikointi on suoritettu [Cole, 2005].

2.2.3. IMAP

IMAP (Internet Message Access Protocol) on menetelmä, jolla voidaan välittää sähköpostia etäpalvelimelta (remote server). Se on suunniteltu etäkäyttöön, jolloin sähköposti pidetään etäpalvelimella ja sitä luetaan useasta eri kohteesta esim. kotoa, toimistosta ja matkapuhelimella. IMAP, toisin kuin POP3, tukee useita kansiota palvelimella, ja se on sopiva myös hitaisiin verkkoyhteyksiin, sillä kaikkia viestejä ei tarvitse ladata etälaitteeseen niiden pysyessä palvelimella. IMAP tukee myös osin MIME:ä [Cole, 2005].

2.3. Roskaposti

Roskapostiksi eli spammiksi katsotaan sähköposti, jota käyttäjä ei ole pyytänyt tai halunnut saada ja joka on lähetty umpimähkään, suoraan tai epäsuoraan vastaanottajalle, jolla ei ole olemassa mitään kanssakäymistä lähettäjän kanssa [Cormack and Lynam, 2006]. Vuonna 2001 on laskettu kaikesta sähköpostista olleen noin 5 % roskapostia [Barracuda Networks, 2007]. McCarthyn [2005] mukaan vuonna 2003 kaikista lähetetyistä sähköposteista 30% oli roskaposteja, seuraavana vuonna määrä oli jo kasvanut 70 %:iin. Vuonna 2006 roskapostien määrän on laskettu olevan jopa 80 % kaikista sähköposteista. Tätä tutkimusta aloittaessani 2007 oli roskapostin määrän arvioitu olevan noin 86 % kaikesta sähköpostista. Barracuda Networks [2007], joka on johtava turvallisuusasiantuntija niin sähköpostin kuin verkkoturvallisuuden alalla, raportoi vuosittaisessa roskapostiraportissaan 12. joulukuuta 2007 roskapostin osuuden olevan jopa 90 - 95 % kaikesta sähköpostista.

Ensimmäiseksi roskapostiksi voidaan katsoa DEC:in (Digital Equipment Corporation) työntekijän, Gary Thuerkin, lähettämä ilmoitus uuden DEC-tietojärjestelmän esittelytilaisuudesta kaikille ARPANET:in käyttäjille vuonna 1978 (kuva 3). Ensimmäinen roskaposti tavoitti täten 100 % kaikista sähköpostin käyttäjistä, tosin käyttäjiä ei tällöin ARPANET:issä, internetin edeltäjässä, ollut vielä montaa.

Gary oli yksi DEC:n markkinointiosaston työntekijöistä. Hän odotti, että vastaanottajat normaalisti vastaisivat kutsuun. Vastaus ei ollut kuitenkaan toivottu, vaan ensimmäisestä roskapostista tuli suuri kohu ja useat vastaanottajat kommentoivat viestinnän moraalia ja tarkoitusperiä. Kommentteja tuli siinä määrin, että silloinen ARPANET oli suorituskykyjensä rajoissa alkeellisten linjanopeuksien ja palvelimien vähäisten muistimäärien vuoksi [Zdziarski, 2005].

```
Mail-from: DEC-MARLBORO rcvd at 3-May-78 0955-PDT
Date:  1 May 1978 1233-EDT
From: THUERK at DEC-MARLBORO
Subject: ADRIAN@SRI-KL

DIGITAL WILL BE GIVING A PRODUCT PRESENTATION OF THE NEWEST MEMBERS OF THE
DECSYSTEM-20 FAMILY; THE DECSYSTEM-2020, 2020T, 2060, AND 2060T.  THE
DECSYSTEM-20 FAMILY OF COMPUTERS HAS EVOLVED FROM THE TENEX OPERATING SYSTEM
AND THE DECSYSTEM-10 <PDP-10> COMPUTER ARCHITECTURE.  BOTH THE DECSYSTEM-2060T
AND 2020T OFFER FULL ARPANET SUPPORT UNDER THE TOPS-20 OPERATING SYSTEM.
THE DECSYSTEM-2060 IS AN UPWARD EXTENSION OF THE CURRENT DECSYSTEM 2040
AND 2050 FAMILY. THE DECSYSTEM-2020 IS A NEW LOW END MEMBER OF THE
DECSYSTEM-20 FAMILY AND FULLY SOFTWARE COMPATIBLE WITH ALL OF THE OTHER
DECSYSTEM-20 MODELS.

WE INVITE YOU TO COME SEE THE 2020 AND HEAR ABOUT THE DECSYSTEM-20 FAMILY
AT THE TWO PRODUCT PRESENTATIONS WE WILL BE GIVING IN CALIFORNIA THIS
MONTH.  THE LOCATIONS WILL BE:

               TUESDAY, MAY 9, 1978 - 2 PM
                   HYATT HOUSE (NEAR THE L.A. AIRPORT)
                   LOS ANGELES, CA

               THURSDAY, MAY 11, 1978 - 2 PM
                   DUNFEY'S ROYAL COACH
                   SAN MATEO, CA
                   (4 MILES SOUTH OF S.F. AIRPORT AT BAYSHORE, RT 101 AND RT 92)

A 2020 WILL BE THERE FOR YOU TO VIEW. ALSO TERMINALS ON-LINE TO OTHER
DECSYSTEM-20 SYSTEMS THROUGH THE ARPANET. IF YOU ARE UNABLE TO ATTEND,
PLEASE FEEL FREE TO CONTACT THE NEAREST DEC OFFICE
FOR MORE INFORMATION ABOUT THE EXCITING DECSYSTEM-20 FAMILY.
```

Kuva 3. Ensimmäinen ei kaupallinen roskaposti [Zdziarski, 2005]

Ensimmäinen kaupallinen roskaposti, joka keräsi laajempaa huomiota oli aviopari Canterin ja Siegelin organisoima roskaposti. Aviopari palkkasi tietokoneohjelmoijan kirjoittamaan ohjelman, joka lähettää mainoksen kaikille tunnetuille uutisryhmille (kuva 4). Tästä syntyi ensimmäinen massasähköpostitusohjelma, jollaisia roskapostin lähettäjät hyödyntävät. Ihmisten reaktiot olivat odotetun kielteiset. Canter ja Siegel (C&S) olivat aikansa kumouksellisimmat ja vihatuimmat suoramarkkinoijat [Zdziarski, 2005].

```
From: Laurence Canter (nike@indirect.com)
Subject: Green Card Lottery- Final One?
Date: 1994-04-12 00:40:42 PST


Green Card Lottery 1994 May Be The Last One!
THE DEADLINE HAS BEEN ANNOUNCED.

The Green Card Lottery is a completely legal program giving away a certain
annual allotment of Green Cards to persons born in certain countries. The
lottery program was scheduled to continue on a permanent basis.  However,
recently, Senator Alan J Simpson introduced a bill into the U. S. Congress
which could end any future lotteries. THE 1994 LOTTERY IS SCHEDULED TO TAKE
PLACE SOON, BUT IT MAY BE THE VERY LAST ONE.

PERSONS BORN IN MOST COUNTRIES QUALIFY, MANY FOR FIRST TIME.

The only countries NOT qualifying  are: Mexico; India; P.R. China; Taiwan,
Philippines, North Korea, Canada, United Kingdom (except Northern Ireland),
Jamaica, Domican Republic, El Salvador and Vietnam.

Lottery registration will take place soon.  55,000 Green Cards will be
given to those who register correctly.  NO JOB IS REQUIRED.

THERE IS A STRICT JUNE DEADLINE. THE TIME TO START IS NOW!!

For FREE information via Email, send request to cslaw at indirect.com

--
*****************************************************************
Canter & Siegel, Immigration Attorneys
3333 E Camelback Road, Ste 250, Phoenix AZ  85018  USA
cslaw at indirect.com   telephone (602)661-3911  Fax (602) 451-7617
```

Kuva 4. Ensimmäinen kaupallinen roskaposti [Zdziarski, 2005]

2.4. Ensimmäisiä roskapostin estomenetelmiä

Automaattisten roskapostinestomenetelmien kehittäminen tuli ajankohtaiseksi, kun ensimmäiset massaroskapostin lähetysmenetelmät yleistyivät 1994 jälkeen. Tavoitteena oli hillitä yritysten tarvitsemaa internetyhteyden kaistanleveyttä, palvelinkapasiteettia, muistitilaa sekä hallinnointiin käytettyä henkilökuntaa.

Ensimmäiset roskapostin suodattimet eivät varsinaisesti olleet suodattimia vaan yksinkertaisesti tutkivat sanoja otsakkeessa, sisällössä ja lähettäjän tai vastaanottajan kentissä. Esimerkiksi ”Free Trial”, ”Call Now” tai vastaavaa sisältävä posti voitiin suodattaa roskapostiksi. Vastaavasti, jos vastaanottajan kentässä oli tuntemattomia nimiä tai useita nimiä, voitiin tätä hyödyntää

suodatuksessa. Suomenkielisiä roskaposteja ei juuri esiintynyt alkuvaiheessa, joten tavalliset suomalaiset käyttäjät saattoivat luokitella kaikki englanninkieliset sähköpostit roskapostiksi. Itse työskentelin tuolloin yrityksessä, jonka kommunikointikieli oli englanti, joten minulle ei tuosta säännöstä ollut hyötyä. Yrityksessä käytössä olevassa sähköpostijärjestelmissä (MS mail / MS Outlook) oli mahdollisuus itse asettaa sääntöjä tulevan postin sisällöstä ja lähettäjästä löytyvien sanojen perusteella. Tulevasta sähköpostista löytyvien sanojen perusteella koetin suodattaa omaa kasvavaa roskapostimääräänі. Aina, kun uusia roskaposteja pääsi suodattimen lävitse, lisäsin uusia läpitulleesta postista löytyneitä sanoja suodattimeen, mutta roskapostin määrä vain lisääntyi ja sanalista alkoi olla vaikea hallita.

3. Roskapostin olemassaolosta

Roskapostin lähettäminen on useissa maissa, kuten Suomessa, pääasiassa laitonta. Tämän vuoksi roskapostit lähetetäänkin nykyään jonkun muun sähköpostia ja verkkopalveluita käyttäen. Roskapostiongelma ei ole ainoastaan siinä, että sähköpostit ovat vastaanottajalle hyödyttömiä, ne saattavat olla myös haitallisia ja sisältää viruksia. Vuonna 2003 tilanne muuttui ratkaisevasti, kun roskapostit sisälsivät viruksia, jotka varsinkin Windows-ympäristön tietoaukkoja hyväksikäyttäen muuttivat käyttäjän laitteiston avoimeksi palvelimeksi roskapostin lähettäjän käyttöön. Toisin sanoen roskapostissa olevat virukset auttoivat roskapostien levittämistä. Vuonna 2003 sähköpostista levinneiden virusten ja kaikkien tietokonevirusten lukumäärien suhdeluku oli 1:33, vuotta myöhemmin suhdeluku oli jo 1:16 [McCarthy, 2005].

Kaikilla internetin käyttäjillä on voimakas mielipide roskapostista. Suurin osa on ehdottomasti roskapostia vastaan. Yksikään internetpalvelujen tarjoaja (ISP) ei halua roskapostin lähtevän juuri heidän verkostaan. Useimmissa maissa roskapostien lähettäminen on myös lailla kielletty, joten internetpalvelujen tarjoajat saattavat myös joutua edesvastuuseen, jos roskapostia lähetetään heidän verkostaan. Roskapostin lähettäminen, samoin kuin tietokonevirusten tehtailu, on kilpajuoksua roskapostin levityskeinojen ja roskapostin torjumiskeinojen välillä. Kun roskapostin torjumismenetelmät kehittyvät torjumaan roskaposteja, levittäjien pitää kehittää uusia menetelmiä, joita aiemmat torjumiskeinot eivät tavoita.

3.1. Lähettäjien motiivit

Mainoksen ja mainostamisen merkitys myynnille on aina ollut merkittävä. Mainostaminen yleisesti kasvattaa tietoisuutta tuotteesta tai palvelusta. Perinteiset mediat kuten lehdet, TV ja radio ovat olleet välineinä uusille markkinointikampanjoille. Markkinoijalle on tärkeää, että mahdollisimman

moni näkee tai kuulee tuotteesta. Vielä tuottoisampaa on, jos mainostettava tuote suunnataan oikealle kohderyhmälle, eli potentiaalisille ostajille.

TV-mainoksien kustannukset ovat kuitenkin suuria. Kun internetin ja sähköpostin käyttö alkoi yleistyä 1990-luvulla, niin huomattiin, että tällä uudella medialla voidaan mainostaa murto-osalla TV-mainosten kustannuksista. Mainoksia alettiin levittää sähköpostitse. Ensimmäiset mainostetut tuotteet tai palvelut olivat pornografisia. Pornografiset sivustot olivat internetin alkuaikoina kaikkein suosituimpia, joten oli luonnollista, että näitä alettiin myös mainostaa. Sivuston tekijät palkkasivat jonkun tekemään ”mainoskampanjan” eli lähettämään massasähköposteja valittuihin sähköpostiosoitteisiin.

Roskapostin lähettäjille maksetaan tulospalkkioita. Jokainen, joka rekisteröityy asiakkaaksi klikattuaan roskapostissa saamaansa linkkiä, tuottaa roskapostin lähettäjälle osuuden rekisteröinnistä. Oletetaan, että roskapostin lähettäjä saa euron jokaisesta rekisteröinnistä. Kun yhtä roskapostia lähetetään miljoona kappaletta ja 20 % kaikista vastaanottajista klikkaa linkkiä, joista 5 % rekisteröityy asiakkaaksi, on rekisteröityneitä 10 000. Näin ollen roskapostin lähettäjä on tienannut 10 000 €. Tässä yksinkertaistuksessa ei ole otettuhuomioon roskapostin estomenetelmistä johtuvaa karsintaa, mutta roskapostin alkuaikoina niitä ei juuri ollutkaan. Roskapostin lähettäminen oli ja on edelleen kannattavaa [Spammer-X, 2004].

3.2. Sähköpostiosoitteiden kerääminen

Roskapostia voidaan lähettää manuaalisesti tai koneellisesti. Yleensä jokin ohjelma huolehtii lähettämisestä tiedettyihin sähköpostiosoitteisiin, joita on saatettu hankkia joskus kyseenalaisinkin keinoin. Sähköpostiosoitteet voivat olla myös jonkin toisen ohjelman tuottamia. Jos sähköpostiosoite on internetissä jollain hakukoneella löydettävissä olevalla www-sivulla, on se vapaasti

roskapostien lähettäjien siepattavissa. Samoin jos sähköpostiosoite on muotoa <etunimi>.<sukunimi>@<yhteisö>.fi, on sen muodostaminen ohjelmallisesti suhteellisen helppoa. Kun posti sitten lähetetään hyväksytysti tähän muodostettuun osoitteeseen, on se silloin validi ja sitä voidaan hyödyntää jatkossa. Roskapostia varten osoitteita haetaan varmasti:

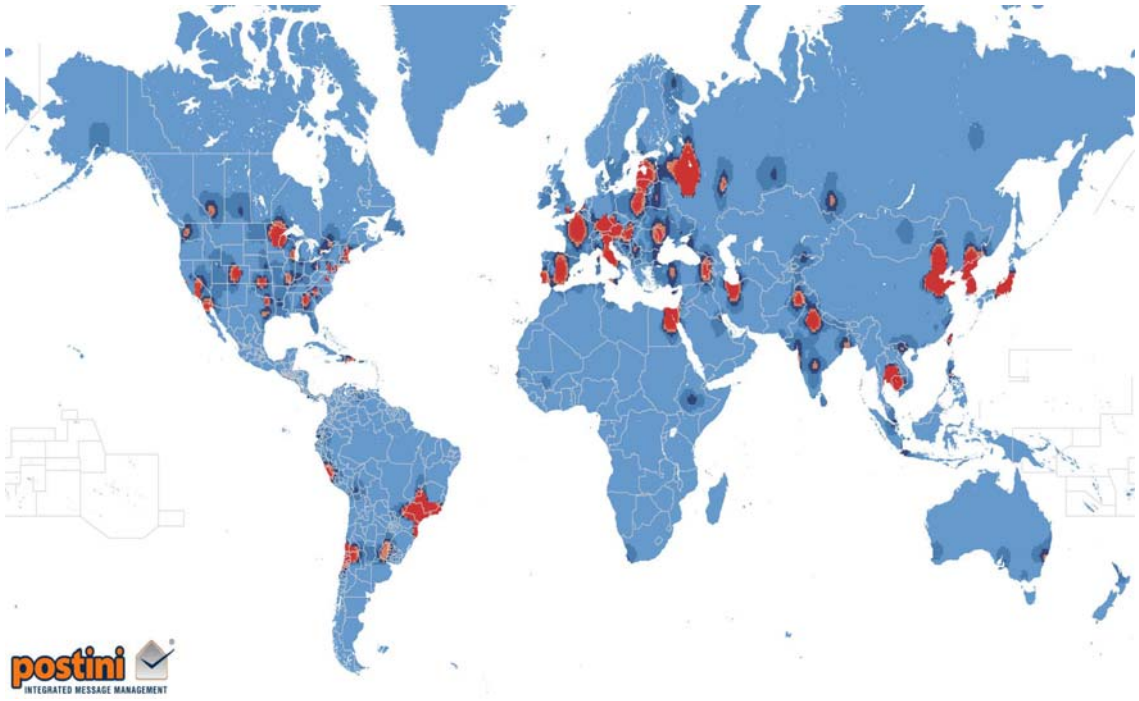
- verkkosivuilta
- vieraskirjoista
- uutisryhmistä
- IRC -palvelimilta
- keskustelupalstoilta (chateista)
- erilaisista avoimista tietokannoista
- sähköpostilistojen arkistoista
- sivujen vieraskirjoista
- blogeista (verkkopäiväkirjoista)
- ylipäätään mistä tahansa tekstimuotoisesta materiaalista, joka löytyy verkosta.

Sähköpostiosoitteita voidaan myös hakea aggressiivisesti mm. menetelmällä, jota kutsutaan DHA:ksi (Directory Harvest Attact). DHA on menetelmä, jolla roskapostin lähettäjät yrittävät kerätä toimivia sähköpostiosoitteita verkko-palvelimilta. Hyökkäys suoritetaan sähköpostipalvelimen hakemistoon, josta koetetaan arvata siellä olevia standardimuotoisia sähköpostiosoitteita. Ohjelma voi joko arpoa satunnaisen kirjainjonon tai muodostaa yhdistelmän yleisistä etu- ja sukunimistä tai niiden alkukirjaimista. Tämän jälkeen ohjelma lisää osoitteeseen sen verkkopalvelimen tunnuksen, jonne hyökkäys suunnataan. Muodostetut osoitteet ovat esimerkiksi muotoa aseerrt@domain.com, pvirtanen@domain.com tai pekka.virtanen@domain.com.

DHA:n tehokkuus perustuu siihen, että SMTP-yhteyden aikana sähköposti-palvelin joko hylkää tai hyväksyy muodostetun sähköpostiosoitteen. Varsinai-

sen DHA-sähköpostin sisältö saattaa olla vaikka vain ”Hello”, sillä sähköpostin lähettäjä ei halua roskapostisuodattimien estää postin saapumista perille. Kaikki hyväksytyt sähköpostiosoitteet roskapostin lähettäjä lisää omaan roskapostituslistaansa. Roskapostin lähettäjät myös myyvät ja vaihtavat listoja toimivista sähköpostiosoitteista keskenään. [Clyman, 2004].

DHA:ta voidaan käyttää suoraan myös roskapostien lähettämiseen, mutta silloin se ei välttämättä saavuta toivottua määrää vastaanottajia. Tämä menetelmä aiheuttaa suunnattoman kuorman sähköpostin välittäjille ja internettiin. Kuvassa 5 nähdään alueet, joista tällä hetkellä suurin osa DHA hyökkäyksistä on lähtöisin.



Kuva 5. DHA-hyökkäysten maantieteellinen jakauma (punaiset alueet) perustuen IP-osoitteista saatuihin tietoihin [Postini, 2007].

3.3. Virukset ja roskaposti

Viruksilla ja roskaposteilla on ainakin kaksi yhteistä tekijää. Roskapostit saattavat sisältää erilaisiin tarkoituksiin kehitettyjä viruksia ja toisaalta virukset voivat auttaa roskapostin levittämisessä esimerkiksi pitämällä roskapostin lähettäjän henkilöllisyyden salassa.

Duntemannin [2004] mukaan on olemassa kaksi tietokoneohjelmatyyppiä, jotka monistavat itseään, virukset ja madot (worms). Itseänsä monistava tarkoittaa sitä, että ohjelma yrittää levitä toisiin tietokoneisiin kopioimalla itseään ja liittämällä itsensä lailliseen tietokoneohjelmaan. Leviäminen tapahtuu esim. sähköpostin välityksellä. Viruksen toimiessa tietokoneessa on tietokone tällöin saastunut (infected). Virusten ei tarvitse alustaa kovalevyä, tuhota käyttö-järjestelmää, muuttaa internetasetuksia tai korruptoida tiedostoja, kuten vuosia sitten oli tapana. Virusten ei välttämättä tarvitse tehdä mitään muuta kuin levitä koneesta toiseen jollakin uudella haasteellisella tavalla.

Useimmilla viruksilla on jokin tietty tehtävä saastuttamassaan koneessa. Tätä ohjelman osaa kutsutaan viruksen hyötykuormaksi (virus payload). Nykyään virukset toimivat hienostuneemmin ja niiden hyötykuormana saattaa olla muodostaa saastuttamastaan tietokoneesta roskapostien uudelleenlähettäjä (spam relay). Uudelleenlähetyksen tarkoituksena on estää alkuperäistä roskapostin lähettäjää jäämästä kiinni toiminnastaan. Tällaista roskapostin uudelleen lähettäjää kutsutaan roskapostizombieksi. Vuonna 2004 noin 40 % kaikista roskaposteista tuli juuri tällaiselta roskapostizombielta [Duntemann, 2004].

Roskapostizombiet ovat oikeastaan loisia, joiden toimintaedellytyksinä on isäntäkoneen toimiminen mahdollisimman hyvin. Tällöin ei käyttäjä huomaa tätä tietokoneelle aiheutunutta ylimääräistä kuormaa ja poista sitä. Joidenkin virusten on havaittu hakevan myös salasanoja ja sähköpostiosoitteita, joita

virusohjelma sitten toimittaa roskapostin lähettäjien internetosoitteeseen. Virus voi siis toimia käyttäjän tietämättä tietokoneessa kuukausia tai peräti vuosia. Mikä pahinta, tietokone saattaa rikkoa lakia lähettämällä kiellettyä materiaalia tai roskapostia edelleen.

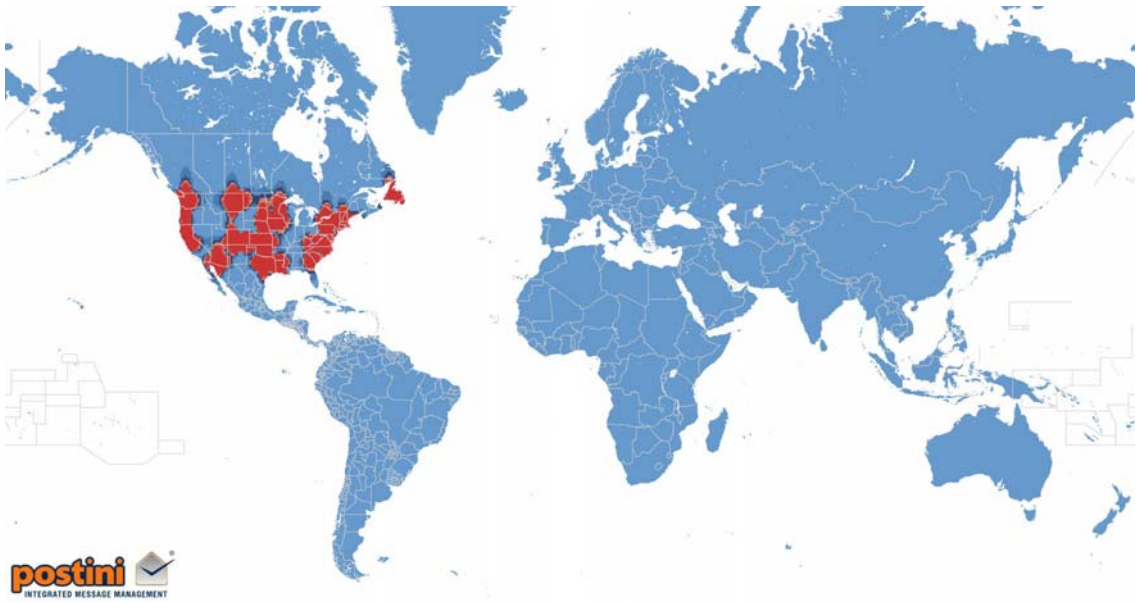
Duntemannin [2004] mukaan on olemassa viisi tapaa saastuttaa tietokoneensa viruksilla:

1. Käynnistää tietokone saastuneelta levykkeeltä, zip-asemalta tai flash-asemalta.
2. Sallia makrojen käyttö toimistosovelluksissa ja käynnistää tällainen sovellus.
3. Sallia ohjelmien asentaminen suoraan selaimesta.
4. Käynnistää saastunut ohjelma, jonka olet saanut internetistä tai ystävältä.
5. Avata saastunut sähköpostin liitetiedosto.

Kaikki vaihtoehdot eivät ole yhtä todennäköisiä. Harva käynnistää tietokonettaan enää levykkeeltä. Sen sijaan kohdat 4 ja 5 ovat suurimmat virusten leviämisväylät. Periaatteessa kysymys on samasta asiasta kuin jos saastunut ajettava tiedosto avataan ja käyttöjärjestelmä käynnistää sen. Sähköpostin liitetiedostot saattavat olla piilotettuja ajettavia tiedostoja [Duntemann, 2004]. Omakohtainen kokemus tällaisesta piilotetusta ajettavasta tiedostosta on eräs sähköposti, jossa tiedoston nimi alkoi ”fun.jpg” ja useamman rivinvaihdon jälkeen päättyi määreeseen ”.exe”. Sähköposti-ikkunassa näkyi vain alku ja liite näytti vaarattomalta jpeg-kuvalta. Klikkaus ei avannut kuvaa, vaan käynnisti ohjelman.

Suurin osa viruksista leviää tahattomasi sähköpostin välityksellä, kun vastaanottajan sähköpostiosoite on tallennettuna lähettäjän tietokoneelle ja virus lähettää itsensä kaikkiin isäntäkoneen tuntemiin sähköpostiositteisiin.

Kuvasta 6 havaitaan, että virusaktiviteettia oli marraskuussa 2007 lähes pelkästään Yhdysvalloissa.



Kuva 6. Virusten lähettäjien sijainti (punaiset alueet) perustuen virusten lähettäjien IP-osoitteisiin [Postini, 2007].

3.4. Käyttäjän toimenpiteet

Mitä vaikeammaksi tehdään roskapostin lähettäjien keinot ansaita rahaa roskapostia lähettämällä, sitä vähemmän roskapostin lähettäjiä tulee mukaan alalle ja edelleen vähemmän roskaposteja myös lähetetään. Tämä on yksi epäsuora tapa estää roskapostia.

Käyttäjän ei tulisi tukea roskapostin lähettämistä. Roskapostiin vastaaminen varmistaa sähköpostiosoitteen joutumisen toimivalle sähköpostilistalle, joita roskapostin lähettäjän käyttävät ja myyvät toisilleen. Roskapostissa saattaa olla linkki tai sähköpostiosoite, josta voidaan kieltää vastaavan sähköpostin lähettäminen jatkossa. Roskapostin ollessa kyseessä on suhtauduttava varauksellisesti myös tämän linkin toimivuuteen. Saattaa olla, että roskapostin lähettäjä varmistaa vain tällä tapaa sähköpostiosoitteen toimivuuden. Samoin

gallupkysely, treffipalveluun liittyminen tai yleensä mihinkään arveluttavaan sähköpostiin vastaaminen ei ole suositeltavaa.

Roskapostissa olevista mainoksista ei myöskään kannata ostaa mitään. Ensinnäkin sähköpostiosoite tulee edellä mainitulla tavalla julki ja sen lisäksi mainostettaviin tuotteisiin ei voi luottaa. Luotettavaltakin kuulostavat tuotteet saattavat olla väärennettyjä tai vähintään ylihintaisia. Roskapostissa olevasta mainoksesta tilatut tuotteet eivät välttämättä koskaan saavu perille. Rekisteröityminen tämän kautta jollekin sivustolle luottokortilla ei välttämättä pääty koskaan. Roskapostin ohjaama luotettavalta kuulostava internetsivusto saattaa olla väärennetty.

Sähköpostin laittamista kotisivuille tai antamista yleensä internetin keskustelupalstoille tms. pitää välttää. Internetiin laitettavassa sähköposti-osoitteessa olisi vältettävä tekstimuotoista osoitetta ja sen sijaan käytettävä esimerkiksi kuvaa tai mainintaa etunimi.sukunimi@domain.fi ja kirjoittaa nimet erikseen. Tällöin sähköpostiosoitetta ei voida ohjelmallisesti etsiä internetistä. Jos sähköpostiosoite kaikesta huolimatta pitää antaa oikeassa muodossa, esimerkiksi nettikauppaan, kannattaa käyttää tilapäistä sähköpostiosoitetta, jonka voi poistaa käytöstä tarvittaessa tai muutaman kuukauden kuluttua, ja luoda sitten uusi tilapäinen osoite seuraavaa käyttökertaa varten.

Sähköpostiosoite kannattaa valita sellaiseksi, jota on vaikea kalastaa aggressiivisella keräysmenetelmällä (DHA). Sähköpostinimeksi kannattaa valita jokin sana tai nimi, joka ei löydy sanakirjasta tai yleisestä nimiluettelosta. Tällainen voisi olla esimerkiksi jokin edellisten yhdistelmä tai väärinkirjoitettuja sanoja kuten cat4me@domain.com, kissatalo@domain.com tai yukka69@domain.com. Täysin satunnaista lyhyehköä merkkijonoa ei kannata käyttää [Duntemann, 2004].

4. Roskapostin estomenetelmiä

Roskapostin estomenetelmiä ja niiden luokitteluja on useita. Roskapostia voidaan ehkäistä mm. keräämällä sallittuja ja ei-sallittuja lähettäjiä (lähettäjän sähköpostiosoite, IP-osoite, palveluntarjoaja), säätämällä roskapostin kieltäviä lakeja, tutkimalla sähköpostin sisältöä ja keräämällä tunnettujen roskapostien ominaisuuksia tarkistenumeroksi. Tässä työssä käsitellään pääasiassa roskapostin estomenetelmiä, jotka perustuvat roskapostin tunnistamiseen ja roskapostin lähettäjän tunnistamiseen.

Roskapostin sisältöön perustuvat estomenetelmät voidaan jakaa neljään eri luokkaan:

- sisältöperustaiseen suodatukseen (content-based filtering)
- tilastollisiin menetelmiin perustuviin, (statistic filtering)
- heuristisiin suodattimiin (heuristic filtering)
- tarkistenumeroperustaisiin suodatusmenetelmiin (checksum-based filtering).

Roskapostin lähettäjän tunnistamiseen perustuvat estomenetelmät voidaan jakaa

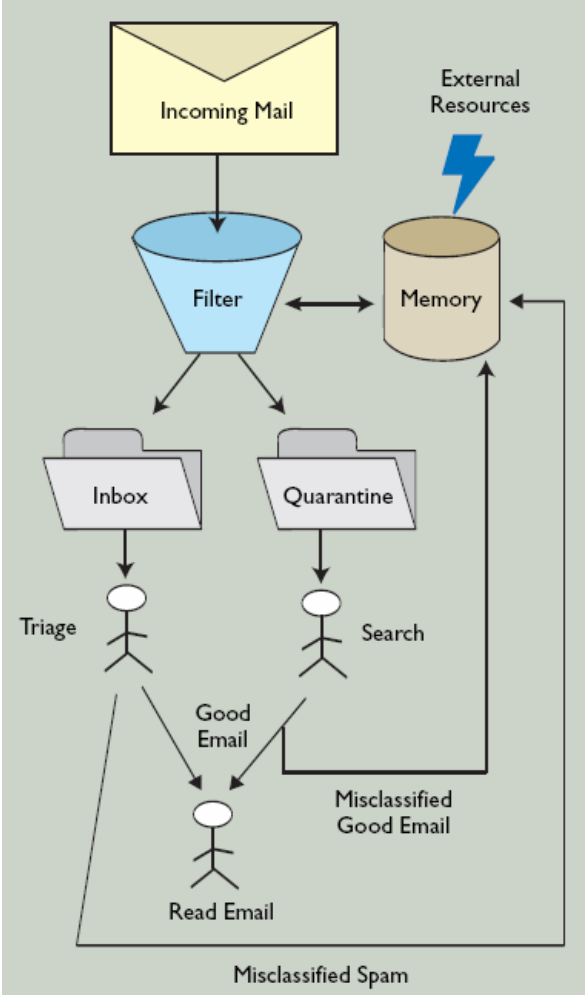
- olemassa oleviin mustiin listoihin (black listing)
- olemassa oleviin valkeisiin listoihin (white listing).

Seuraavissa kohdissa tarkastellaan edellä mainittuja menetelmiä yksityiskohtaisemmin.

4.1. Roskapostin sisältöön perustuvat estomenetelmät

Roskapostin sisältöön perustuvan suodatusprosessin pääperiaatteet on esitetty kuvassa 7. Sähköpostin saapuessa palveluntarjoajan laitteistoon suodatin (filter) lajittelee sähköpostin kahteen kansioon ennen sen toimittamista varsinaiselle sähköpostin saajalle: saapuneisiin (inbox), jota luetaan säännöllisesti, ja karanteenissa oleviin (quarantine), jota tutkitaan määräajoin. Suodatin on tehnyt virheen, jos roskaposti on päätynyt saapuneeet-kansioon (false negative) tai oikea sähköposti on mennyt karanteenissa olevien kansioon (false positive). Suodatinta voidaan parantaa tarkastelemalla näitä virheellisiä suodatustapahtumia ja tehdä muutoksia suodattimeen. Samaa menetelmää käytettäessä voidaan uuteen versioon saada edellistä versiota parempi suodatuskyky [Goodman et al., 2007]. Suodattimen parantaminen vaatii toimenpiteitä joko käyttäjältä tai järjestelmän ylläpitäjältä. Automaattisia, esim. tekoälyyn, perustuvia korjaamismenetelmiä voidaan myös käyttää. Perus-periaate on, että järjestelmä kehittyy roskapostien levittämismenetelmien kehittyessä.

Loogisesti ajatteleva ihminen suodattimena tunnistaisi roskapostin nopeasti ja virheitä ei suodattamisprosessissa juurikaan tapahtuisi. Roskapostin määrästä johtuen ei ihmisvoimien käyttäminen sähköpostin lajitteluun ole järkevää tai edes mahdollista. Haasteena onkin roskapostin tunnistaminen ohjelmallisesti tai muun automatisoidun menetelmän avulla.



Kuva 7. Roskapostin sisältöön perustuvan suodatuksen pääperiaatteet [Goodman et al., 2007].

4.1.1. Sisältöperustainen suodatus

Yksinkertainen suodatin tutkii, esiintyvätkö tietyt sanat sähköpostissa, esimerkiksi ”seksi”, ”Viagra” tai ”ilmainen”. Suodattimen toimivuus riippuu siitä, miten kattava ja toimiva sanalista on suodattimeen etukäteen asetettu. Pelkästään tietyn sanan ilmeneminen postissa saattaa jo kertoa postin olevan roskapostia. Roskapostien lähettäjät ovat kuitenkin oppineet kiertämään tämän suodattimen välttämällä tiettyjä sanoja tai vaihtamalla niiden kirjoitusasua. Tästä esimerkkinä on avainsanojen tahallinen väärinkirjoittaminen, esimerkiksi sana ”Viagra” voidaan kirjoittaa ”V1agra”, “Via'gra”, “V I A G R A”, “Vaigra”, “\/iagra” tai “Vi@graa”. Kilpajuoksu on loputon roskapostin lähettäjien

kekseliäisyyden ja suodattimeen asetettujen sanojen välillä. Siksi tämän menetelmän käyttäminen pelkästään on tehotonta.

4.1.2. Tilastollisiin menetelmiin perustuvat suodattimet

Tilastolliset menetelmät perustuvat myös sähköpostin sisällön tarkasteluun, mutta yhden sanan sijaan tarkastellaan sanajoukkoja. Näitä kerättyjä sanajoukkoja käsitellään tilastollisin menetelmin.

Bayesin menetelmä perustuu sähköpostin sana-alkioiden esiintymistodennäköisyyteen roskapostissa ja oikeassa sähköpostissa. Sana-alkioiden todennäköisyyden laskemista varten tarvitaan sähköpostiaineisto, josta todennäköisyys voidaan laskea. Tätä suodattimen opetusta varten Graham [2002] keräsi 4000 roskapostiviestiä ja 4000 oikeaa sähköpostia. Osa aineistosta oli tullut suoraan Grahamille ja osa joillekin muille. Kaikkien sähköpostien sanat ja välimerkit käytiin läpi, mukaan lukien myös otsakkeet, HTML-koodi, Java-scriptit ja kaikki mistä yleensä voitiin muodostaa sana-alkioita. Merkittäväksi muodostui tekstissä olevien pornografisien termien ohella mm. HTML-koodissa tekstin tausta- ja pohjaväriä määrittävä koodi (”FF0000” HTML: kirkas punainen). Kaikelle näille aineiston osille, sana-alkioille, laskettiin todennäköisyysarvo sille, kummasta aineistosta ne todennäköisemmin löytyvät. Yhdessä sähköpostissa olevien kaikkien sana-alkioiden todennäköisyyksien keskiarvo lopulta ratkaisi tuloksen. Jos esimerkiksi sana ”sex” löytyi 70 %:sta kaikista aineiston roskaposteista määräytyi sana-alkion roskaposti todennäköisyydeksi 70 %. Tilastollisesti laskettuna sähköpostin, joka sisältää sanan ”sex”, todennäköisyys olla roskaposti muodostuu kaikista postissa olevista sana-alkioista, eikä tämän yksittäisen sanan merkitys ole ratkaiseva. Jos muut viestissä olevat sanat ovat alle 50% todennäköisyyden omaavia sanoja ja niitä on riittävästi, niin viestiä ei luokitella roskapostiksi [Graham, 2002].

Ongelmaksi muodostuivat sanat, joita ei aineistossa ollut ja joille ei löydy valmista luokitusta. Graham [2002] havaitsi yrityksen ja erehdyksen kautta, että sopiva arvo uusille sanoille on 40 %. Aikaisemmin esiintynyt sana on todennäköisesti harmiton.

A) Koko aineistossa olevia sanoja	
Madam	0.99
Promotion	0.99
Republic	0.99
shortest	0.047225013
Mandatory	0.047225013
standardization	0.07347802
Sorry	0.08221981
Supported	0.09019077
people's	0.09019077
Enter	0.9075001
Quality	0.8921298
Organization	0.12454646
Investment	0.8568143
Very	0.14758544
valuable	0.82347786

B) Grahamin sähköpostissa olevia sanoja	
Perl	0.01
Python	0.01
Tcl	0.01
Scripting	0.01
Morris	0.01
Graham	0.01491078
Guarantee	0.9762507
Cgi	0.9734398
Paul	0.027040077
Quite	0.030676773
pop3	0.042199217
Various	0.06080265
Prices	0.9359873
managed	0.06451222
Difficult	0.071706355

Taulukko 1. Viisitoista mielenkiintoisinta sanaa A) kerätyssä aineistossa ja B) Grahamin omassa sähköpostissa ja niiden roskapostitodennäköisyydet [Graham, 2002].

Taulukosta 1 huomataan, että luokittelun kannalta on merkitystä, kenelle sähköposti on lähetetty. Grahamin roskapostista ei todennäköisesti löydy sanaa “Graham”, eikä hänen henkilökohtaiseen elämään tai työhön liittyviä sanoja. Kummankaan taulukon sana-alkioiden todennäköisyydet eivät sinällään ole yleiskäyttöisiä.

Bayes-suodattimen teho ei perustukaan pelkästään sanoihin, vaikka niiden todennäköisyyksiä lasketaankin, vaan tällä tavoin havaitaan sanojen todennäköisyyksistä viestintätyyli. Viestintätyyleistä erottuvat ne, jotka ovat muokattu pääasiassa massaviestiksi, ja ne, jotka ovat tarkoitettu yksilölle.

Roskapostin kirjoittajan on vaikea ottaa tätä huomioon roskapostia kirjoittaessaan. Suodattimen selkeä etu on myös se, että se voidaan opettaa aina käyttäjäkohtaisesti. Esimerkiksi, jos perheenjäsenten tai kollegoiden nimet esiintyvät sähköpostissa, alentaa se selvästi roskapostin todennäköisyyttä.

Graham [2002] asettaa toivonsa tulevaisuudessakin Bayes-suodattimille, sillä suodattimet kehittyvät sitä mukaa kuin roskaposti. Esimerkiksi merkkijono ”c0ck” omaa huomattavan suuren roskapostin todennäköisyyden ”cock”-sanaan verrattuna ja siten varmemmin suodattuu pois roskapostina.

POPFile, SpamProbe, Bogofilter, DSPAM ja dbacl ovat esimerkkejä kaupallisista tuotteista, jotka perustuvat kaikki Bayesin menetelmään. Näitä on kehitetty ja tehostettu perusmenetelmästä mm. ”kohinanvaimennuksella” (noice reduction), jolloin tietystä osoitteesta tulevat ääritapaukset saadaan suodatussääntöihin mukaan, ja ottamalla mukaan painotetusti otsakekenttä. Parannetuissa menetelmissä valitaan myös tarkoin, mitkä sanat otetaan mukaan tarkasteluun. Valinta voi kohdistua esimerkiksi 15-20 useimmin esiintyvään sanaan, ei koko sähköpostin sanoihin.

Roskapostin lähettäjät ovat keksineet keinoja ohittaa Bayesin tilastollinen tarkastelu liittämällä roskapostinsa perään pitkä lista postiin sinänsä liittymättömiä sanoja satunnaiseen järjestykseen. Tällaista listaa kutsutaan ”sanasalaatiksi”, ja sen tarkoituksena on kumota varsinaisen viestin sanojen roskapostitodennäköisyys.

4.1.3. Heuristisiset suodattimet

Heuristiset eli kokemukseen perustuvat suodattimet ovat osoittautuneet käytännössä myös erittäin tehokkaiksi. Näissä menetelmissä ratkaisevaa on suodattimen puoliautomaattinen tai täysin automaattinen oppiminen. Se, miten itse suodatin on toteutettu, ei ole ratkaisevaa. Suodatin on voitu toteuttaa yhtä tai useampaa perusmenetelmää käyttäen. Heuristiset suodattimet kehittyvät oppimalla.

4.1.3.1 SpamAssassin

SpamAssassin on tunnettu sääntöpohjainen suodatin, joka on tarkoitettu ensisijaisesti yrityskäyttöön ja Unix-ympäristöön. SpamAssassinia voidaan käyttää muissakin järjestelmissä, joissa on eriytetty sähköpostipalvelin (mail server). Unixissa tämä voidaan toteuttaa Procmail-ohjelmalla, joka välittää sähköposteja CIS Unix -käyttäjille. Normaalisti Procmail välittää postin oletus-INBOX:iin. Kotihakemistoon voidaan kuitenkin määritellä .procmailrc-tiedosto, jonka avulla tulevaan sähköpostiin voidaan tehdä sääntöjä jälleenlähettämistä, kääntämistä toiseen osoitteeseen tai poistamista varten. Tässä sisään tuleva posti ohjataan suodatettavaksi SpamAssassin-ohjelmalla.

SpamAssassin soveltuu suuriin yhteisöihin, jossa on paljon sähköpostin käyttäjiä samassa sähköpostijärjestelmässä, kuten esimerkiksi suuryrityksissä, joissa järjestelmän ylläpitäjä huolehtii sääntöjen päivityksestä. SpamAssassin hyödyntää useita eri sääntöihin perustuvia testejä jokaiselle saapuvalle postille. Kun olemassa oleva suodatus ei riitä ja järjestelmän ylläpitäjää havaitsee jonkin tietyn roskapostin kuormittavan järjestelmää, hän voi lisätä juuri tähän roskapostiin tehoavan uuden säännön estämään sitä menemästä käyttäjille. Automaattista päivitystä tässä järjestelmässä ei ole. Säännöt tutkivat esim. otsakekentän ja varsinaisen tekstin sisältöä ja muotoilua. Eri osatekijöistä lasketaan arvo, joka kuvaa roskapostin todennäköisyyttä. Mitä korkeampi on

tulos, sitä todennäköisemmin kyseessä on roskaposti. Postit merkitään ennen kuin ne toimitetaan eteenpäin. Postit voidaan toimittaa normaali-INBOX:iin tai roskapostia varten olevaan erilliseen INBOX:iin riippuen siitä, ylittääkö saapunut posti roskapostin todennäköisyydelle asetetun raja-arvon. Käytännössä SpamAssassin lisää roskapostin todennäköisyyden sähköpostiotsakkeeseen (mail header) esimerkiksi:

- X-MailScanner-SpamCheck: spam, SpamAssassin (score=6.7, required 5))
- X-MailScanner-SpamCheck: not spam, SpamAssassin (score=-0.8, required 5).

SpamAssassinin teho perustuu siihen, että se voidaan muokata kunkin käyttäjän tarpeiden mukaiseksi. Nämä ohjeet voidaan tallentaa edelleen käyttäjän kotihakemistoon. Toisaalta on huomattava, että suuret määrät sääntöjä ja ohjeita jokaisella käyttäjällä vievät paljon levytilaa [Sand, 2002]. Sovellukset kuten AntibodyMX, McAfee, SpamKiller, Spamnix, SpamEliminator, MailLaunder, SmarterMail Enterprise ja Mail Them Pro käyttävät SpamAssassinin heuristista menetelmää sähköpostin suodatuksessa.

4.1.3.2 CRM114

Ensimmäinen CRM114-versio (the Controllable Regex Mutilator) oli hypoteettinen suodatin, jonka luonnos tehtiin 1998. Alkuperäinen tarkoitus ei ollut roskapostien suodatus vaan erilaisten aihealueiden lajittelu sähköposteista. CRM114 eroaa kaikista muista suodattimista ohjelmointikielensä ansiosta. CRM114 on ohjelmoitavissa olemaan mikä tahansa suodatin tai useampia suodattimia yhtä aikaa [Zdziarski, 2005].

Yksi CRM114:n suunnitteluolettamuksista on ollut se, että yksittäiset sanapiirteet eivät ole yhtä tärkeitä kuin sanojen ominaisuudet ja ominaisuuksien kasaantumat. Alkuperäinen CRM114:n koekoodi perustui kirjainmonikkoihin (letter tuple), eikä sanamonikoihin (word tuple). Testaus kevyesti naamioituja roskaposteja vastaan osoitti, että kirjainmonikoihin

perustuvan koodin erottelutarkkuudeksi saatiin yli 98 %. Tämä vakuutti, että monikkopiirteisiin perustuvat käsittelyt ovat merkittävästi parempia kuin yksittäisten merkkien piirteisiin perustuvat. Ennakkokäsityksistä poiketen todettiin, että oppiva luokittelija voi pystyä huomattavasti parempaan tarkkuuteen kuin ihmisen luoma heuristiikkajärjestelmä koskaan [Assis et al.,2005].

CRM114 ei varsinaisesti ole suodatin vaan suodatinoptimoitu ohjelmointikieli. Se on toteutettu kerroksittain: ohjelmakirjasto, ohjelmointikieli, valmiiksi ohjelmoidut suodattimet ja toimintaa ohjaavat parametritiedostot. Toimintaa voidaan helposti säätää parametritiedostoa muokkaamalla ilman ohjelmointitaitoja. CRM114:ssä valmiina oleva tilastollinen suodatus perustuu vanhoihin menetelmiin. Siinä missä Bayes-tekniikka pohjautuu Thomas Bayesin (1702 - 1761) oppeihin, pohjautuu tämä CRM114:ssa Andrei Markovin (1856 – 1922) teoriaan, Markovin ketjuun, jonka perusajatuksena on muistava satunnaisprosessi ja jonka edistyksellisiä tilastollisia menetelmiä voidaan osin hyödyntää roskapostinsuodatukseen. Markovin piilomalli (Hidden Markov Model, HMM) on tavallisesta Markovin mallista johdettu laajennus, jossa sanoja tärkeämpää on sanojen järjestys. Malli perustuu fraaseihin, ilmaisuihin ja sanajärjestyksiin. Markovin mallin on todettu olevan noin 40 % tehokkaampi kuin perinteinen, yksittäisiin sanoihin perustuva tilastollinen malli [Zdziarski, 2005].

CRM114-menetelmän tilastollisen mallin kiertäminen on astetta vaikeampaa kuin Bayesin menetelmässä, sillä roskapostin perään lisätty ”sanasalaatti”, joka sisältää vain satunnaisia ”hyviä” sanoja, ei riitä. Jotta tämän menetelmän kiertäminen olisi mahdollista, on roskapostin perään lisättävä oikean sanajärjestyksen ja ”järkevää” tekstiä sisältävä tarina.

4.1.4. Tarkistenumeroperustaiset suodattimet

Tarkistusnumeroperustaiset suodattimet (Checksum Based Filters, CBF) perustuvat roskapostin sisältöön, joka on aina lähes muuttumaton yhden lähettäjän tietyssä roskapostissa (massaroskaposti useille käyttäjille). Viestistä poistetaan vastaanottajan sähköpostiosoite sekä mahdollinen jäljite (web bug), joka on evästeen (cookie) kanssa toimiva tunniste sähköpostissa. Loppu viestin sisällöstä muunnetaan pitkäksi tarkistenumeroksi, joka toimii jatkossa sähköpostin tunnisteena. Järjestelmä vaatii siis ihmisen, roskapostin vastaanottajan, varmistavan postin roskapostiksi.

Tunnisteita verrataan vertaisverkossa (peer-to-peer network) ja raportoidaan, onko viesti nähty, onko se roskaposti ja kuinka moni on nähnyt sen. Tämän seurauksena syntyy suuri tietokanta, josta löytyy olemassa olevien roskapostien tunnisteet. Algoritmin on kehittänyt järjestö Rhyolite Software23, joka tunnetaan nimellä Distributed Checksum Clearinghouse. Menetelmä toimii sekä sähköpostin vastaanottajan että Internet-palvelun tarjoajan järjestelmissä. Menetelmällä voidaan pysäyttää roskaposti heti, kun joku on havainnut kyseisen postin. Toisaalta menetelmässä on myös heikkouksia, sillä se perustuu tunnisteen laskentatavan kykyyn tunnistaa roskapostien lähettäjien tekemät satunnaiset muutokset roskapostiin ja kykyyn erottaa ne normaaleista sähköposteista.

Tarkistusnumeroperustaisilla suodattimilla voidaan estää iso osa roskaposteista, mutta tunnistusprosentti jää selvästi alhaisemmaksi kuin muissa menetelmissä. Tämän menetelmän suurin etu onkin siinä, että suurta osaa sähköposteista ei edes välitetä eteenpäin, kun internetpalvelun tarjoaja on saanut siitä tiedon. Näin säästetään välitettävän datan määrää ja vältetään roskapostien aiheuttamia verkkotukoksia [Vohra, 2005].

Tarkistusnumeroperustaisen menetelmän kiertäminen on asetta vaikeampaa kuin CRM114:n tai Bayesin menetelmän. Roskapostin perään liitetyn ”järkevän” tekstin pitää olla satunnaista ja vielä jokaisessa yksittäisessä roskapostissa erilainen. Tällaista satunnaista tekstiä saadaan helposti kopioitua esimerkiksi erilaisista uutisryhmistä. Suodatin ei välttämättä ole sidottu sähköpostin sisältöön vaan voidaan tunnistaa myös erilaisia rakenteita roskapostissa, kuten ”sanasalaatin” sijainti ja määrä roskapostissa.

Kaikki nämä aikaisemmin mainitut roskapostin sisällön tunnistamiseen perustuvat menetelmät ovat kykenemättömiä torjumaan roskaposteja, joissa varsinainen viesti on kuvamuodossa ja suodattimien harhauttamiseksi on lisätty satunnaista järkevää tekstiä, kuten kuvassa 8. Tällainen roskaposti voidaan kuitenkin tunnistaa muita menetelmiä käyttäen.

Kuva 8. Roskaposti, jossa kaikki sisältöperusteisten suodattimien tunnistamat sanat ovat kätketty kuvaan ja muu viestissä oleva teksti ja otsikko ovat valittu satunnaisesti Raamatusta.

4.2. Roskapostin lähettäjän tunnistamiseen perustuvat estomenetelmät

Roskapostin estomenetelmät voivat perustua myös johonkin muuhun kuin sisällön tunnistamiseen. Lähettäjän tunnistamiseen perustuvat menetelmät ovat toimivia vain rajatuissa tapauksissa. Näitä menetelmiä voidaan kuitenkin tehokkaasti käyttää yhdessä sisältöperusteisten suodatusmenetelmien kanssa, esimerkiksi heuristisen menetelmien tukena ja näin tehostaa suodatusta. Lähettäjän tunnistamiseen perustuvia menetelmiä on olemassa useita ja uusia kehitetään jatkuvasti lisää. Tässä luvussa esitellään muutamia tunnetuimpia menetelmiä.

Osa sähköpostipalveluja tarjoavista yrityksistä suhtautuu välinpitämättömästi roskapostitusta ammatikseen harjoittaviin asiakkaisiinsa. Tällainen palveluntarjoaja voidaan asettaa vastuuseen roskapostin levittämisestä. Näistä palveluntarjoajista ja roskapostittajista kerätään listoja ja tietokantoja, joita epäviralliset yhteisöt voivat levittää keskenään esimerkiksi internetin keskustelupalstoilla.

Musta- ja valkealistaus (blacklisting and whitelisting) ovat menetelmiä, joissa kerätään tunnettuja internetosoitteita ja verkkotunnuksia, joista roskapostia lähetetään tai ei lähetetä. On olemassa useita julkaistuja mustia listoja, joita palveluntarjoajat tai käyttäjät voivat ladata sähköpostin lajittelumekanismiinsa. Reaaliaikaisen mustan listan (real time blackhole list, RBL) kehittäminen alkoi 1997, kun Paul Vixie alkoi kerätä vapaasti käytettävää listaa internetpalvelujen tarjoajista, joiden verkoista lähetetään roskaposteja [Zdziarski, 2005]. Mustalla listalla olevat IP-osoitteet tai verkkopalvelutunnukset voidaan sulkea suoraan pois roskapostina ilman, että niitä tarkistetaan millään muulla roskapostin suojausmenetelmällä. Vastaavasti valkoisella listalla olevia IP-osoitteista tai verkkopalvelutunnuksista tulevat sähköpostit voidaan päästää suoraan läpi niitä erikseen tarkistamatta. Musta- ja valkealistoja voidaan kerätä useilla eri tavoilla. Yleisesti verkkoyhteisöt (internet community) tiedottavat toisilleen uusista roskapostia lähettävistä osoitteista. Mustalistaus on periaatteeltaan yksinkertainen menetelmä ehkäistä roskapostia, mutta ei kuitenkaan täysin aukoton, joten tähän menetelmään usein yhdistetään jokin suodatusmenetelmä, esim. CRM114 tai useiden suodattimien yhdistelmä alentamaan mahdollisien virheiden määrää.

Mustalistauksen kiertämiseen ovat kekseliään roskapostin lähettäjät kehittäneet useita keinoja [Spammer-X, 2004]. Kaikki nämä menetelmät perustuvat siihen, että alkuperäisen lähettäjän IP-osoite ei tule missään vaiheessa ilmi ja että sähköposti lähetetään jonkun toisen, luotettavan yhteisön tai palveluntarjoajan

toimesta. Kuitenkin vastaanottavan tahon roskapostin torjuntaohjelmisto saattaa aktivoitua siitä huolimatta, vaikka sähköpostin lähettäjä ei ole mustalla listalla saadessaan satoja yhteydenottopyyntöjä samasta IP-osoitteesta.

Välipalvelimien (proxy server) tai SMTP-palvelimen kaappaus omaan käyttöön on ollut eräs tapa välttää roskapostin lähettäminen mustalistatusta IP-osoitteesta. Tämä on kuitenkin vähentynyt välipalvelimien turvallisuuden paranemisen myötä.

Internet sivuilla käytettävän CGI:n (Common Gateway Interface) käyttö roskapostin lähettämiseen luotettavasta IP-osoitteesta on edelleen suosittua. Esimerkiksi ”ota yhteyttä” –sivulla voisi olla seuraava HTML-koodi:

```
…
<form method=”POST” action=”http://domain.com/fl/cgi-bin/vast-mail.cgi”>
<input type=”hidden” name=”recipient” value=”webmaster@domain.com”>
<input type=”hidden” name=”required” value=”realname, subject, email”>
…
```

Roskapostin lähettäminen tältä sivulta on erittäin helppoa, sillä webmasterin sähköpostiosoite on muuttujana ja se voidaan täten korvata satunnaisella sähköpostiosoitteella. Tarvitaan ainoastaan ohjelma, joka lähettää POST-viestin sopivilla muuttujilla CGI:lle. Näin tunnistamaton roskaposti on valmis lähtemään tämän sivuston omistajalta.

Langattoman lähiverkon (WLAN) käyttäminen roskapostin levitykseen on uusi ja suhteellisen helppo tapa. Tämä johtuu siitä, että langattomat verkot ovat vielä erittäin huonosti suojattuja. Roskapostin lähettäjän on helppo käyttää kannettavan tietokoneen havaitsemia suojaamattomia langattomia lähiverkkoja omiin tarkoituksiinsa. Kaikki edellä mainitut keinot ovat laittomia. Kuitenkin roskapostin lähettäjä on enemmän kiinnostunut voitoista ja siitä, että oma IP-osoite ei paljastu mustalistojen kerääjille.

4.3. Suodattimien yhdistelmät

Jos yhdellä suodattimella voidaan estää 95 % roskapostista, niin kahden toisistaan riippumattoman suodattimen yhdistelmä suodattaa jo 98 % roskapostista ja kolmen suodattimen yhdistelmä peräti 99 %. Tämä menetelmä on saanut roskapostia vastaan taistelevat tahot suunnittelemaan monimutkaisia suodatinketjuja, joiden eri osat keskittyvät tietyntyyppisiin roskaposteihin. Tällainen sisältöpohjaisten suodattimien yhdistelmä liitettynä mustalistatarkasteluun antaa erittäin hyvän lopputuloksen. Nykyään käytössä olevat suodattimien yhdistelmät koostuvat seuraavista menetelmistä:

- **Otsakkeen tarkastelu**, jossa tarkastellaan, ovatko sähköpostin lähettäjätiedot oikeita ja muuttamattomia.
- **Sisältöanalyysi**, jossa haetaan roskapostille ominaisia sanoja tai sanojen yhdistelmiä viestin sisällöstä.
- **Mustalistatarkastelu**, jossa lähettäjän IP-osoitetta verrataan staattiseen tai dynaamiseen mustaan listaan. Mustalistan ylläpitäjiä esim. MAPS, ORBS ja SpamHouse.
- **Bayes-suodatin tai Markovin malli**, jotka käsittelevät sisältöä tilastollisin menetelmin päätelläkseen roskaposti todennäköisyyden.
- **Tarkistenumeroperustainen tarkastelu**, jossa tarkistenumeroita ylläpitävien yritysten ylläpitämistä ja toimittamista tiedoista tutkitaan, onko sähköposti tunnettu roskaposti. Tarkistenumeroita ylläpitäviä yrityksiä on esim. Razor, Pyzor ja DCC.

Tällaisella yhdistelmällä pystytään suodattamaan 95 – 98 % tulevasta roskapostista [Spammer-X, 2004].

5. Suurten yritysten käyttämät menetelmät, tapaus Nokia

Haastattelin tätä tutkimusta varten Nokian sähköpostiturvallisuudesta vastaavaa Lasse Jokista [Jokinen, 2007]. Tavoitteena oli selvittää, miten yrityksissä on ennen varauduttu ja varaudutaan nyt roskapostin ehkäisemiseen.

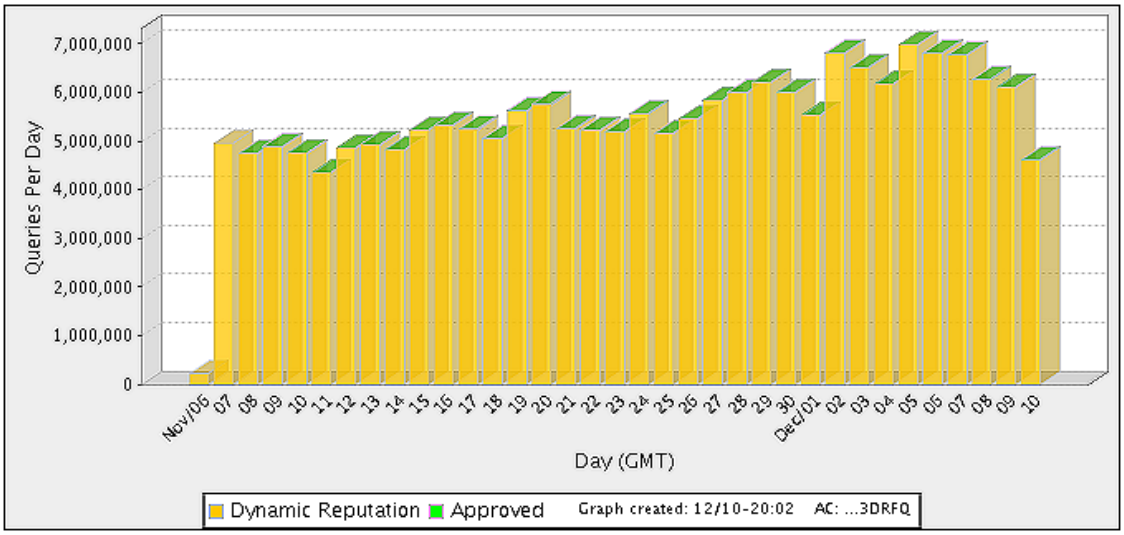
Nokialaisen sähköpostikansiossa suurin osa posteista tulee Nokian sisältä ja vain pieni osa on internetsähköpostia. Aivan sähköpostin alkumetreiltä on kaikkien Nokian ulkopuolelta tulevien sähköpostiosoitteiden, internetsähköpostin, eteen liitetty sana ”ext” kertomaan vastaanottajalle, että sähköposti saattaa tulla tuntemattomasta lähteestä. Kaikkeen ”ext”-sähköpostiin on ohjeistettu suhtautumaan varauksellisesti.

Ennen vuotta 1997 ei Nokialla juuri kiinnitetty huomiota roskapostiin tai virustorjuntaan, mutta juuri kyseisenä vuonna roskapostin määrä moninkertaistui ja roskapostissa olevien virusten määrä kasvoi huolestuttavasti. Vuonna 1998 pääasiallinen estomenetelmä oli mustalistaus.

Vuonna 2000 Nokialla siirryttiin käyttämään yrityksessä kehitettyä *Nokia* Message Protector tuotetta, joka oli sisällytetty Nokian palomuuriratkaisuihin. Alussa suurin ongelma oli virukset, joiden torjumiseen ensisijaisesti keskityttiin. Vasta muutaman vuoden kuluttua pääpaino siirtyi roskaposteihin. Nokialla alettiin käyttää kaupallista sääntöpohjaista sisältöön perustuvaa suodatinta. Suodattimen säännöstö oli sisällytetty palvelua tarjoavan yrityksen tekemään määrittelytiedostoon, joka perustui silloiseen sähköpostivirtaan ja siinä esiintyviin roskaposteihin. Määrittelytiedosto oli dynaaminen ja säännöstöä ylläpitävä yritys lähetti päivityksen kerran kahdessa kuukaudessa asiakkailleen. Tässä suodattimessa oli viisi eri osa-aluetta, joiden roskaposti todennäköisyyttä voitiin erikseen määritellä. Jokaisella osa-alueella oli oma

raja-arvonsa, jota voitiin tilanteen mukaan säätää. Näistä osa-alueista voitiin sitten päätellä yhden postin roskapostin todennäköisyys joko siitä, että yhden osa-alueen raja-arvo ylittyi tai kaikkien osa-alueiden yhteenlaskettu raja arvo ylittyi. Tämä vaati jatkuvaa ylläpitoa ja säätämistä, sillä tilanne muuttui nopeasti. Aina uuden määrittelytiedoston saapuessa roskapostin määrä väheni merkittävästi ja jälleen kasvoi pikku hiljaa, ennen kuin uusi määrittelytiedosto saatiin ajettua järjestelmään.

Nokia käyttää pääasiassa mustalistausta ja valkealistausta roskapostin suodatuksessa. Jo 30 – 35 % roskapostista voidaan suodattaa IP-osoitteeseen perustuen. IP-osoite tai verkkotunnus (domain name) riittää useissa tapauksissa luokittelemaan sähköpostin roskapostiksi. Mustalistatuista lähteistä saapuva sähköposti estetään kokonaan saapumasta vastaanottajalle ja valkealistattu posti toimitetaan aina perille, tosin virustarkastelu tehdään ja roskapostitodennäköisyys lasketaan sekä liitetään aina postin otsakkeeseen. Menetelmä on erittäin tehokas. Nokia ei itse ylläpidä mustalistaa, vaan siitä huolehtii ulkopuolinen yritys. Sama yritys hoitaa useiden suurien yhtiöiden sähköpostin suodatukseen käytettyjen mustalistojen ylläpitoa. Nykyinen mustalista on dynaaminen, joten sitä voidaan tarvittaessa muuttaa nopeassakin syklissä. Aiemmin käytettyjen staattisten listojen ongelmana oli juuri niiden hidas päivitettävyys. Roskapostien ollessa kyseessä muutama päivä on jo liian pitkä aikaväli. Roskapostin elinkaari on tyypillisesti vain päivä tai pari. Roskapostiaktiviteettia saattaa tulla tietystä osoitteesta vain lyhyen aikaa, joten dynaaminen lista on tehokkaampi ja myös paljon lyhyempi kuin staattinen lista. Ongelmaksi staattisessa listassa muodostuu se, että oikeat sähköpostit suodattuvat pois vanhentuneen mustanlistan takia. Esimerkiksi korealainen operaattori on ilmoittanut kaikkien aliverkkojensa osoitteet staattiselle mustalle listalle, jolloin oikeat sähköpostit suodattuvat myös pois. Kuvassa 9 nähdään mustalistattujen ja hyväksyttyjen sähköpostien välinen suhde selkeästi.



Kuva 9. Päivittäiset yhteyspyynnöt sähköpostin lähettämiseksi noin kuukauden ajalta. Vihreällä on merkitty valkolistatut, hyväksytyt yhteyspyynnöt ja keltaisella mustalistatut, hylätyt yhteyspyynnöt. (Kuva: Lasse Jokinen, Nokia Oyj)

Eri IP-ryhmistä tai palvelimesta tulevia viestejä tarkastellaan ja lasketaan, kuinka monta yhteydenottoa ne pystyvät tekemään minuutin aikana sekä kuinka monta väärää vastaanottaja osoitetta niille annetaan. Tämän perusteella sähköposti laitetaan ”jäähylle” satunnaiseksi ajaksi. Jäähyllä tarkoitetaan sitä, että roskapostin lähettäjä saa muodostettua esim. 20 yhteyttä minuutissa, jonka jälkeen seuraavien yhteyksien muodostumista viivästytetään 5-10 min jokaista. Tämän jälkeen tarkistetaan, onko vastaanottaja olemassa. Jos vastaanottaja on olemassa, sähköposti otetaan vastaan ja viestiä arvioidaan Cloudmarkilla, joka on sääntöpohjainen suodatin. Cloudmark hyödyntää mm. alakohdassa 4.1.3. kuvattua SpamAssasinin ominaisuuksia. Tämä tarkastelu on erittäin nopea.

Näiden tarkastelujen jälkeen sähköpostipalvelin joko lähettää roskapostin lähettäjälle SMTP-viestin ”ei otettu vastaan” (550 - not accept for delivery – permanent failure) tai laittaa postin jäähylle SMTP:n antaessa vastineen ”lähetä myöhemmin uudelleen” (421- please try again later). Roskaposteja ei siis edes oteta vastaan ja suuret määrät roskapostia voidaan torjua ennen kuin käyttäjä edes näkee niitä. Tässä kohtaa noin 96 % roskaposteista on eliminoitu.

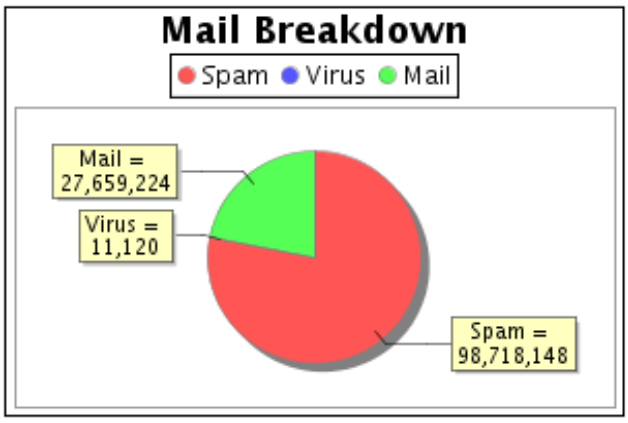
Edellisten tarkistuksen jälkeen siirrytään tarkistenumeroperustaiseen suodatukseen. Hyväksytyistä sähköposteista lasketaan sähköpostin tarkistenumero etukäteen määritellyllä menetelmällä. Tarkistenumero lähetetään ulkopuoliselle firmalle, joka hallinnoi roskapostien tarkistenumeroita. Tämä ulkopuolinen firma kerää roskapostien tarkistenumeroita kuten alakohdassa 4.1.4 on mainittu ja palauttaa sähköpostin roskapostin totuusarvon (1/0). Tarkistenumeroiden roskapostin totuusarvot pohjautuvat sen hetkiseen dynaamisen sähköpostivirtaan ja niistä saatuihin kokemuksiin.

Seuraavaksi sähköpostia tarkastellaan viiden eri kaupallisen sähköpostin sisältöön perustuvan suodatusmenetelmän avulla. Jos sähköposti hyväksytään näissä kaikissa suodattimissa puhtaaksi, toimitetaan se vastaanottajalle. Näiden kaupallisten tuotteiden roskapostin suodatus- tai tunnistusmenetelmistä ei valitettavasti saa mitään tietoa. Jos menetelmät paljastettaisiin, niin roskapostin lähettäjät voisivat varautua näihin tarkistuksiin ja muuttaa roskapostejansa siten, että suodattimet eivät enää toimisi.

Sähköpostin otsakkeeseen lisätään tietoa vastaanottajaa varten esimerkiksi seuraavasti:

X-Nokia-AV: Clean	- Antivirus tarkistus OK
X-pstn-spam: W	- W-valkealistattu, Y-Roskaposti, N-Ei roskaposti
X-Spam-Score: 99.00%	- Todennäköisyys sille ,onko kyseessä roskaposti.

Virustarkistus käydään läpi kaikille eteenpäin lähetettäville sähköposteille. Tällä hetkellä virusten määrä roskapostisuodatuksen jälkeen kaikissa sähköposteissa on alle 0,0001 %, eli virukset suodattuvat erittäin tehokkaasti samalla kuin roskapostit suodattuvat (kuva 10).



Kuva 10. Yhden kuukauden sähköpostijakauma, jossa virukselliset sähköpostit ovat selkeä vähemmistö (Kuva: Lasse Jokinen, Nokia Oyj).

Menetelmä ei kuitenkaan ole aukoton, joten myös oikeat sähköpostit saattavat suodattua. Mustalle listalle on saattanut myös joutua yrityksiä, jotka ovat itse tai heidän palveluntarjoajansa on ollut ilkivallan kohteena.

Tulevaisuudessa Nokian roskapostin suodattamisessa tullaan hyödyntämään uuden SendMail 8.13:n mukana tuomia ominaisuuksia ja tarkistuksia. Jossain vaiheessa tullaan ottamaan selektiivisesti käyttöön aidonnetut sähköpostit, jotka on kuvattu kohdassa 7.1. Tarkoitus on estää mahdollisimman monta yhteydenottoa yhdyskäytävän (gateway) kautta. Kaikki sisään pääsevät yhteydenotot vaativat prosessointitehoa. Näiden käyttö yksistään ei todennäköisesti ole nopean aikavälin ratkaisu. SendMailin uusien ominaisuuksien ja aidonnettujen sähköpostien käytön tulisi yleistyä sähköpostinkäyttäjien keskuudessa, jotta yksinään nämä menetelmät voisivat estää roskaposteja tehokkaasti

Aikaisemmin DHA-hyökkäykset olivat yksi isoimmista ongelmista. Ennen sähköpostia voitiin lähettää ja vastaanottaa yrityksen miltä tahansa palvelimelta. Tällöin DNS:stä voitiin saada palvelimien nimiä, joihin taas lähetettiin miljoonia viestejä. Näistä viesteistä onnistuttiin löytämään suuri määrä toimivia sähköpostiosoitteita. Joissakin tapauksessa kuormitus kasvoi

niin suureksi, että järjestelmä ylikuormittui ja jopa palomuuri saattoi lopettaa toimintansa, kun pyyntöjä oli yli 500 palvelimelta ja yhteyksiä 75 000. Yhteyksiä kasvatettiin 100 000:een, mutta sekään ei riittänyt. Markkinoilta ei juuri löydy palomuuria, joka pystyisi näin moneen yhteyteen. Tällöin sähköpostin tutkinta tehtiin vasta, kun viesti oli otettu vastaan. Tämä puolestaan varasi suunnattomasti yhteyksiä. Roskapostin lähettäjät olivat myös tyytyväisiä saadessaan nopeasti viestinsä perille.

6. Roskapostia koskeva lainsäädäntö eri maissa

Roskaposti tavoittaa vastaanottajat globaalisti riippumatta siitä, mitkä ovat paikalliset lait ja määräykset. Tämä on merkittävä asia niin roskapostin lähettäjien kuin niiden vastaanottajien kannalta. Valitettavasti vain muutamat maat ovat ottaneet huomioon roskapostin lähettämisen ja tehneet tästä toiminnasta laitonta. Sellaiset maat kuten Venäjä, Intia ja Brasilia ovat kyllä havainneet roskapostiongelman ja näissä maissa on vaadittu roskapostin estäviä lakeja, mutta tällä hetkellä roskapostin lähettäminen ei ole laitonta kyseisissä maissa. Suomessa on säädetty suoramarkkinointia koskeva laki, joka määrittelee roskapostit laittomiksi: *Automatisoitujen soittojärjestelmien sekä telekopiolaitteiden, sähköpostiviestien, tekstiviestien, puheviestien, ääniviestien tai kuvaviestien avulla toteutettua suoramarkkinointia saa kohdistaa vain sellaisiin luonnollisiin henkilöihin, jotka ovat antaneet siihen ennalta suostumuksensa*. [Sähköisen viestinnän tietosuojalaki, 2004] Palveluntarjoajien ylläpitämiä sisältöön perustuvia suodatusmenetelmiä koskee myös perustuslain 10 §:n säätämä luottamuksellisen viestin loukkaamattomuuspykälä, joka asettaa palveluntarjoajan vastuuseen virheellisistä suodatuksista.

Lähes kaikissa maissa sähköpostiosoitteiden varastaminen katsotaan laittomaksi ja esimerkiksi suojaamattomien välipalvelimien käyttö roskapostin lähettämiseksi katsotaan rangaistavaksi luvattomaksi tietokoneen käytöksi. Japanissa suuri osa ihmisistä käyttää sähköpostiansa matkapuhelimesta tai muusta kannettavasta laitteesta, jolloin roskapostit tulevat kalliiksi käyttäjille esimerkiksi matkapuhelinlaskun datasiirtojen muodossa. Tämä on saanut Japanin tiukentamaan lainsäädäntöään roskapostin suhteen.

Kuva 11. Roskapostin lähetysalueet (punaisella suurin lähetysten määrä) perustuen lähettäjän IP-osoitteisiin 24h sisällä 6.11.2007 [Postini, 2007].

Kuvasta 11 nähdään, että suurin osa roskapostista tulee juuri maista, joissa roskaposteja koskevaa lainsäädäntöä ei ole, tai sen täytäntöönpano on puutteellista. Roskapostin lähettäjiä ei pystytä estämään toimimasta.

6.1. USA:n CAN-SPAM-laki

Valvonta ja kiinnijääminen eivät kuitenkaan ole vielä tehonneet toivotulla tavalla. Esimerkiksi Yhdysvalloissa vuoden 2004 alussa voimaan tullut CANSPAM Act -laki (Controlling the Assault of Non-Solicited Pornography and Marketing Act) tekee roskapostin lähettämisestä liittovaltion rikoksen. CANSPAM Act määrittelee roskapostituksen ja kieltää tilaamattoman sähköpostimarkkinoinnin [CANSPAM Act, 2003].

CANSPAM:n mukaan lakia rikkoo jokainen joka edesauttaa tietoisesti kotimaista (USA) tai ulkomaista mainostamista ja

1) tunkeutuu suojattuun tietojärjestelmään ilman asianmukaisia oikeuksia ja tahallisesti lähettää useita kaupallisia sähköposteja suoraan tai tämän tietojärjestelmän kautta.

2) käyttää suojattua tietokonetta lähettämään tai välittämään sähköposteja tarkoituksena huijata tai harhaanjohtaa vastaanottajaa tai verkkopalvelun tarjoajaa mainitun sähköpostin alkuperästä.
3) muuttaa olennaista sähköpostin otsaketietoa useissa kaupallisista sähköposteista ja tahallisesti panee alulle tällaisen sähköpostien lähetyksen.
4) rekisteröityy, käyttäen vääriä tietoja, jota oleellisesti väärentää todellisuudessa rekisteröityvän henkilön identiteetin vähintään viidessä sähköpostitunnuksessa tai verkkopalvelun käyttäjätunnuksissa tai vähintään kahdessa verkkotunnuksessa ja tahallisesti edesauttaa useiden kaupallisten sähköpostin lähettämisen näitä virheellisesti rekisteröityjä palveluita hyväksi käyttäen tai
5) virheellisesti väittää olevansa verkkotunnuksen haltija tai hänen laillinen edustajansa ja tahallisesti myötävaikuttaa vähintään viidestä verkko-osoitteesta lähetettyjen useiden kaupallisten sähköpostien lähettämisen sähköpostimarkkinoinnin [CANSPAM Act, 2003].

6.2. EU-lait

Euroopan Unionilla ei ole olemassa roskapostin kieltävää lakia, mutta on olemassa direktiivejä, jotka vastaavat CANSPAM:n sisältöä. Useassa EU-maassa roskapostin lähettäminen on kuitenkin kielletty. Iso-Britanniassa roskapostin lähettäminen yrityksille on sallittu ja roskapostin lähettäminen on kielletty ainoastaan henkilökohtaisiin sähköposteihin. Italiassa roskapostin lähettäminen on kielletty ja siitä voidaan tuomita 90 000 euron sakot sekä 3 vuotta vankeutta [Spammer-X, 2004].

6.3. Laillinen roskaposti

Kaikki sähköpostissa tulevat mainokset eivät suinkaan ole roskapostia. Kuvassa 12 on esimerkki mainoksesta, jossa on selkeästi kerrottu, mistä sähköpostiosoite on saatu, sekä ohjeet välttyä uusilta mainoksilta niin

halutessaan. Samoin otsaketiedoista nähdään, että ”from” ja ”return-path” ovat samat. Myös Nokian lisäämistä kentistä voidaan todeta, että lähettäjä ei ole mustalla listalla (X-pstn-spam: N), eikä muussakaan roskapostin tarkastelussa ole havaittu mitään roskapostin ominaisuuksia (X-Spam-Score: 0.00 %).

Microsoft Mail Internet Headers Version 2.0
Received: from esebh103.NOE.Nokia.com … by trebe101.NOE.Nokia.com …
Received: from esebh107.NOE.Nokia.com … by esebh103.NOE.Nokia.com…
Received: from mgw-mx05.nokia.com … by esebh107.NOE.Nokia.com …
Received: from sanomamx2.dmz.2ndhead.net … by mgw-mx05.nokia.com …
Received: from ATKTUKI04 … by sanomamx2.dmz.2ndhead.net …
From: "ext Helsingin Sanomat" <helsinginsanomat@hs.fi>
To: <HARRI.SUNDSTROM@NOKIA.COM>
Subject: Hanki talveen valopilkku
Date: Tue, 20 Nov 2007 18:18:22 +0200
Message-ID: <14ecf01c82b90$f8741020$10257f9e@corp.sanoma.fi>
X-Nokia-AV: Clean
X-pstn-spam: N
X-Spam-Score: 0.00%
Return-Path: helsinginsanomat@hs.fi
Content-Transfer-Encoding: 8bit
Content-Type: text/html; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

Kuva 12. Esimerkki mainoksesta, joka ei ole roskapostia.

Roskapostin lähettäjät osaavat myös kiertää nämä CANSPAM:n asettamat rajoitukset. Kiinnittämällä huomion muutamaan seikkaan roskapostissa voi roskapostin lähettäjä välttyä CANSPAM:n määrittelemiltä sakoilta tai vankilatuomiolta. Kun seuraavat seikat otetaan huomioon, on roskaposti laillinen:

1) Muuttamaton sähköpostin otsake. Varmistetaan, että otsakkeessa ei ole virheellistä tietoa.
2) Sähköpostissa on toimiva linkki, jonka avulla vastaanottaja voi halutessaan poistaa nimensä postituslistalta. Vaihtoehtoisesti toimiva sähköpostiosoite, jonne voi ilmoittaa halunsa poistua postituslistalta. Linkin tai sähköpostiosoitteen tulee toimia vähintään 10 päivää lähetyksen jälkeen ja nimi pitää poistaa postituslistalta 10 päivän sisään.
3) Selkeästi mainitaan, että kyseessä oleva sähköposti on mainos ja mainittava, mistä tuotteesta on kysymys.
4) Jos postissa on arkaluontoista materiaalia, esim. pornokuvia, pitää se selkeästi mainita sähköpostin otsikossa (subject).
5) Sähköpostia saa lähettää vain niille, jotka ovat erikseen antaneet suullisen tai kirjallisen luvan lähettää kaupallista sähköpostia.
6) Sähköposti ei saa tulla toisten tietokoneiden kautta, ellei lähettäjällä ole lupaa tähän.
7) Sähköpostiosoitteiden myyminen ja ostaminen on kiellettyä [Spammer-X, 2004].

7. Tulevaisuuden näkymiä

Nykyisissä sähköpostijärjestelmissä ei ole alun perin otettu huomioon roskapostin mahdollisuutta. Nykyisen sähköpostijärjestelmän puutteita pyritään korjaamaan erilaisilla roskapostin estomenetelmillä. Samaan aikaan roskapostien lähettäjät kehittävät omia menetelmiään jatkuvasti ohittaakseen olemassa olevia suodattimia ja käyttävät hyväksi nykyisen järjestelmän puutteita. Tulevaisuudessa ainoa mahdollisuus on kehittää kokonaan uudentyyppisiä sähköpostijärjestelmiä, joissa roskapostin lähettäminen alun perinkin on mahdotonta.

Internetin IP-osoitteita ja tunnuksia hallinnoiva ei kaupallinen ICANN (Internet Corporation for Assigned Names and Numbers) on pyrkinyt kokonaisvaltaiseen internet-verkon hallintaan. Valitettavasti ICANN ei pysty vaikuttamaan roskapostien leviämiseen säädöksillään. Esimerkiksi Kiinassa käytetään nimipalvelimia (domain name) ja järjestelmiä, jotka toimivat kiinalaisella merkistöllä. ICANN standardien mukaan näin ei saa olla, mutta Kiina voi tietenkin päättää omista standardeistaan omassa maassaan [Robertson, 2006].

Käytäntöjä pitää kehittää niin, että roskapostien lähettäminen on estetty. Tällä hetkellä yleisin internetin sähköposti protokolla, SMTP, ei tarjoa sähköpostin aidonnusta (authentication). Roskapostin lähettäjän on nyt helppo käyttää jonkun muun sähköpostiosoitetta lähettäessään roskapostejaan. Tämä on erittäin yleistä roskapostin lähettäjien keskuudessa. Jos esimerkiksi tietyn henkilön sähköpostia on käytetty roskapostin lähettämiseen, hän saattaa saada virheilmoituksen, siitä että sähköpostia ei ole voitu toimittaa vastaanottajalle. Tämä siis tapahtuu vaikka kyseinen henkilö ei ole koskaan lähettänyt ko. sähköpostia.

SPF (Sender Policy Framework) on avoin standardi, joka tarjoaa teknisen menetelmän estämään lähettäjän osoitteen väärentämisen. SPF suojaa otsakkeessa olevan lähettäjän osoitteen. Verkkotunnuksen omistaja (domain owner) julkaisee SPF käytännön verkkotunnuspalvelimellaan (DNS).

Esimerkiksi SPF-tietue voisi olla:

example.net. TXT "v=spf1 mx a:pluto.example.net include:gmail.com -all"

jossa tietueen osat merkitsevät seuraavaa:

v=spf1	SPF versio1
mx	vastaanottava sähköpostipalvelin saa lähettää postia
a:pluto.example.net	pluto.example.net on myös authorisoitu lähettämään
include:gmail.com	kaikki gmail.com:in laillistamat postit ovat OK
-all	muut laitteet eivät ole sallittuja lähettämään postia.

Sähköpostia vastaanottava palvelin voi nyt tarkistaa sähköpostissa mainitulta palvelimelta, onko ko. sähköposti SPF sääntöjen mukainen. Jos posti ei tule säännöissä mainitulta palvelimelta, on lähettäjä tekaistu ja sähköposti todennäköisesti roskapostia. Valitettavasti tämä menetelmä vaatisi toimiakseen sen, että kaikkien sähköpostipalvelimien tulisi tukea SPF-tarkistusta [SPF, 2007].

Verkkoavaimen (Domain Key) käyttö on tehokas sähköpostin autentikointimenetelmä, joka tarjoaa lähes päästä päähän kuittauksen ja varmennuksen sähköpostille. Eräs tällainen järjestelmä on Yahoon ja Ciscon kehittämä DKIM (Domain Keys Identified Mail). Tässä järjestelmässä nimipalvelujärjestelmä (Domain Name System, DNS) varmistaa sähköpostin lähettäjän oikeellisuuden salainen-julkinen -avainpareilla (public private key pair) ja vastaanottava sähköpostijärjestelmä vastaavalla menetelmällä tunnistaa sekä lähettäjän että sähköpostin muuttumattomuuden välityksen aikana. Jos sähköposti ei ole asianmukaisesti autentikoitu, ei sähköpostia toimiteta perille. Jos sähköposti kuitenkin havaitaan autentikoinnin jälkeen roskapostiksi (esim. Bayes-suodatin + käyttäjä), niin sekä lähettäjän että palveluntarjoajan tiedot on tallennettu ja nämä voidaan asettaa edesvastuuseen roskapostin lähettämisestä.

Molemmat voidaan laittaa luotettavalle mustalle listalle. Tutkimukset DKIM-järjestelmän käyttöönottamiseksi ovat meneillään.

Kuvassa 12 on verkkoavaimen käytön toimintaperiaate, joka etenee seuraavasti:

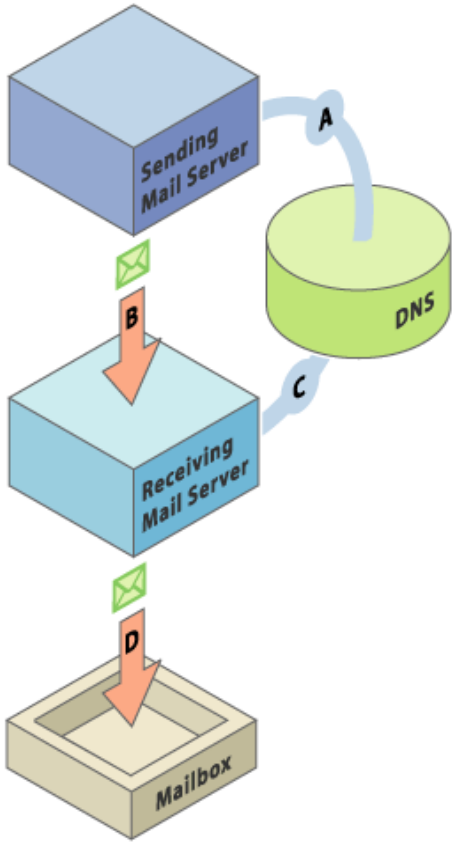
A) Verkkopalvelimen omistaja, internetpalveluntarjoaja tai yritys, generoi salainen-julkinen -avainparin (tai useita) lähetettävää sähköpostia varten. Julkinen avain julkaistaan DNS:ssä ja salainen avain asetetaan saatavaksi verkkoavainta käyttävälle sähköpostipalvelimelle.

B) Kun autentikoitu käyttäjä lähettää sähköpostia, verkkoavainta käyttävä sähköpostipalvelin muodostaa digitaalisen tunnisteen salaisesta avaimesta ja varsinaisesta sähköpostista. Digitaalinen tunniste liitetään sähköpostin otsakkeeseen ja sähköposti lähetetään vastaanottajan sähköpostipalvelimelle.

C) Verkkoavainta käyttävä vastaanottava sähköpostipalvelin purkaa tunnisteen ja pyytää sähköpostin otsakkeessa lähettäjäksi mainitulta verkkopalvelimelta julkisen avaimen.

D) Vastaanottava sähköpostijärjestelmä tarkistaa tunnisteen sekä soveltaa sille asetettuja muita sääntöjä. Jos lähettävä verkkopalvelin tunnistetaan ja roskapostitestaus menee lävitse, voidaan sähköposti toimittaa vastaanottajalle.

Verkkoavaimet tarkistetaan aina vastaan ottavassa sähköpostipalvelimessa, mutta loppukäyttäjän sähköposti ohjelmaakin voidaan muokata tarkistamaan näitä tunnisteita ja reagoimaan saamiinsa tuloksiin [Delany, 2007].



Kuva 12. Verkkoavaimen käytön toimintaperiaate [Delany, 2007].

Verkkoavaimet tarkistetaan aina vastaan ottavassa sähköpostipalvelimessa, mutta loppukäyttäjän sähköposti ohjelmaakin voidaan muokata tarkistamaan näitä tunnisteita ja reagoimaan saamiinsa tuloksiin.

8. Yhteenveto ja johtopäätökset

Tässä tutkimuksessa on esitelty roskapostin tunnistamiseen ja roskapostin lähettäjän tunnistamiseen perustuvien menetelmien periaatteita. Roskapostin tunnistamiseen perustuvat menetelmät on jaettu neljään eri luokkaan: sisältöperustaiseen, tilastollisiin menetelmiin perustuviin, heuristisiin ja tarkistenumeroperustaisiin suodattimiin. Näistä suodattimista tilastolliset menetelmät ovat tällä hetkellä kaikkein suosituimpia. Bayes-tekniikka, joka perustuu roskapostin todennäköisyyden laskemiseen sähköpostissa esiintyvien yksittäisten sanojen avulla, on ylivoimaisesti eniten käytetty. Tämä johtunee siitä, että Bayes-tekniikka on suhteellisen helppo ymmärtää ja implementoida sekä verrattain tehokas, nopea ja helposti automatisoitavissa. Roskapostin lähettäjän tunnistamiseen perustuvat menetelmät voidaan karkeasti jakaa mustalistaukseen ja valkealistaukseen, jotka ovat toistensa vastakohtia. Nämä määrittelevät, onko lähetetty sähköposti roskapostia vai ei.

Jokaiseen yksittäiseen menetelmään löytyy keino kiertää suodatin. Roskapostin estomenetelmien heikkouksia ovat mm.

- Pelkkä sisältöperusteinen suodatin ei tunnista tahallaan väärin kirjoitettuja sanoja.
- Tilastollisiin menetelmiin perustuvat suodattimet erehtyvät roskapostista, jossa on paljon varsinaiseen viestiin kuulumatonta tekstiä, joka muuttaa viestin roskapostin todennäköisyyden vääräksi.
- Heuristisia suodattimia voidaan harhauttaa pitkälti samoilla menetelmillä kuin tilastollisia suodattimia. Lisätyn ylimääräisen tekstin on oltava ”järkevää”. Oppimis- ja opettamisominaisuuden vuoksi suodatin on pidempään käyttökelpoinen.

- Tarkistenumeroperustaista suodatinta voidaan harhauttaa myös lisäämällä roskapostiin ylimääräistä tekstiä. Lisättävä teksti pitää vain olla erilainen jokaisessa yksittäisessä roskapostissa.
- Mustalista saattaa sisältää oikeita sähköposteja, jos internetpalvelun tarjoaja on joutunut hyökkäyksen tai muun laittomuuden kohteeksi.
- Valkealistalla olevasta IP-osoitteesta on laittomasti onnistuttu lähettämään roskapostia.

Tällä hetkellä paras ratkaisu on käyttää useiden estomenetelmien yhdistelmiä. Esto olisi myös tehokkainta tehdä mahdollisimman aikaisin, ennen kuin haitallinen posti pääsee kulkemaan pitkän matkan tietoliikenneverkossa. Näin säästyttäisiin roskapostin verkolle aiheuttamalta kuormalta. Mahdollisesti jo internetpalveluntarjoaja tai välityspalvelun tarjoaja voisivat poistaa havaitsemansa haittapostin heti sen havaittuaan.

Aloittaessani tätä tutkimusta keräsin itselleni tulleista sähköposteista suuren määrän aineistoa, jota suunnittelin käyttäväni erilaisten roskapostin suodatus ja estomenetelmien testaukseen. Huomasin kuitenkin nopeasti, että roskapostien luonne muuttuu jatkuvasti. Mitä myöhäisemmän vaiheen sähköposteja keräsin, sitä vaikeampi niistä oli tunnistaa roskaposteja suodattimilla tai muillakaan menetelmillä. Roskapostin lähettäjät olivat keksineet uusia keinoja suodattimien harhauttamiseksi. Samoin oma sähköpostini on myös suodatettua, joten kaikkia minulle lähettyjä sähköposteja en saanut myöhempään aineistooni. Tilanteen muuttuessa nopeasti päätin luopua empiirisestä tutkimuksesta.

Roskapostin määrä kaikesta sähköpostista saadessani tämän työn valmiiksi oli laskennallisesti 86 % ja joissakin lähteissä annetaan jopa 90 - 95 % lukuja. Suuria määriä verkkokapasiteettia, reitittimiä, palvelimia, ohjelmistoja ja muistitilaa käytetään vain, että roskaposti ei tukkisi koko internetin viestijärjestelmää.

Roskapostin saamiseksi kuriin kannattaa taistella, sillä siitä hyötyvät lähes kaikki osapuolet. Pois lukien tietysti roskapostilla mainostavat tahot ja roskapostin lähettäjät.

Uusi ongelma on myös kannettavissa laiteissa, kuten matkapuhelimissa tai kämmentietokoneissa, yleistyvä reaaliaikainen sähköposti. Kaikki ilmarajapinnassa tapahtuva liikennöinti maksaa käyttäjälleen rahaa, joko yhteysmaksuna tai datan määränä. Jos matkapuhelimeen tulevan roskapostin määrä on huomattava, joutuu vastaanottaja myös maksamaan tämän roskapostin vastaanottamisesta. Japanissa yleisesti käytössä oleva langaton sähköposti on jo nyt aiheuttanut lainsäädännön tiukentumisen roskapostin lähettäjiä vastaan. Tähän puututaan aikanaan varmasti myös Suomessa sekä muualla maailmassa.

Nykyinen sähköpostijärjestelmä ei ole suunniteltu sellaiseen käyttöön kuin siltä odotetaan nyt, eikä se tue roskapostin estämistä. Seuraavan sukupolven sähköpostin pitää pystyä vastaamaan näihin haasteisiin. Pitää luoda järjestelmä jossa roskapostin lähettäminen on tehty mahdottomaksi.

Viiteluettelo

[Assis et al.,2005] Fidelis Assis, William Yerazunis, Christian Siefkes and Shalendra Chhabra, CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track, *NIST Text REtrieval Conference (TREC)*, (November 2005). Available: http://crm114.sourceforge.net/NIST_TREC_2005_paper.-pdf, Checked: 09.02.2007.

[Barracuda Networks, 2007], Barracuda Networks Releases Annual Spam Report 12.12.2007, Available http://www.barracudanetworks.com/ns/-news_and_events/index.php?nid=232, Checked 18.12.2007

[CANSPAM Act, 2003] The CAN-SPAM Act of 2003, Pub. L. No. 108-187, 117 Available: http://www.spamlaws.com/federal/can-spam.shtml, Checked: 2.5.2007.

[Clyman, 2004] John Clyman, Understanding Directory Harvest Attacks, *PC Magazine* 4/6/2004, **23** 6, (2004), 64.

[Cole, 2005] Cole, Eric. *Network Security Bible.* Hoboken, NJ, USA: John Wiley & Sons, Incorporated, 2005

[Cormack and Lynam, 2006] Gordon Cormack and Thomas Lynam TREC 2005 Spam Track Overview. In: *The Fourteenth Text Retrieval Conference (TREC 2005) Notebook*, 2006.

[Delany, 2007] Delany Mark, *Description of DomainKeys*, Available: http://antispam.yahoo.com/domainkeys, Checked 23.10.2007

[Duntemann, 2004] Duntemann, Jeff. *Degunking Your Email, Spam, and Viruses.* Scottsdale, AZ, USA: Paraglyph, 2004.

[Goodman et al., 2007] Joshua Goodman, Gordon Cormack and David Heckerman, Spam and the ongoing battle for the inbox. *Communications of the ACM* **50,** 2 (February 2007), 25-33.

[Graham, 2002] Paul Graham, A plan for spam, Presented at *2004 Spam Conference*, Cambrigde, UK. Available: http://www.paulgraham.com/-spam.html Checked 02.02.2007.

[Hauben, 2007] Michael Hauben, *History of ARPANET*. Available: http://www.-dei.isep.ipp.pt/~acc/docs/arpa.html, Checked 21.12.2007

[Hughes, 1998] Lawrence E. Hughes, *Internet E-Mail : Protocols, Standards and Implementation.* Artech House, 1998

[Jokinen, 2007] Lasse Jokinen, haastattelu 13.11.2007.

[McCarthy, 2005] Vance McCarthy, Sendmail Raises Bar on Integrated Mail Security, Integration Developers News, 2005, Available: http://-www.idevnews.com/IntegrationNews.asp?ID=170, Checked: 20.4.2007.

[Peter, 2004] Ian Peter, *The Internet History Project, 2004*. Available: http://www.-nethistory.info/History%20of%20the%20Internet/email.html, Checked 2.1.2008

[Postini, 2007] Postini Corporation, *Daily Email Statistics of Spam, Virus and Harvest Accacks*, Available: http://www.postini.com/stats/index.php, Checked 6.11.2007

[Roberts, 2006] Paul F Roberts, Birth of a killer application. *InfoWorld* **28**, 44 (October 2006), 14.

[Sand, 2002] Paul A. Sand, Spam Filtering with SpamAssassin. Available: http://pubpages.unh.edu/notes/spamassassin.html , Checked 20.4.2007.

[Spammer-X, 2004], Spammer-X, *Inside the SPAM Cartel*, Syngress Publishing, Inc, 2004

[SPF, 2007] Sender Policy Framework, Available: http://www.openspf.org/, Checked 2.11.2007

[Sähköisen viestinnän tietosuojalaki, 2004] Sähköisen viestinnän tietosuojalaki 16.6.2004/516. Suoramarkkinointi 7 luku, 26 § Suoramarkkinointi luonnol-liselle henkilölle. Saatavissa: http://www.finlex.fi/fi/laki/ajantasa/2004/-20040516 , Tarkistettu: 2.5.2007.

[Vohra, 2005] Kaiesh Vohra, *The Identification of Unsolicited Electronic Mail*. University of Edinburgh School of Informatics (2005). Available: http://www.kaiesh.com/anna/KaieshVohra2005-Antispam.pdf, Checked 20.4.2007.

[Zdziarski, 2005] Jonathan A Zdziarski, *Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification.* No Starch Press, 2005.