Seminar

Technischer Datenschutz in Kommunikationsnetzen
(Kurs 02560)

Wintersemester 2004/05
03. Mai 2005

Lehrgebiet Kommunikationssysteme
des Fachbereichs Elektrotechnik und Informationstechnik
der FernUniversität Hagen

Spam-Problematik

Autor: Martin Eller

Betreuer: Prof. Dr.-Ing. Firoz Kaderali
Dipl.-Ing. Alex Didier Essoh

Inhaltsverzeichnis

1 Spam: Eine Einführung

1.1 Der Begriff „Spam“

Der Begriff „Spam“ bezeichnet eigentlich Dosenfleisch der Firma Hormel Foods Corporation, USA. Dabei handelt es sich um ein Kunstwort bestehend aus „**s**piced **p**ork and h**am**“. Der Bezug zur Problematik der unverlangt zugesandten E-Mails ergibt sich aus einem Sketch der britischen Komikertruppe „Monty Python“:

> »Der Sketch spielt in einem Restaurant, dessen Speisekarte ausschließlich Gerichte mit Spam enthält: Der Dialog zwischen dem Gast, der ausdrücklich ein Gericht ohne Spam bestellen möchte, und der Kellnerin, die immer wieder neue Speisenvariationen mit Spam vorschlägt, wird permanent durch einen intonierenden Wikinger-Chor unterbrochen. Insgesamt kommt das Wort „Spam“ mehr als 120 mal in dem Sketch vor. Die lautstarken Gesänge des Chores machen am Ende des Sketches jegliche Unterhaltung im Restaurant unmöglich - genau so, wie Spam-Mails die Kommunikation per E-Mail erschweren.«[1]

Spam

Aus dieser Analogie entstand der Begriff „Spam“ für das Medium Internet: Erst im Bereich des Usenet (News), später auch im Bereich der E-Mail-Übertragung; stets bezieht er sich auf die massenhafte Versendung von Artikeln/E-Mails.

Eine weitere „Übersetzung“ des Begriffes „Spam“ für den Bereich des Internets lautet: *Send phenomenal amounts of mail.*

Der eigentlich fachlich korrekte Ausdruck für diese unverlangt zugesandte E-Mails ist **UCE** (Unsolicited Commercial E-Mail) oder **UBE** (Unsolicited Bulk E-Mail), ich halte mich im weiteren Verlauf dieses Textes jedoch an den gebräuchlicheren Ausdruck „Spam“.

UBE/UCE

Das sind viele Namen für ein und das selbe Problem: Massenhaft versandte E-Mails mit Werbung für Medikamente, Kredite, dubiose Softwarelizenzen, etc. Wohl jeder, der eine E-Mail-Adresse einmal öffentlich gemacht hat, sei es durch das Posten von Beiträgen im Usenet, durch Anmeldung bei Newslettern, oder durch Präsentation auf einer Webseite, ist wohl mit der Problematik schon in Berührung gekommen und hat solche lästigen Spam-Mails erhalten.

Doch warum überhaupt gibt es Spam und wieso ist es so schwer, der Problematik Herr zu werden? Im Folgenden werde ich auf diese Frage näher eingehen.

1.2 Motive

Zunächst einmal stellt sich natürlich die Frage, was Spam für den Absender überhaupt interessant macht.

Durch den Inhalt einer Spam-E-Mail werden Produkte oder Webseiten beworben[1].

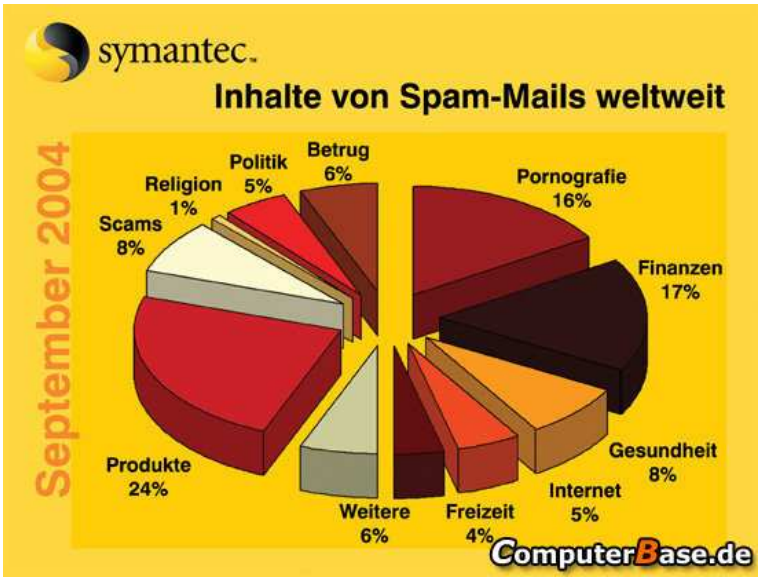

Abbildung 1: Inhalte von Spam

Die Produkte sollen bestellt, bzw. die Webseiten sollen per Browser „angesurft“ werden. Durch die Bestellung und den Verkauf des Produktes bzw. durch den Aufruf der Webseite (Werbung/Dialer) von möglichst vielen verschiedenen Stellen aus, verdient der Absender Geld. Für einen möglichst hohen Gewinn muss die zuvor getätigte Investition möglichst gering sein. Da heutzutage ein breitbandiger Internetzugang keine hohen Kosten und der

[1] Mit Spam wird hauptsächlich für Dinge geworben, für die es sich die Investition in „übliche“ Werbung nicht lohnt, weil das beworbene Produkt entweder praktisch wertlos oder aber sogar illegal ist

E-Mail-Versand keine Portokosten verursacht, ist die Nutzung des Mediums Internet-E-Mail sehr reizvoll.

Wenn man z.B. 2 Millionen Spam-E-Mails versendet und nur 4 Empfänger das beworbene Produkt für z.B. 50$ erwerben, sind dies 200$ Einnahmen bei sehr geringen Ausgaben.

1.3 Ursache: Das SMTP-Protokoll (RFC 2821)

Wie der Name (*Simple* Mail Transport Protocol) schon andeutet, soll mit Hilfe dieses Protokolls in erster Linie ein *einfacher* Weg zur Übermittlung von E-Mail bereitgestellt werden.

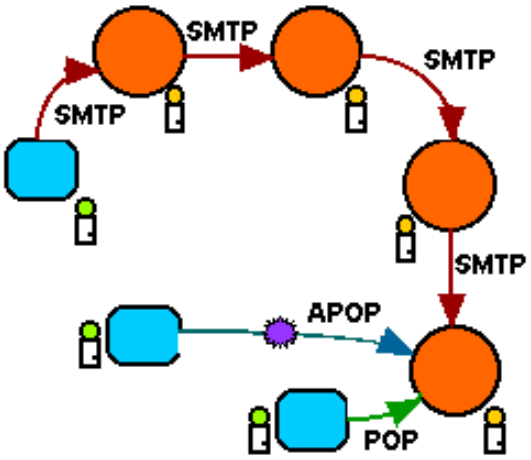

Abbildung 2: E-Mail-Übertragung (Prinzip)

Die Kommunikation läuft hierbei über das Anwendungsprogramm des Anwenders (MUA[2]) zum lokalen Mailserver (MTA[3]), der die E-Mail zum jeweiligen Mailserver des Empfängers weiterleitet (Relay). Die Adresse des zuständigen Mailservers erfährt der MTA ggf. über die Abfrage des MX-Eintrages der Empfängerdomäne vom DNS[4]. Die E-Mail kann durchaus über mehrere Mailserver geleitet werden, bis sie ihr Ziel erreicht. Der Mailserver des Empfängers nimmt die E-Mail schließlich an und speichert sie, der Empfänger

[2]Mail User Agent
[3]Mail Transfer Agent
[4]DNS: Domain Name Server

holt sie mit seinem Anwendungsprogramm dann von diesem Mailserver ab. Bei der Abholung kommt als Protokoll meist POP3 oder IMAP zum Einsatz, die restlichen Mailübertragungen werden per SMTP abgewickelt.

Auf Seiten des MTA werden hierbei zwei Arten der E-Mail-Annahme unterschieden: Die Annahme zur Weiterleitung (Relay) und die Annahme zur (lokalen) Auslieferung (Delivery).

Ein typischer Ablauf der E-Mail-Übertragung vom versendenden Client (MUA) zum Mailserver (MTA) ist in der Abbildung 3 dargestellt. Hierbei wird aus Gründen der Einfachheit davon ausgegangen, dass nur zwei MTAs an der Übertragung beteiligt sind.

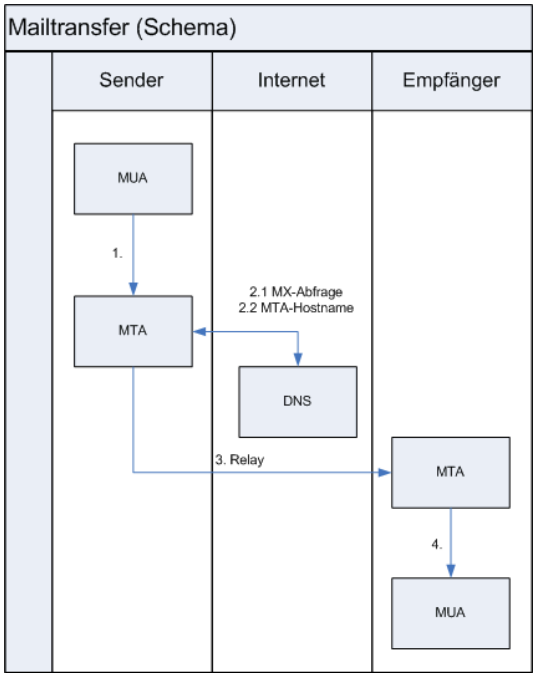

Abbildung 3: E-Mail-Übertragung

Die Kommunikation zwischen MUA und MTA geht aus Abbildung 4 hervor.

Die Antworten des MTA beginnen grundsätzlich mit einem dreistelligen Zahlencode. Dieser Code wird softwareseitig vom MUA ausgewertet, die Bedeutung der Zahlencodes ist Bestandteil des Protokolls SMTP.

In 1: ist hier die typische Begrüssungszeile eines MTA abgebildet.

```
 1. 220 port-goran.ehoch2.de ESMTP Sendmail 8.12.10/8.12.10/SuSE Linux
    0.7; Wed, 23 Mar 2005 17:51:57 +0100

 2. EHLO albernia.ehoch2.de 0

 3. 250-port-goran.ehoch2.de Hello albernia.ehoch2.de [192.168.0.100],
    pleased to me et you
    250-ENHANCEDSTATUSCODES
    250-PIPELINING
    250-8BITMIME
    250-SIZE
    250-DSN
    250-ETRN
    250-DELIVERBY
    250 HELP


 4. MAIL FROM: martin.eller@ehoch2.de

 5. 250 2.1.0 martin.eller@ehoch2.de... Sender ok

 6. RCPT TO: ellerm@bkk-bv.de

 7. 250 2.1.5 ellerm@bkk-bv.de... Recipient ok

 8. DATA

 9. 354 Enter mail, end with ".” on a line by itself

10. From: Martin Eller <martin.eller@ehoch2.de>
    To: Eller, Martin <ellerm@bkk-bv.de>
    Subject: Demo

    Hallo Martin,
    dies ist eine kurze Demo.
    Gruss,
    Martin
    .

11. 250 2.0.0 j2NGpvCl012688 Message accepted for delivery

12. QUIT
```

Abbildung 4: Transscript einer SMTP-Sitzung

In 2: folgt die eigentliche SMTP-Kontaktaufnahme auf ISO-Layer 7: Mit EHLO <hostname> meldet sich der sendende Client (MUA) am MTA an.

In 3: ist die entsprechende Serverantwort zu sehen. Der MTA bestätigt die Kontaktaufnahme und listet auf, welche erweiterten Befehle er anbietet. Im Rahmen dieses Textes werde ich auf diese erweiterten Befehle jedoch nicht näher eingehen.

In 4: fährt der MUA mit der Übertragung fort und übermittelt zunächst die Absenderadresse.

In 5: erfolgt wiederum die Bestätigung des MTA: Der Absender ist berechtigt, E-Mails über diesen Server einzuliefern[5].

In 6: Übergibt der MUA die Empfängeradresse. An dieser Stelle könnten auch mehrere RCPT TO:-Zeilen folgen.

In 7: bestätigt der MTA den Empfang der Adresse.

In 8: teilt der MUA mit, dass ab jetzt die eigentliche E-Mail folgt.

In 9: bestätigt der MTA dies.

In 10: Folgt nun die E-Mail: Sie teilt sich auf in zwei Teile, die Header und den Body. Die Header bestehen hier aus den (nochmaligen) Angaben des Absenders, des Empfängers und des Betreffs. Abgetrennt durch eine Leerzeile folgt dann der Mailbody. Das Ende des Bodys wird durch eine Zeile, die nur einen ».« enthält.

In 11: bestätigt der MTA den Empfang der E-Mail.

In 12: beendet der MUA die Verbindung mit QUIT.

Besonderes Augenmerk möchte ich hier auf drei Zeilen lenken: `EHLO`, `MAIL FROM:`, und `From:`. In diesen Zeilen identifiziert sich der Absender bzw. der einliefernde Client beim Mailserver.

keine Überprüfung

SMTP sieht keine Überprüfung der hier gemachten Angaben vor, so dass hier grundsätzlich beliebige Angaben gemacht werden *können*. Die einzige verlässliche Information ist die meist, aber nicht immer angegebene IP-Adresse des einlieferndes Rechners in der Antwortzeile des Mailservers nach dem EHLO-Befehl.

Hier ist es der annehmenden MTA-Software überlassen, aus welchen IP-Adressbereichen sie E-Mails zur Weiterleitung entgegen nimmt und ob sie

[5] Dieser Test ist *nicht* im SMTP vorgeschrieben.

die Angaben bzgl. des Absenders prüft (z.B. per DNS-Anfrage zur Domain des mutmaßlichen Absenders).

Zusätzlich können an Stelle nur einer `RCPT TO:` Zeile beliebig[6] viele angegeben werden. So muss der Mailtext nur einmal übertragen werden, die Vervielfältigung an alle Adressaten übernimmt der Mailserver.

Von Seiten des Versenders von Spam-E-Mails liegen die Vorteile auf der Hand:

offener Relay

Ist erst einmal ein Mailserver gefunden, der ungeprüft Mails zur Weiterleitung annimmt (ein sogenannter Offener Relay), genügt es, dem Server alle Adressen und einmal die Werbebotschaft zu übertragen. Idealerweise wird noch eine Adresse aus dem Adresspool als (gefälschte) Absenderadresse ausgewählt und zu Beginn der Werbebotschaft noch ein paar erfundene Mailheader ergänzt, schon ist es dem ungeübten E-Mail-Empfänger nahezu unmöglich, den wahren Absender der Spam-E-Mail zu ermitteln.

Ganz nebenbei werden die ohnehin recht geringen Kosten der Massenübertragung dem Betreiber des offenen Relays überlassen...

1.4 Problematik

Wie auch aus Abbildung 5 hervorgeht, steigt der Spam-Anteil des gesamten E-Mail-Volumens stetig an.

Ham

Heute wird oftmals mehr Zeit dafür verwendet, Spam von normaler E-Mail („Ham“ genannt) zu trennen, als für das Lesen und Bearbeiten der E-Mail, was natürlich gerade im professionellen Umfeld sehr ärgerlich ist.

Bot-Netze

Eine weitere, erst neuerdings verstärkt auftretende Erscheinung ist die Nutzung sogenannter Bot-Netze zum Spam-Versand. Hierbei handelt es sich um Systeme, die durch Ausnutzung von Exploits oder Fehlkonfigurationen kompromittiert wurden und nun z.B. durch IRC-Channels[7] ferngesteuert werden. Diese Bot-Netze werden z.B. an Spam-Versender quasi als „Wegwerf-Mailserver“ vermietet, die hierüber ihre Spam-Mails versenden können, ohne durch Real Time Blacklists geblockt zu werden.

[6] Begrenzt wird die Zahl de facto durch die annehmende Software

[7] IRC: Internet Relay Chat. Ein Dienst, der textbasierte Echtzeitkommunikation per Internet zur Verfügung stellt

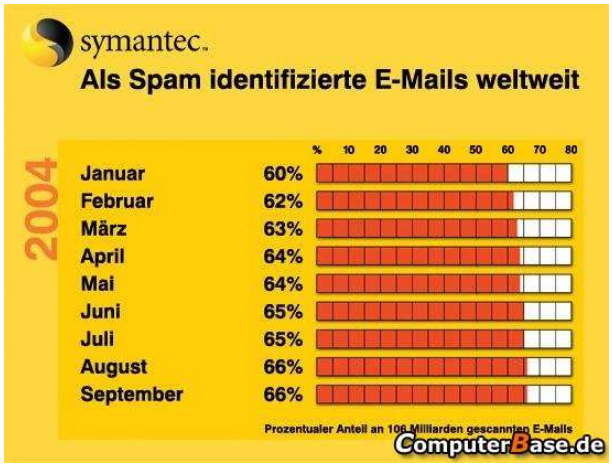

Abbildung 5: Anteil von Spam am E-Mail-Verkehr

2 Aktuelle Gegenmaßnahmen

Um die Postfächer weitestgehend spam-frei zu halten, wurden verschiedene Gegenmaßnahmen entwickelt, von denen ich im folgenden einige beschreibe.

2.1 Real Time Blacklists (RTBs oder RBLs)

Real Time Blacklists sind Listen mit IP-Adressen von Mailservern, die als offene Relays bzw. Proxies erkannt worden sind, oder die bereits anderweitig als Spamversender aufgefallen sind. Diese Listen werden per Internet zur Verfügung gestellt und vom annehmenden Mailserver in Echtzeit abgefragt. Wird die IP-Adresse eines einliefernden Servers in der RBL gefunden, wird die Annahme sämtlicher E-Mails verweigert.

Die E-Mail-Übertragung erfolgt im Prinzip wie bereits in Abbildung 3 aufgezeigt, allerdings wird unmittelbar nach der Anmeldung des Clients mit EHLO dessen IP-Adresse gegen eine RTB abgeglichen und die Verbindung ggf. mit einer Fehlermeldung getrennt.

Da die Zahl der offenen Relays nicht zuletzt durch konsequentes Blacklisting immer weiter abnimmt und die Kosten für übertragene Daten ebenfalls immer weiter sinken, wird in letzter Zeit zunehmend versucht, Spam-Mails direkt auszuliefern, also den Versand nicht mehr einem gefundenen/gekaperten Mailserver zu überlassen, sondern jede einzelne E-Mail über Abfrage des MX-

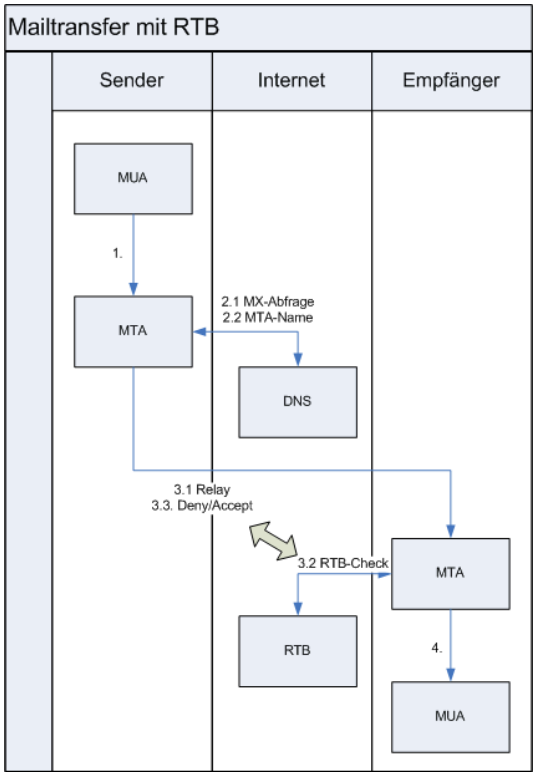

Abbildung 6: Real Time Blacklist (RTB)

Eintrages der Zieldomäne direkt zuzustellen.

2.2 DNS-Domänenabfragen

Wie bereits in der Einführung beschrieben, ist eine Überprüfung der Angaben des Mail einliefernden Servers grundsätzlich nicht vorgeschrieben.

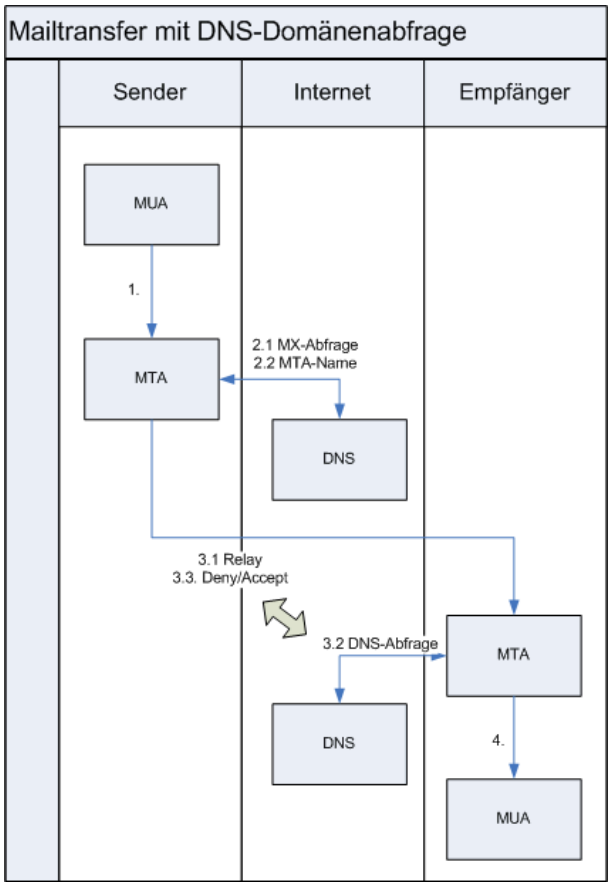

Abbildung 7: DNS-Domänenabfragen

Um die Maileinlieferungen einzuschränken und es Spamversendern schwerer zu machen, führen mittlerweile viele Mailserver bereits während der Einlieferung von E-Mail eine Überprüfung des Rechnernamens des einliefernden Rechners und/oder eine Überprüfung der Domain des angeblichen Absenders durch. Nur wenn die übergebenen Namen per DNS aufgelöst werden können, die entsprechenden Domänen also existieren, wird überhaupt Mail angenommen. Die Abfrage setzt also bereits ein, wenn der RCPT TO:-Eintrag empfangen wird.

Als Reaktion darauf, wurden als Absender nicht mehr beliebig erzeugte Absenderdomänen für Spam genutzt, sondern bekanntermaßen existierende Domänennamen wie z.B. hotmail.com, gmx.net, oder microsoft.com.

2.3 Callouts

Gewissermaßen als Weiterentwicklung der DNS-Domänenabfragen wird heute oft das sogenannte „Sender-Callout“ durchgeführt. Hierbei wird bei Einlieferung einer E-Mail überprüft, ob die angegebene Absenderadresse (Sender) existiert und die Annahme bei nicht existierender Adresse verweigert.

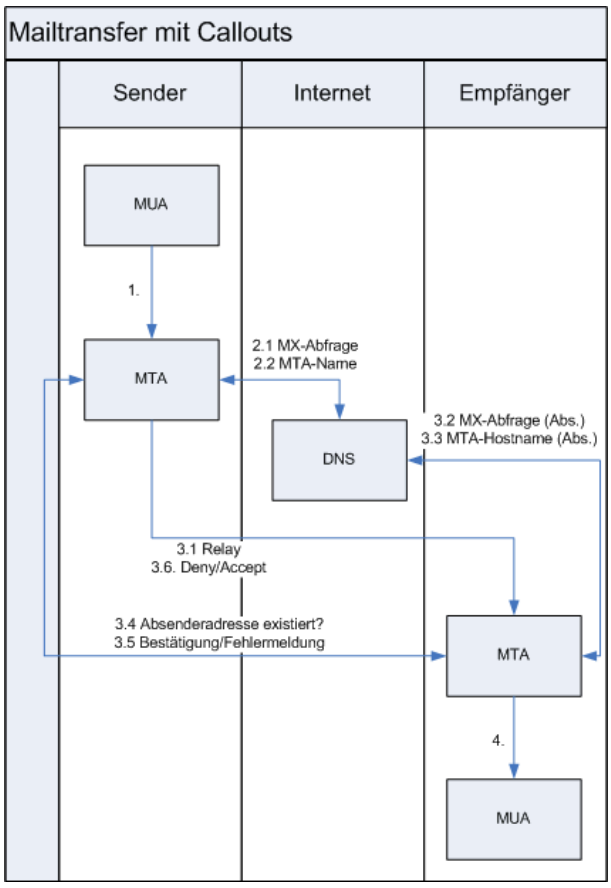

Abbildung 8: Callouts

Als Folge dieser Sender-Callouts werden mittlerweile zunehmend existierende

E-Mail-Adressen als gefälschte Absenderadresse für Spam-Mails verwendet. Die Folge dieses Verhaltens ist, dass der vermeintliche Absender sämtliche Rückmeldungen (Fehlermeldungen, „Delivery Status Notifications (DSN)“ und auch Beschwerden von Empfängern) erhält, ohne dass er sich dagegen wehren kann.

2.4 Filter

Hierbei handelt es sich um das eher „klassische“ Gegenmittel. Eingehende E-Mails werden gegen Filterregeln geprüft und gegebenenfalls aussortiert.

Ursprünglich waren diese Filterregeln statisch und mussten daher regelmäßig an die aktuell eintreffenden Spam-Mails angepasst werden, um weiter effektiv zu arbeiten. Beispiele für solche Regeln sind zum Beispiel die Filterung auf die Worte „sex“ oder „viagra“ im Betreff. Außerdem arbeiteten diese Filter „binär“, kannten also nur die Zustände „ist Spam“ und „ist kein Spam“. Dadurch kam es zwangsläufig zu Fehl-Erkennungen in beiden Richtungen.

Im Laufe der Zeit hat sich ein steter Wettbewerb zwischen Versendern und Filterern von Spam eingestellt, so dass statische Filter in immer kürzeren Abständen anzupassen waren. So gingen Spam-Versender z.B. dazu über, statt „Viagra“ „V I A g R a“ in die Betreffzeile zu schreiben, oder einzelne Buchstaben durch ähnlich aussehende Ziffern (z.B. 0 statt O) zu ersetzen, um statische Filterregeln zu umgehen. Mittlerweile beherrschen die meisten statischen Filter reguläre Ausdrücke, mit denen sich auch solche „Wortspielereien“ erkennen lassen.

Ein anderer Filteransatz sind die „Bayes-Filter“ , die nicht mehr länger binär arbeiten, sondern auf Grund verschiedener Eigenschaften einer E-Mail errechnen, mit welcher Wahrscheinlichkeit es sich hierbei um Spam handelt. Bayes-Filter

> »Bayes'sche Filter sind „lernend“ [...] und setzen auf Worthäufigkeiten in bereits vom Benutzer erhaltenen und klassifizierten E-Mails. Ein bayes'scher Filter wird durch seinen Benutzer trainiert, indem dieser seine E-Mails in erwünschte (Ham) und unerwünschte (Spam) einteilt. Der bayes'sche Filter stellt nun eine Liste mit Wörtern zusammen, die in unerwünschten E-Mails vorkommen. Hat der Benutzer E-Mails mit den Begriffen „Sex“ und „Viagra“ als Spam gekennzeichnet, haben alle E-Mails mit diesen Begriffen eine hohe Spamwahrscheinlichkeit. Begriffe aus erwünschten E-Mails wie „Verabredung“ oder „Bericht“ führen dann

zu geringer Spamwahrscheinlichkeit. Allerdings reichen einzelne Schlüsselworte nicht aus, sondern die Gesamtsumme der Bewertungen der einzelnen Wörter macht es aus.

Der Filter schafft bereits nach kurzem Training mit zirka 30 E-Mails erstaunlich hohe Trefferquoten - auch wenn für die produktive Nutzung mindestens ein paar hundert Mails beider Kategorien empfohlen wird. Er wird von vielen Providern zum Abfangen von Spam verwendet.«[2]

2.5 Distributed Checksum Clearinghouse (DCC)

Unter „Distributed Checksum Clearinghouse“ wird ein völlig anderer Ansatz zur Spambekämpfung gemacht: Die Natur von Spam-Mail ist es bekanntlich, dass ein und dieselbe E-Mail an eine sehr große Anzahl von Empfängern verschickt wird. Das macht man sich auf zentralen Mailservern zu Nutze und bildet Checksummen über alle einlaufenden E-Mails und meldet diese, zusammen mit der Anzahl der Empfänger an eine zentrale Klärungsstelle („Clearinghouse“). Mails, die über einem definierten Schwellwert liegen, werden als Spam klassifiziert (mit Ausnahme von Checksummen, die auf einer Whitelist gepflegt werden). Im Gegenzug können Mailserver zu einer Checksumme abfragen, ob es sich um Spam handelt. [3]

Um diesem Verfahren zu begegnen, sind Spam-Versender allgemein dazu übergegangen, in den Body der Spam-Mail (meist ans Ende) zufällig erzeugte Zahlen/Buchstabenreihen einzufügen, um die Generierung einer eindeutigen Checksumme zu verhindern.

Als Reaktion sind die Verfahren zur Ermittlung der Checksummen unschärfer geworden (Fuzzy-Checksum), was natürlich die Treffergenauigkeit ebenfalls reduziert.

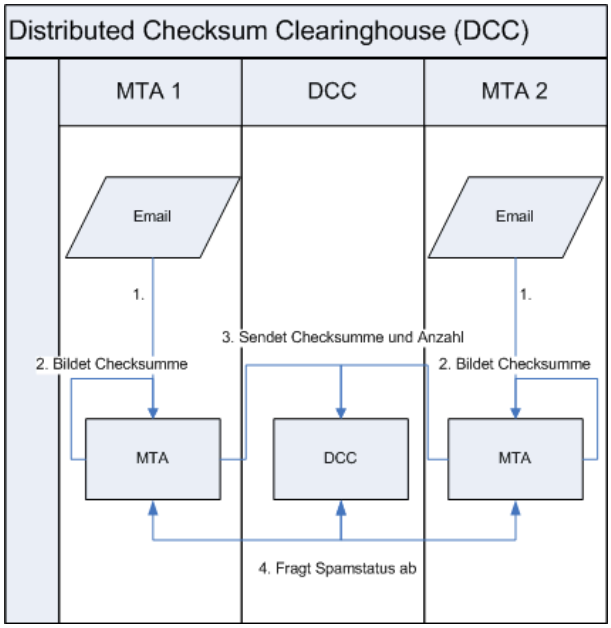

Abbildung 9: Distributed Checksum Clearinghouse (DCC)

2.6 Greylisting

In der Regel versucht ein Spam-Versender nur ein mal pro Adressat, die E-Mail zu senden, außerdem wechselt der Spam-Versender öfters seine IP-Adresse während des Versandes, um die Rückverfolgung zu erschweren.

Genau dieses Verhalten nutzt das Greylist[8]-Verfahren aus, in dem zu jeder E-Mail genau drei Informationen, ein sogenanntes Triplet, untersucht werden: Die IP-Adresse des einliefernden Rechners, die Absender-Adresse und die Empfänger-Adresse.

Hat der Greylist-Server dieses Triplet bereits in seiner Datenbank, wird die E-Mail zugestellt. Ist das Triplet jedoch noch unbekannt, wird die E-Mail mit der Fehlermeldung 471 (Temporarily refused, try again later) RFC-konform zurückgewiesen und das Triplet mit kurzer Verzögerung in die Datenbank eingetragen. Erfolgt nun ein erneuter Zustellversuch[9], so ist das

[8]Das Greylist-Verfahren hat seinen Namen daher, dass es einen Kompromiss zwischen Whitelist- und Blacklistverfahren darstellt

[9]Ein RFC-konformer Mailserver wird nach Erhalt der Fehlermeldung 471 die E-Mail zwischenspeichern und nach einer definierten Zeit einen erneuten Versuch unternehmen

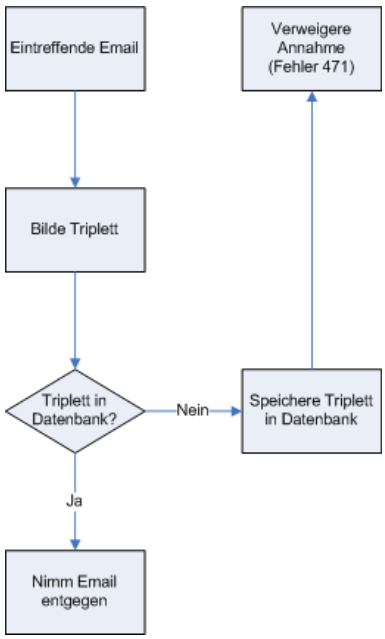

Abbildung 10: Greylisting

Triplet bekannt und die E-Mail wird zugestellt. Meist realisiert man die Greylist-Datenbank so, dass auch Alter und Auftritt-Häufigkeit der Triplets gespeichert werden. So lässt sich die Wartung der Datenbank automatisieren, in dem alte, lange nicht mehr aufgetretene Triplets, oder Triplets von nur einmal versuchter Mailzustellung wieder gelöscht werden. [4]

3 Aktuelle Entwicklungen zur Spam-Bekämpfung

Aktuelle Entwicklungen adressieren die Möglichkeit, Absenderadressen beliebig zu fälschen, in dem sie einen Mechanismus anbieten, die Echtheit der Absenderadresse zu bestätigen.

3.1 SenderID / Sender Policy Frame (SPF)

SenderID (ursprünglich „CallerID“) ist eine Entwicklung von Microsoft zur Überprüfung, ob eine E-Mail aus der Domäne heraus versandt wurde, deren Absender sie trägt. SenderID kommt in erster Linie in der Mailsoftware des Empfängers zum Einsatz und Überprüft den „From:“-Header des Mailbodys. Das Ergebnis dieser Überprüfung wird als zusätzliche Information genutzten Filterprogrammen zur Verfügung gestellt, um die Spam-Erkennung zu verbessern. [5]

„Sender Policy Frame“, ehemals „Sender Permitted From“, verfolgt den gleichen Ansatz wie SenderID, setzt jedoch schon im Mailserverbereich an. Entsprechend wird hier der „MAIL FROM:Header des Mail-Envelopes ausgewertet.

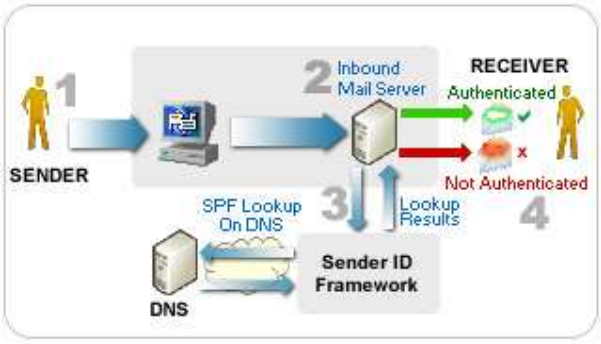

Abbildung 11: Funktionsweise des SPF

Bei beiden Verfahren erfolgt ein Abgleich der IP-Adresse des einliefernden Rechners mit einer Liste von IP-Adressen, die vom Eigentümer der Absenderdomäne als zum Versand von E-Mail berechtigt eingetragen wurden.

Mittlerweile sind beide Verfahren zum „Sender Policy Framework (SPF)“ kombiniert worden, so dass Microsofts SenderID auf das SPF zurückgreift. [6][5]

Literatur

[1] http://www.vis-recht.bayern.de/de/left/themen/aufdraengung/spam-begriff.htm

[2] http://de.wikipedia.org/wiki/Bayesscher_Filter

[3] http://http://www.rhyolite.com/anti-spam/dcc/

[4] http://projects.puremagic.com/greylisting/whitepaper.html

[5] http://www.microsoft.com/mscorp/safety/technologies/senderid/overview.mspx

[6] c't 15/04, S. 142-144

Abbildungsverzeichnis