Technische Filter

Herbert Kuchen, Christian Hermanns, Michael Poldner
European Research Center for Information Systems, Universität Münster
Leonardo Campus 3, 48149 Münster

Zusammenfassung

Der Einsatz von Firewalls und Filtern auf Anwendungsebene ist heutzutage ein notwendiges Instrument, um sich gegen Massen von unerwünschten Werbemails, Phishing und Viren zu schützen sowie Webseiten mit illegalen Inhalten ausfindig zu machen. Die vorliegende Arbeit vermittelt die grundlegende Arbeitsweise von Filtern auf Anwendungsebene, beleuchtet unterschiedliche und historisch gewachsene Filtertechnologien und diskutiert populäre Tricks, mit denen versucht wird, Filter-Software zu umgehen. Die Arbeitsweise von Filtern wird am Beispiel von Spam-Filtern erläutert, ist aber auf andere Anwendungen übertragbar.

I. Einleitung

Mit der Verbreitung und der kommerziellen Nutzung des Internets wurde schnell der Ruf nach brauchbaren Technologien laut, um den lästigen Popup-Werbefenstern und der leidigen Flut von täglichen Werbemails zu entgehen, den heimischen Rechner vor Viren, Würmern und Trojanern und seine Kinder vor Internetseiten mit nicht jugendfreien oder Gewalt verherrlichenden Inhalten zu schützen. Es wird geschätzt, dass mittlerweile in Deutschland alleine rund 500 Millionen Spam-Mails pro Woche verschickt werden, was einem Anteil von über 50 % des Email-Gesamtaufkommens entspricht. Als Reaktion auf diese Problematik versprechen inzwischen eine Vielzahl von Firmen Hilfe in Form von Filter-Software, und bei vielen Postfächern werden mehr oder weniger effektive Spam-Filter vorgeschaltet, um der Flut von E-Mails Herr zu werden. Damit die Emails nicht in den verbreitet eingesetzten Spam-Filtern hängen bleiben, bedienen sich Spammer immer neuer Tricks. Effektives Filtern von Spam basiert auf einem umfassenden Mix von verschiedenen Filtertechnologien, der richtigen Filterarchitektur sowie einer geeigneten Platzierung des Filters.

Die vorliegende Arbeit vermittelt die grundlegende Arbeitsweise von Paket- und Spam-Filtern, beleuchtet unterschiedliche und historisch gewachsene Filtertechnologien und diskutiert populäre Tricks, mit denen Spammer versuchen, Filter-Software zu umgehen.

Der Aufbau der Arbeit gliedert sich wie folgt: Kapitel II befasst sich mit den Grundlagen der Kommunikation im Internet und beschreibt die Arbeitsweise von Firewalls. Kapitel III gibt einen kurzen, historischen überblick über populäre Spam-Filter und diskutiert deren zugrunde liegende Technologien, Stärken und Schwächen. Kapitel IV beschäftigt sich mit Angriffen auf aktuelle statistische Spam-Filter. Eine kurze Zusammenfassung bietet Kapitel V.

II. Grundlagen der Kommunikation im Internet

In den folgenden Abschnitten werden die technischen Grundlagen der Kommunikation im Internet erläutert, der Aufbau und die Eigenschaften von IP-Paketen sowie die Network Address Translation beschrieben und die grundlegende Funktionsweise von Firewalls vorgestellt. Eine gute Einführung in dieses Themengebiet bieten [3,6].

1. Das Internet

Das Internet ist ein weltweiter Zusammenschluss von Computernetzwerken und ermöglicht über das Internet Protocol (IP) einen paketbasierten Austausch von Daten. Als „Netzwerk von Netzwerken“ besteht das Internet aus einer Vielzahl an privaten, öffentlichen, akademischen und staatlichen Netzwerken, die zusammen eine große Menge unterschiedlicher Dienste zum Austausch von Informationen zur Verfügung stellen. Zu den bekanntesten Diensten zählen unter anderem Email, Online Chat, Dateiübertragung und World Wide Web (WWW).

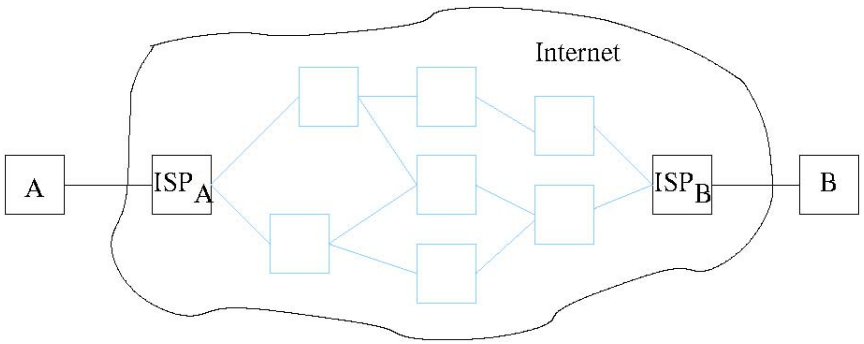


Abbildung 1: Das Internet, ein Zusammenschluss von Netzwerken.

Die Begriffe WWW und Internet werden oft fälschlicherweise als Synonyme verwendet. In Wahrheit handelt es sich beim WWW nur um eine von vielen Anwendungen im Internet. Um ein Privat- oder Firmennetzwerk mit dem Internet zu verbinden, werden häufig die Dienste eines Internet Service Providers (ISP) in Anspruch genommen. Ein ISP ermöglicht den leitungsbasierten Zugang (über DSL- oder ISDN-Leitungen) zum Internet (vgl. Abb. 1).
Die Tabelle 2 zeigt eine Übersicht über die wichtigsten Dienste des Internets, sowie deren Anteil am Gesamtvolumen des Datentransfers.

DIENST	PROTOKOLL	PORTS	ANTEIL
World Wide Web	HTTP	80	45%
Email	SMTP, POP3, IMAP	25, 110, 143	12 %
Peer-to-Peer-Systeme	eDonkeye, Gnutella, etc		24 %
Telefonie(VoIP)	H.323, SIP	5060	7 %
Streaming-Dienste			7 %

Tabelle 2: Anwendungen und Protokolle

2. Kommunikation im Internet

Um die über das Internet zur Verfügung gestellten Dienste nutzen zu können, bedarf es der Kommunikation zwischen dem Anbieter eines Dienstes (Server) und einem Dienstnutzer (Client). Sowohl der Client als auch der Server werden durch Anwendungssoftware und nicht durch Hardware repräsentiert. Beide nehmen in einer Kommunikationsbeziehung eine gleichberechtigte Rolle ein. Wenn Programme sowohl Client- als auch Server-Tätigkeiten durchführen, dann spricht man von einer Peer-to-Peer-Kommunikation. Damit zwei Programme über das Internet kommunizieren können, müssen sie in die Lage versetzt werden, Nachrichten bzw. Daten miteinander auszutauschen. Diese komplexe Aufgabe wird von Protokollen übernommen und umfasst je nach Art des Protokolls unterschiedliche Funktionen, wie z.B. die Gewährleistung von Zuverlässigkeit, Sicherheit und Effizienz. Die Probleme, die ein Protokoll zu bewältigen hat, reichen von der konkreten physikalischen Übertragung elektrischer Signale bis hin zu wesentlich abstrakteren Aufgaben, wie der Kommunikation auf der Anwendungsebene. Aufgrund der unterschiedlichen Abstraktionsebenen der Aufgaben, werden diese im *Open Systems Interconnection (OSI)-Referenzmodell* der *International Organization for Standardization* (ISO) verschiedenen vertikal angeordneten Schichten zugeordnet. Tabelle 3 enthält ein vereinfachtes Modell der für die Kommunikation im Internet relevanten Schichten. Die unterschiedlichen Protokolle sind jeweils einer Schicht des OSI-Modells zugeordnet und stehen untereinander in einer vertikalen Kommunikationsbeziehung. Dabei stellt ein Protokoll einer direkt übergeordneten Schicht Dienste zur Verfügung und nimmt die Dienste einer untergeordneten Schicht in Anspruch. Welche Dienste und Funktionen durch die jeweiligen Schichten bereitgestellt werden, kann Tabelle 3 entnommen werden. Die für die Kommunikation im Internet wichtigsten Protokolle sind das Vermittlungsschichtprotokoll IP und das Transportschichtprotokoll TCP, das sich auf IP abstützt. Das TCP-Protokoll erlaubt einem sendenden Programm, eine Nachricht an eine beliebige Anwendung im Internet zu verschicken. Auf der Senderseite wird diese Nachricht in einzelne IP-Pakete zerlegt. IP, welches für die Weiterleitung der Pakete verantwortlich ist, nimmt die IP-Pakete entgegen und leitet sie anhand der IP-Adresse über einen benachbarten Vermittlungsknoten im Netz schließlich an den Rechner des Empfängers weiter. Die einzelnen Pakete können so auf unterschiedlichsten Wegen über das Internet zum Empfänger gelangen. Der IP-Empfangsprozess des Empfängers erkennt, dass die Pakete für die eigene IP-Adresse bestimmt sind und leitet sie daraufhin an das übergeordnete TCP weiter. TCP auf der Empfängerseite ist dafür verantwortlich, die empfangenen Pakete in der richtigen Reihenfolge wieder zusammenzusetzen und die vollständige Nachricht an die Anwendungsschicht weiterzuleiten.

Die Protokolle der Anwendungsschicht, wie HTTP (WWW) und FTP (Datenübertragung), bedienen sich der durch TCP und IP zur Verfügung gestellten Dienste für den Nachrichtenaustausch. Für die Protokolle höherer Schichten ist die Funktionsweise der untergeordneten Schichten dabei vollkommen transparent.

SCHICHT	PROTOKOLLE	DIENSTE
Anwendungsschicht	HTTP, FTP, SMTP, POP3, IMAP, DNS,…	Austausch von Anwendungs-daten (z. B. Webseiten)
Transportschicht	TCP, UDP	Ende zu Ende Verbindung, Segm./Zusammenfügen der Nachrichten in/aus Pakete(n)
Vermittlungsschicht	IP	Vermittlung der Pakete durch das Internet an Empfängerad-resse
Sicherungsschicht		Ggf. Korrektur der physikali-schen Übertragung
Bitübertragungsschicht		Physikalische Übertragung

Tabelle 3: OSI-Schichten der Internetprotokolle

3. Aufbau eines IP-Pakets

IP-Pakete werden auch als Datagramme bezeichnet und stellen das wesentliche Element der Internetkommunikation dar. Sämtliche Nachrichten und Informationen werden in IP-Paketen verpackt durch das Internet geleitet. Aufgrund der beschränkten Größe eines IP-Paketes wird eine Nachricht oft auf mehrere Pakete verteilt. Die Pakete werden von einem Knoten (Rechner) des Internets zum nächsten geleitet, bis sie ihre Zieladresse erreichen. Je nach Netzwerkkapazität können die einzelnen Pakete einer Nachricht dabei vollkommen unterschiedliche Wege durch das Internet nehmen. Diese paketbasierte Kommunikation hat gegenüber der leitungsbasierten Kommunikation den entscheidenden Vorteil, dass die vorhandenen Netzwerkkapazitäten deutlich besser genutzt werden können. Bei einer leitungsbasierten Kommunikation müsste eine exklusive Verbindung vom Sender zum Empfänger der Nachricht geschaltet werden. Da die Netzwerkressourcen für die Dauer des Nachrichtenaustauschs blockiert wären, stünden sie für weitere Kommunikationsverbindungen nicht zur Verfügung.
Ein IP-Paket besteht, wie in Abbildung 4 dargestellt, aus einem Header- und einem Datenteil. Der Header enthält alle für das IP und damit für die Vermittlung des Paketes relevanten Informationen. Zu diesen gehören unter anderem die Länge des Paketes, die Paketnummer, um die vollständige Nachricht zu rekonstruieren, die Lebenszeit, die verhindern soll, dass ein Paket zeitlich unbegrenzt weitergeleitet wird, eine Prüfsumme, die IP-Adresse des Absenders und die IP-Adresse des Empfängers. Die minimale Größe des Headers beträgt 20 Byte, die maximale Größe des gesamten Paketes (Header + Datenteil) beträgt 64 Kilobyte.

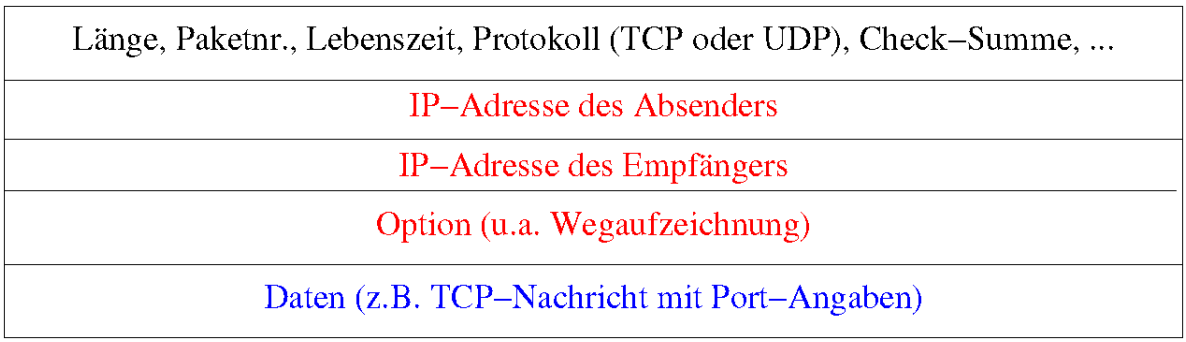


Abbildung 4: Aufbau eines IP-Paketes

4. Eigenschaften von IP-Adressen

Eine IP-Adresse dient zur Identifizierung eines Gerätes in einem IP-Netzwerk, wie dem Internet. Das Internet Protocol Version 4 (IPv4) ist die heutzutage am weitesten verbreitete Version des IP und verwendet IP-Adressen mit einer Länge von 4 Byte (32Bit). Die IPv4 Adressen werden üblicherweise durch vier dezimale Blöcke, von denen jeder für ein Byte steht, zusammengefasst, z.B. 192.168.1.1.

Da die IPv4 Adressen eine Länge von 32Bit haben, gibt es maximal 2^{32} bzw. 4294967296 eindeutige IP-Adressen. Viele dieser Adressen sind in der Praxis jedoch nicht nutzbar, da sie für Sonderfunktionen oder einzelne Organisationen reserviert sind. Weil zu Beginn der Entwicklung des Internets nicht davon ausgegangen wurde, dass das Internet einmal die heutige Größe erreichen würde, wurden große Adressräume einzelnen (meist amerikanischen oder europäischen) Organisationen zugeteilt, die heute nur einen Bruchteil dieser Adressen benötigen. Das rasante Wachstum des Internets hat dazu geführt, dass die öffentlichen IP-Adressen besonders im heutigen IT-Wachstumsmarkt Asien immer knapper werden. Zur Lösung dieses Problems sind verschiedene Techniken für den sparsamen Umgang mit IP-Adressen entwickelt worden. Zu diesen gehören unter anderem die dynamische Vergabe von IP-Adressen und die Network Address Translation (NAT) (s. Abschnitt 2.5).

Bei der dynamischen Vergabe von IP-Adressen wird eine IP-Adresse einem Gerät nur für die Dauer der Verbindung mit einem IP-Netzwerk, wie dem Internet, zugewiesen. Beendet ein Gerät seine Verbindung zum Netzwerk, so wird diese Adresse wieder freigegeben und kann einem anderen Gerät zugeordnet werden. Bei diesem Verfahren kann einem Gerät bei jeder Verbindung mit dem Internet eine andere IP-Adresse zugewiesen werden. Im Gegensatz zur statischen Vergabe von IP-Adressen, bei der einem Gerät eine IP-Adresse für einen theoretisch unbegrenzten Zeitraum zugewiesen wird, reduziert sich durch dieses Verfahren die Gesamtzahl der gleichzeitig benötigten Adressen. Die dynamische Adressvergabe wird vor allem von ISPs eingesetzt, die dadurch für die Geräte ihrer Kunden (dies können je nach Größe des ISPs mehrere Millionen sein) einen deutlich geringeren Adressraum bereithalten müssen.

Der unzureichende Adressraum des IPv4 ist ein entscheidender Grund, der zur Entwicklung des IPv6 geführt hat. Das IPv6 verwendet für die IP-Adressen 128Bit (16Byte), dies entspricht 6,65 10^{23} Adressen pro Quadratmeter der Erdoberfläche. Es ist allerdings offen, wann IPv4 durch IPv6 abgelöst wird.

5. Network Address Translation

Network Address Translation (NAT) bezeichnet ein Verfahren, bei dem ein Unternehmen nach außen hin nur eine einheitliche IP-Adresse verwendet und intern mit lokalen IP-Adressen arbeitet. Bei einer Kommunikation mit der Außenwelt werden die lokalen IP-Adressen in eine einheitliche IP-Adresse umgewandelt und umgekehrt. Wie bereits erwähnt, wird die NAT dazu eingesetzt, um knapper werdende öffentliche IP-Adressen einzusparen. Obwohl das private Netzwerk nach außen nur eine einzige öffentliche IP-Adresse besitzt, ist es mit Hilfe der NAT dennoch möglich, gleichzeitig mehrere Verbindungen zwischen den Geräten des privaten Netzwerks und beliebigen Geräten der Außenwelt herzustellen. Um ein ankommendes Paket an den richtigen Empfänger im privaten Netzwerk weiterleiten zu können, protokolliert die NAT-Einheit die IP-Adressen sowie die Ports der bestehenden TCP/IP-Verbindungen in einer NAT-Tabelle.

Die Funktionsweise der NAT wird durch die Abbildung 5 verdeutlicht, die den schematischen Aufbau eines Unternehmensnetzwerkes zeigt. Das Netzwerk besteht aus mehreren Geräten, die eine lokale IP-Adresse besitzen. Eines dieser Geräte ist der Router, der eine Vermittlerrolle zwi-

schen den Geräten des privaten Firmennetzwerkes und anderen Netzwerken des Internet übernimmt. Eingehende IP-Pakete werden vom Router anhand der IP-Adresse an die entsprechenden Gräte des Firmennetzwerkes vermittelt, und ausgehende IP-Pakete werden von den Geräten des Firmennetzwerkes über den Router an den Router eines weiteren externen Netzes übergeben. Besäße jedes der Geräte des Firmennetzwerkes eine eigene öffentliche IP-Adresse, so könnte die Kommunikation der Geräte des Firmennetzwerks mit anderen Geräten des Internets allein durch einen Router gewährleistet werden. Das Unternehmen besitzt jedoch für die Kommunikation mit anderen Geräten des Internets nur eine einzige öffentliche IP-Adresse (194.8.7.4), wodurch die gleichzeitige Kommunikation mehrerer Geräte des Firmennetzwerkes mit externen Geräten nicht ohne weiteres möglich ist.
Hierfür wird eine NAT-Einheit benötigt, welche die IP-Adressen der zwischen dem Firmenrouter und dem externen Router ausgetauschten IP-Pakete manipuliert. Angenommen einem Firmenrechner mit der lokalen IP-Adresse 10.0.0.1 würde eine Webseite von einem Host mit der IP-Adresse 128.0.0.4 anfragen. Der Firmenrechner würde in diesem Fall das IP-Paket, welches unten links in der Abbildung 5 dargestellt ist an den Firmen-Router senden. Das IP-Paket enthält die eigene IP-Adresse, die IP-Adresse des Hosts, von dem die Webseite geladen werden soll, sowie ein TCP-Paket, das neben den Ports von Sender und Empfänger die Webseitenanfrage an den Webserver des Ziel-Hosts enthält. Die Portnummern dienen dazu, die miteinander kommunizierenden Programme auf beiden Rechnern zu identifizieren. Auf dem Firmenrechner handelt es sich dabei um einen Webbrowser, während auf dem Ziel-Host der eben erwähnte Webserver an der Kommunikation beteiligt ist. Das an den Router verschickte Paket enthält den Absender-Port 80. Da es sich bei der Empfängeradresse um eine externe IP-Adresse handelt, leitet der Router das IP-Paket an die NAT-Einheit weiter. Die NAT-Einheit ersetzt daraufhin die lokale Absenderadresse durch die öffentliche IP-Adresse des Unternehmens; nur so wird sichergestellt, dass die Antwort auf dieses Paket auch wieder beim Unternehmensnetzwerk ankommt. Um bei einer Antwort, die die öffentliche IP-Adresse des Unternehmens enthält, die private IP-Adresse des eigentlichen Empfängers ermitteln zu können, benutzt die NAT-Einheit eine NAT-Tabelle. In dieser Tabelle sind für jede bestehende TCP-Verbindung unter einem Index die tatsächliche IP-Adresse sowie der zugehörige Port des Empfängers gespeichert. Besteht für eine Verbindung, d.h. eine Kombination aus lokaler IP-Adresse und Port, kein Eintrag in der Tabelle, so wird ein neuer Eintrag erzeugt. Anschließend wird der Sender-Port des TCP-Pakets durch den Indexwert der Verbindung in der NAT-Tabelle ersetzt. Das Paket, welches in Abbildung 5 unten rechts dargestellt ist, wird anschließend an den externen Router weitergeleitet.
Erhält die NAT-Einheit eine Antwort vom Webserver des Ziel-Hosts (2.Paket von oben), so ermittelt es zunächst den Wert des Empfänger-Ports des TCP-Paketes. Bei diesem Wert handelt es sich nicht um den Port des Empfängers, sondern um den Index, unter dem die tatsächliche IP-Adresse des Empfängers mit der richtigen Port-Nummer in der NAT-Tabelle gespeichert ist.
Nachdem die Empfängeradresse und der Port durch die NAT-Einheit ersetzt wurden, wird das Paket (das oberste in Abbildung 5) an den Router weitergeleitet. Der Router kann das Paket aufgrund der ersetzten Empfängeradresse an das richtige Gerät im Firmennetz weiterleiten. Als problematisch erweist sich das NAT-Verfahren oft im Zusammenhang mit Sicherheitsmechanismen und Verschlüsselungsverfahren. Das Umschreiben der IP-Adressen kann dazu führen, dass die IP-Pakete nicht mehr akzeptiert werden bzw. nicht mehr korrekt zuzuordnen sind.
Da die NAT zur Überwindung des Mangels an öffentlichen IP-Adressen eingeführt wurde, ist davon auszugehen, dass dieses Verfahren mit der Einführung des neuen IP-Protokolls IPv6 an Bedeutung verlieren wird. Wie bereits in Abschnitt 4 erwähnt, bietet IPv6 einen riesigen Adressraum, welcher Spartechniken für IP-Adressen überflüssig macht.

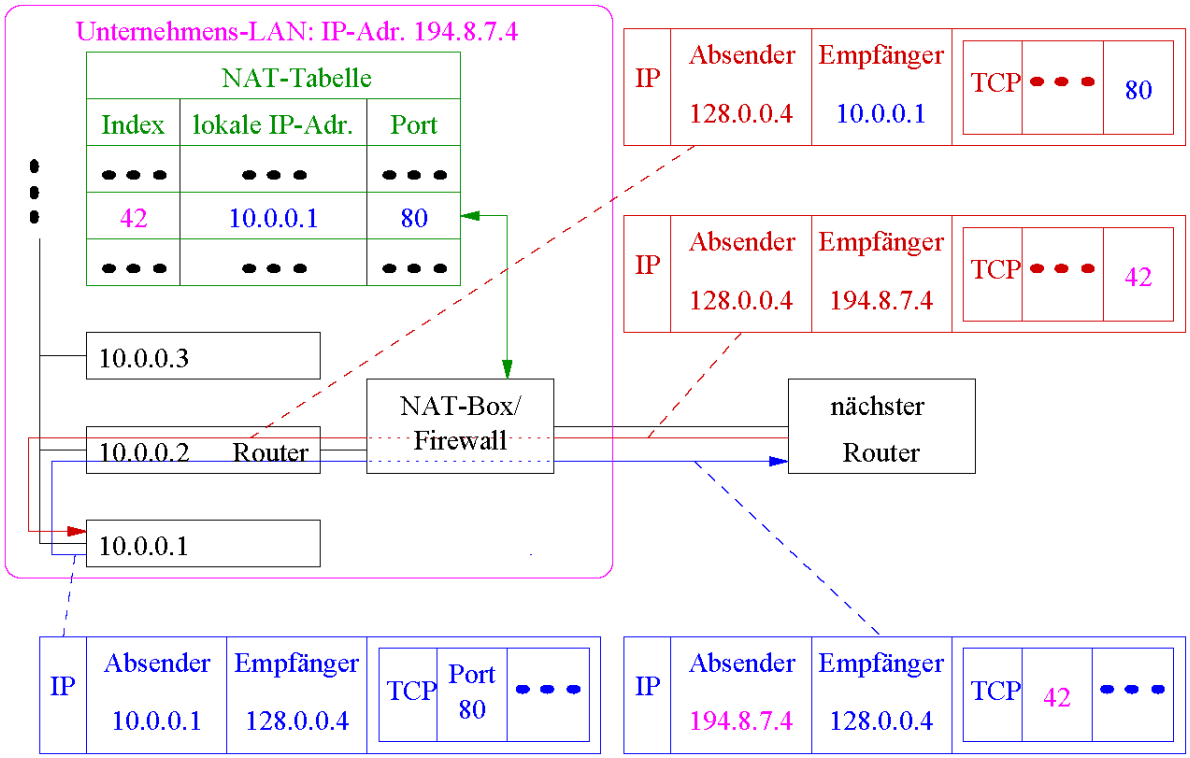
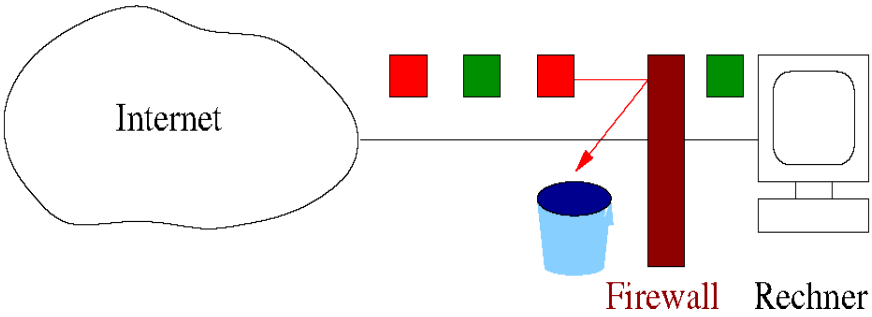


Abbildung 5: Funktionsweise der Network Address Translation

6. Firewalls

Eine Firewall ist ein System aus Hardware- und Softwarekomponenten, das dazu dient, den Datenverkehr zwischen unterschiedlichen Netzwerken zu überwachen und zu kontrollieren. Durch die Kontrolle des Datenverkehrs soll in der Regel ein sicheres Netzwerk, wie ein Firmen- oder Heimnetzwerk, vor Zugriffen aus einem unsicheren Netzwerk, wie dem Internet, gesichert werden. Eine Firewall funktioniert dabei in zwei Richtungen: Sie soll sowohl den ungewollten Verkehr von einem externen Netzwerk in ein geschütztes Netzwerk als auch den ungewollten Verkehr aus einem geschützten Netzwerk in ein externes Netzwerk unterbinden.
Ebenso wie die Kommunikation in Netzwerken auf verschiedene Schichten aufgeteilt ist (s. Abschnitt 2), lassen sich auch die Mechanismen einer Firewall zur Kontrolle der Kommunikation verschiedenen Schichten zuordnen. Schutzmechanismen (wie zum Beispiel ein Virenscanner), die den anwendungsorientierten Schichten des OSI-Modells zugeordnet sind, nehmen eine inhaltliche Klassifikation des Datenverkehrs vor. Bei den Mechanismen, welche den transportorientierten Schichten zuzuordnen sind, findet keine inhaltliche Untersuchung des Datenverkehrs statt. Hier werden lediglich die Transportinformationen über Sender und Empfänger dazu benutzt, den Datenverkehr zu kontrollieren. Die Filterregeln, der Transportebene sind in der Regel wesentlich einfacher aufgebaut, als die Filterregeln auf Anwendungsebene.



6: Firewall

Abbildung 6 zeigt die Funktionsweise einer Firewall auf Transportebene. In diesem Szenario dient die Firewall dazu, unerwünschten Datenverkehr zwischen einem einzelnen Rechner und dem Internet zu unterbinden. Ziel ist der Schutz des Rechners und seiner Daten vor Zugriffen aus dem Internet. Um den Datenverkehr zwischen Internet und Rechner in erlaubten und unerlaubten Datenverkehr unterteilen zu können, verwendet die Firewall ein Regelwerk, mit dem sich die IP-Pakete kategorisieren und somit filtern lassen.
Ein Ausschnitt des Regelwerkes der Firewall ist in der Tabelle 7 abgebildet, welche unterschiedliche Kriterien für die Klassifikation von IP-Paketen zeigt. Diese Kriterien können die Richtung (eingehend oder ausgehend), die Sender- oder Empfänger-Adressen und deren Ports betreffen. Werden diese Parameter von einem IP-Paket nicht erfüllt, so wird dieses verworfen und nicht durch die Firewall weitergeleitet.

Regel	Richtung	Quell-IP	Ziel-IP	Quellport	Zielport	Aktion
A	ein	extern	intern	1023	25	zulassen
B	aus	intern	extern	25	1023	zulassen
C	aus	intern	extern	1023	25	zulassen
D	ein	extern	intern	25	1023	zulassen
E	beliebig	beliebig	beliebig	Beliebig	beliebig	beliebig

Tabelle 7: Regelsatz für eine Firewall

III. Filter

Im Jahre 1994 tauchten erstmals automatisiert verschickte Spam-Mails auf, und es wurden erste neue Technologien entwickelt, mit dem erklärten Ziel, der immer schneller anwachsenden Welle von Spam-Mails entgegenzuwirken, um Unternehmungen große Mengen an Bandbreite, Server Ressourcen und Zeit für das Beschwerdemanagement einzusparen. Die ersten Ansätze waren noch vergleichsweise primitiv, doch mit der Zeit entstanden viele nützliche Technologien, die zum Teil noch heute im Kampf gegen den Spam eingesetzt werden. Dieses Kapitel gibt einen Überblick über einige der populärsten Ansätze sowie deren Stärken und Schwächen ([1,5])

1. Regeln

Von 1994 bis 1997 gab es kaum eine brauchbare Technologie, um effektiv gegen Spam vorzugehen. Die erste Waffe gegen Spam waren Tools, die mit einer sehr simplen Sprachanalyse versuchten, Spam-Mails von seriösen Emails zu unterscheiden. Bei diesem Ansatz wurde der komplette Text einer Email gescannt und nach Spam-typischen Wörtern und Textphrasen durchsucht, wie z. B. `Viagra`, `Call Now!` oder `Free Trial!`. Diese Wörter und Phrasen wurden in

einer vom Anwender gepflegten Wortliste verzeichnet. Der Nutzer konnte über eigene Regeln festlegen, welche Aktion bei einer erkannten Übereinstimmung zwischen Wortliste und Emailtext ausgeführt werden sollte. Beispielsweise konnte auf diese Weise eine als Spam klassifizierte E-mail auf Wunsch unmittelbar gelöscht oder in einen Spam-Ordner verschoben werden.
Das Filtern von Spam auf Basis von einzelnen Wörtern hatte anfangs eine Erfolgsrate von ca. 80 Prozent, und die Wahrscheinlichkeit, dass hiermit auch seriöse Emails aussortiert wurden, war sehr gering. Mit der Zeit wurden Spam-Mails jedoch immer häufiger verschickt, und die Pflege der Wortlisten war sehr aufwändig und zeitintensiv. Aus diesem Grund konnten sich erste kommerzielle Tools auf dem Markt durchsetzen, die als zusätzlichen Service eine täglich aktualisierte Wortliste zum automatischen Download beinhalteten.
Einer der Vorteile von regelbasierten Filtern ist, dass sie einfach zu implementieren sind und vor allem einfach an die Bedürfnisse der Anwender angepasst werden können. Dies ist jedoch auch als großer Schwachpunkt zu sehen, da die Wartung und Pflege des Regelsatzes und der Wortliste sehr viel Zeit in Anspruch nimmt. Ein weiterer Nachteil von regelbasierten Filtern ist, dass obwohl viele der in der Wortliste hinterlegten Wörter und Phrasen in hohem Maße Spam-spezifisch sind, dennoch einige seriöse Mails fälschlicherweise als Spam erkannt und durch die hinterlegten Regeln entsprechend verarbeitet werden. Wenn der Nutzer sicherstellen will, dass keine seriösen Emails verloren gehen, muss er seinen Spam-Ordner in regelmäßigen Abständen nach wie vor manuell prüfen.

2. Schwarze Listen

Paul Vixie entwickelte 1997 die erste frei verfügbare Index-basierte Realtime Blackhole List (RBL) und machte damit den ersten wichtigen Versuch, Spam auf der Ebene des Internet Service Providers (ISP) zu begegnen. Die Blackhole List war ein System zum Vortäuschen von Netzwerkausfällen (blackholes) zum Zweck der Limitierung des Transports von Spam-Mails. Eine Blackhole List führte dazu eine Liste von bekannten Netzwerken, aus denen Spam-Mails verschickt wurden (die sog. „schwarze Liste"), und bot den Nutzern die Möglichkeit, die Kommunikation mit diesen Netzwerken zu unterbinden. Solche schwarzen Listen mussten im Allgemeinen manuell erstellt und verwaltet werden. Wenn ausreichend viele Beschwerden über ein bestimmtes Netzwerk vorlagen, wurde das Netzwerk auf die schwarze Liste gesetzt. Ab diesem Zeitpunkt wurden alle Nachrichten, die aus einem der auf der schwarzen Liste stehenden Netzwerke verschickt wurden, von den Abonnenten der Blackhole List ignoriert.
Vixie nannte sein System MAPS (Mail-Abuse Prevention System) RBL. Die Nutzer konnten über eine Internetverbindung auf das MAPS RBL zugreifen und ihre Router automatisch mit den Blackhole Lists konfigurieren, so dass Nachrichten, die aus den indizierten Netzwerken verschickt wurden, nicht mehr weitergeleitet wurden. Spam-Mails wurden auf diese Weise von den Routern wie von einem schwarzen Loch „verschluckt". Nach außen sah es so aus, als habe der Spammer einen totalen Netzwerkausfall. Kleinere Unternehmen ohne eigene Router konnten ihre Mail Server so konfigurieren, dass die Internetadressen der Absender von eingehenden Emails in den schwarzen Listen nachgeschlagen wurden. Bei einem Treffer wurde die Email abgewiesen. Von 1997 bis 1999 war MAPS RBL das Werkzeug der Wahl, um Spam-Mails zu bekämpfen. Im Jahre 2001 griff eine Anti-Spam-Organisation SPEWS diese Idee auf und führte ihre eigene Blackhole List ein. Im Laufe der Zeit folgten weitere Organisationen. Jede dieser Blackhole Lists wurde unterschiedlich restriktiv erstellt, so dass Internet Service Provider nun die freie Wahl hatten, wie streng ihre Spam-Filter-Politik sein sollte.
Blackhole Lists erlauben dem Anwender, Spam nach ihrem Ursprung und nicht nach ihrem Inhalt zu filtern. Ein wesentlicher Vorteil für den Nutzer ist darin zu sehen, dass er nicht mehr seine

eigenen inhaltsbasierten Filterregeln erstellen muss, sondern sich bequem auf zentral hinterlegte Informationen stützen kann. Zudem ermöglichen Blackhole Lists Bandbreite und Server-Ressourcen einzusparen. Jedoch zeigen sich zwei signifikante Schwächen dieses Systems. Ein wesentlicher Kritikpunkt ist, dass Spam versendende Netzwerke erst dann in die schwarze Liste eingetragen werden, nachdem sie bereits eine Zeit lang ungehindert Spam verschickt haben. Insbesondere heutzutage ist das ein Problem, da es aufgrund gestohlener Einwahl-Accounts oder fehlerhaft konfigurierter Mail Server, mit denen Emails an beliebige Empfänger außerhalb des eigenen Netzwerks gesendet werden können, einfach möglich ist, kurzfristig Spam von einem anderen Host aus zu versenden, wenn der alte auf die schwarze Liste gesetzt wird. Dies führt zu dem zweiten Problem: die Qualität der Wartung von Blackhole Lists. Um dem Problem mit neuen Spam-Hosts entgegenzuwirken, wurde in vielen Listen der Adressraum von Einwahl- oder ISDN-Nutzern vermerkt. Das große Problem ist, dass sich dieser Adressraum ständig ändert. Unternehmen ändern ihre Internet-Adresse oder geben ihr Geschäft auf (insbesondere zwischen 2000 und 2002) und Privatleute wechseln ihren Internet-Provider. Die Folge sind veraltete, kaum aktuelle schwarzen Listen, aufgrund deren viele seriöse Emails blockiert werden.

3. Weiße Listen

Weiße Listen (Whitelists) sind das Gegenstück zu schwarzen Listen. Bei diesem Ansatz werden grundsätzlich alle Emails abgelehnt, deren Absender nicht explizit in der Whitelist aufgeführt ist. Alle Kontakte in der Whitelist werden als vertrauenswürdig eingestuft. Zwar bietet dieser Ansatz einen sehr effektiven Schutz vor Spam, unabhängig vom Inhalt von Emails, da aber alle Nachrichten mit unbekanntem Absender vom Filter abgelehnt werden, werden auch Nachrichten von Personen blockiert, die den Anwender aus seriösen Gründen kontaktieren wollen, ohne dass der Empfänger etwas davon mitbekommt. Des Weiteren nimmt die Pflege dieser Listen sehr viel Zeit in Anspruch. Aus diesen Gründen erscheint es einsichtig, dass sich der Einsatz von Whitelists in der Regel nur für Privatpersonen eignet, die einen festen Bekanntenkreis haben, mit dem sie kommunizieren. Für Unternehmen ist dieser Ansatz unbrauchbar.
Viele Whitelisting-Systeme basieren nur auf Email-Adressen, damit insbesondere unerfahrene Nutzer ihre Kontakte einfach und bequem der Whitelist hinzufügen können. Jedoch sind Email-Adressen leicht zu fälschen. Beispielsweise werden viele Anwender von Antiviren-Software die entsprechenden Support-Adressen wie `support@mcafee.com` in ihre Whitelist eintragen. Da beim Empfang einer Email keine Authentifizierung des Absenders verlangt wird, können Spammer die Email-Adresse des Absenders fälschen und auf diese Weise ungehindert Spam an den Nutzer senden. Die betroffenen Email-Adressen können aber nicht aus der Whitelist entfernt werden, ohne damit gleichzeitig auch alle grundsätzlich erwünschten Nachrichten des Originalabsenders zu blockieren.

4. Statistische Filter

Statistische Filter nutzen Techniken der künstlichen Intelligenz, um Emails in Spam oder Nicht-Spam zu klassifizieren. Hierbei vergleichen die selbstlernenden Filter neue Emails mit bereits erlernten Fakten, gewichten gefundene Merkmale mit einer bestimmten Punktzahl und stufen eine Email als Spam ein, wenn die Gesamtpunktzahl einen gewissen Schwellenwert übertritt. Indikatoren für Spam können eine übermäßige Verwendung von Sonderzeichen und Großbuchstaben, versteckte HTML-Texte, Unsubscribe-Zeilen (angebliche Möglichkeit, sich von der Abonnentenliste abzumelden) und eine große Häufigkeit von bestimmten Schlagworten sein. Durch den Lernprozess passen sich die Filter mit der Zeit dem typischen Email-Verhalten des Anwenders an. Wenn der Nutzer beispielsweise in einer Bank arbeitet, werden Emails, die wie-

derholt das Wort Kredit enthalten, nicht als Spam eingestuft. Die Qualität des Filters hängt dabei stark von der Qualität des Trainings ab. Wenn ein Filter nicht gut trainiert ist, kann er einen hohen Anteil an fehlerhaft klassifizierten Emails erzeugen.

Token	Spam	Nicht-Spam	Wahrscheinlichkeit
Call	90	10	0.9
free	108	12	0.9
hormones	94	6	0.96
It’s	60	60	0.5
now	96	24	0.8
Viagra	92	8	0.92

Tabelle 8: Wortliste

Token	Spam-Wahrscheinlichkeit
Call	0.9
free	0.9
It’s	0.5
now	0.8

Tabelle 9: Token und Token-Werte

Viele Spamfilter analysieren den Text einer Email mit statistischen Methoden und berechnen auf Basis von einzelnen Wahrscheinlichkeiten, mit denen bestimmte Wörter in Spam- oder Nicht-Spam-Nachrichten auftreten, eine Gesamtwahrscheinlichkeit, mit der die untersuchte Email eine Spam- oder Nicht-Spam-Nachricht ist. Solche Filter basieren auf einer Wortliste, die zu jedem eingetragenen Wort die Häufigkeit angibt, mit der das Wort in Spam- und Nicht-Spam-Nachrichten auftritt. Die Wortliste wird durch Training erstellt und aktualisiert.
Eine bekannte Implementierung eines statistischen Filters ist das Bayesian Content Filtering (BCF). Es beinhaltet Konzepte, die bereits Mitte des 18. Jahrhunderts vom britischen Mathematiker Thomas Bayes in seinem Buch `The Doctrine of Chances` beschrieben und erstmals im Jahre 2002 von Paul Graham [2] zur Erkennung von Spam genutzt wurden.
Als erstes wird der Text und der Header einer Email von einem sogenannten Tokenizer in sehr kleine Komponenten aufgebrochen (Token) und jedem dieser Token ein bestimmter Wert zugeordnet. Der Wert eines Token ist die numerische Repräsentation seiner Zugehörigkeit zu Spam oder Nicht-Spam, basierend auf seinem Vorkommen in der Trainingsbasis. Wenn zum Beispiel ein Wort primär in Spam-Mails vorkommt, wird diesem ein Wert zugeordnet, der eine hohe Spam-Zugehörigkeit widerspiegelt (z.B. 99%). Basierend auf den verdächtigsten und unverdächtigsten Token bezüglich der Wortliste wird die Nachricht anschließend klassifiziert.

Betrachten wir als Beispiel folgende Nachricht: `Call now, it's free!`, die vom Tokenizer in die Komponenten `Call`, `now`, `it's` und `free` zerlegt wird. Tabelle 8 zeigt die Wortliste, in der die extrahierten Token gesucht werden. Die Werte in den Spalten Spam und Nicht-Spam sind Häufigkeiten, die angeben, wie oft ein bestimmtes Token in den zum Training genutzten Spam- bzw. Nicht-Spam-Nachrichten vorgekommen ist. Aus diesen Werten leiten sich die entsprechenden Wahrscheinlichkeiten ab. Tabelle 9 zeigt die aus der Nachricht extrahierten Token

und die nachgeschlagenen Spam-Wahrscheinlichkeiten. Diese Einzelwahrscheinlichkeiten werden nun zu einer Gesamtwahrscheinlichkeit verdichtet (z.B. nach Bayes):

$$\rho_{spam} = \frac{0.9 \cdot 0.9 \cdot 0.5 \cdot 0.8}{0.9 \cdot 0.9 \cdot 0.5 \cdot 0.8 + (1-0.9) \cdot (1-0.9) \cdot (1-0.5) \cdot (1-0.8)} = 0.9969$$

Beim ersten Einsatz des Spam-Filters ist die Wortliste in der Regel leer und muss erst durch Training erstellt werden. Zwar sind manche kommerziellen Filter bereits vortrainiert, jedoch muss die Anpassung an das persönliche Email-Verhalten in jedem Fall erfolgen. Dazu werden ungelesene oder bereits gelesene Emails vom Anwender manuell als Spam oder Nicht-Spam klassifiziert. Im Rahmen des Lernprozesses werden anschließend vom Tokenizer automatisch Token aus den Emails extrahiert, der Wortliste hinzugefügt und die entsprechenden Zeileneinträge aktualisiert. Die Einträge in der Wortliste können zu jedem Zeitpunkt durch weiteres Training aktualisiert und erweitert werden. Wenn zum Beispiel in einer Email das Wort `V1agra` anstatt `Viagra` benutzt und die Email fälschlicherweise als Nicht-Spam eingestuft wird, kann durch ein erneutes Training `V1agra` der Wortliste hinzugefügt werden. `V1agra` ist nun ein deutlich signifikanterer Indikator für Spam als `Viagra`, da dieses Token ausschließlich in Spam-Mails vorkommt.
Statistisches Filtern ist derzeit die beste bekannte Methode, um gegen Spam vorzugehen. Aktuelle und gut trainierte Filter erreichen eine Genauigkeit von über 99,9% [4]. Als Nachteil ist jedoch auch hier zu sehen, dass keine 100%-ige Genauigkeit erreicht wird. Es kann nicht ausgeschlossen werden, dass seriöse Nachrichten fälschlicherweise als Spam oder Spam-Mails fälschlicherweise als Nicht-Spam-Nachrichten klassifiziert werden.

IV. Angriffe auf statistische Filter

Die meisten Angriffe auf statistische Filter beziehen sich auf den Tokenizer, also auf den Teil des Filters, der für das Aufbrechen des Nachrichteninhalts in kleine Komponenten (Token) verantwortlich ist. Der gebräuchlichste Ansatz ist hierbei, die Nachricht zu verschleiern, um auf diese Weise die heuristischen Funktionen des Tokenizers so zu verwirren, dass die Nachricht fehlinterpretiert und als Nicht-Spam klassifiziert wird. Ein Angriff wurde bereits diskutiert. Damit Emails mit dem Wort `Viagra` nicht sofort vom Filter aussortiert werden, schaffen Spammer stattdessen Wortvariationen wie beispielsweise `V!agra, V1agra` oder `Viagr@`, von denen sie ausgehen, dass diese noch nicht in der Wortliste des Filters vermerkt sind. Bis zu dem Zeitpunkt, an dem der Filter mit den entsprechenden Emails und den darin enthaltenen Wortvariationen trainiert wird, gelten diese Wörter als unverdächtig und nehmen keinen Einfluss auf die Klassifikation der Email.
Eine weitere Variante der Wortverschleierung ist das Einfügen von Trennzeichen, wie im Beispiel von `V-i-a-g-r-a`. Allerdings ist dieser Angriff leicht durch Elimination der Trennzeichen zu verhindern.
Die Wortliste eines statistischen Filters speichert Informationen darüber, wie oft bestimmte Token in den zum Training analysierten Emails vorgekommen sind. Aus dieser Information wird die Wahrscheinlichkeit dafür abgeleitet, dass ein betrachtetes Wort in einer Spam-Mail verwendet wird (vgl. Tabelle 8). Ein Spammer kann sich diese Eigenschaft zu Nutze machen und seine eigentliche Spam-Nachricht um einen bunten Mix aus Wörtern erweitern, die durch den Filter als unbedenklich eingestuften wurden, in der Hoffnung, dass der Filter die betrachtete Email daraufhin als Nicht-Spam klassifiziert. In der Regel ist ein solcher „Wortsalat" jedoch nicht signifikant

und wird vom Spam-Filter ignoriert. Folgendes Beispiel zeigt eine Spam-Mail, die beide vorgestellten Techniken kombiniert:

```
Hi,

VbAGRA for less:
http://www.ironequestrian.info
----------------------
and sing, my work would be limited to throwing the switch before
each himself out on his blower-powered bagpipe. Steengo plucked
at a tiny cartridge. He looked up as I passed, stared me straight
in the face. I…
```

Immer häufiger wird die eigentliche Spam-Nachricht nicht als reiner Text verschickt, sondern in einem Bild (z.B. im JPEG-Format) versteckt, das statt des Werbetextes angezeigt wird. Auf diese Weise analysiert der Tokenizer nur den aus unbedenklich eingestuften Wörtern bestehenden Wortsalat. Jedoch bieten Email-Header und HTML-Tags weiterhin eine gute Filtergrundlage, so dass auch dieser Angriff durch aktuelle Filter vielfach abgewehrt werden kann.
Manche Spam-Mails verstecken die Nachricht auch in einem aus Text- und Sonderzeichen „gemalten" ASCII-Bild. Die Idee hierbei ist, dem Tokenizer ausschließlich unbrauchbares Datenmaterial zu liefern, so dass eine Zerlegung des Nachrichtentextes in Token nicht mehr möglich ist. Auf diese Weise wird dem Filter die Basis seiner Analysemethoden genommen, und die Nachricht kann den Filter ungehindert passieren. Allerdings ist das Ergebnis vergleichsweise unattraktiv und kommerziell wenig erfolgreich, wie folgendes Beispiel verdeutlicht:

```
X   X  X     X         XX     XXX       X
 X  X   X    X X      X       X   X     X  X
  X X   X    XxxX     X GXX   XXY      XxxX
  X Y   X   X   X      X   X   X X      X    X
   X    X   X    X       XY    X  X     X    X


http://www.ironequestrian.inf
```

V. Zusammenfassung

In der vorliegenden Arbeit wurden die Grundlagen der Kommunikation im Internet beschrieben und die Arbeitsweise vom Firewalls und unterschiedlichen Spam-Filtern beschrieben. Statistisches Filtern ist derzeit die beste bekannte Methode, um gegen Spam vorzugehen. Aktuelle und gut trainierte Filter erreichen eine Genauigkeit von über 99,9%. Aktuelle Filter kombinieren jedoch in der Regel mehrere Filtertechniken, um einen noch effektiveren Schutz gegen Spam bieten zu können. Die Techniken wurden hier am Beispiel der Spam-Abwehr erläutert, sind aber genau so bei der Erkennung von Webseiten mit illegalen Inhalten, auch innerhalb von E-Shops, verwendbar.

Literatur

[1] *J.Aycock, N.Friess*, Spam Zombies from Outer Space, TR2006-808-01, University of Calgary, 2006.

[2] *P.Graham*, A Plan for Spam, http://www.paulgraham.com/spam.html, 2002.

[3] *A.S.Tanenbaum,*Computernetzwerke, Prentice Hall, 3.Aufl., 2000.

[4] *W.S.Yerazunis*, The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past, MIT Spam Conference, Cambridge, Massachusetts, 2004.

[5] *J.A.Zdziarski*, Ending Spam, No Starch Press, 1st Ed., 2005.

[6] *E.D.Zwicky, S.Cooper, D.B.Chapman*, Building Internet Firewalls, O`Reilly, 2nd Ed., 2000.