MECANISMO ANTI-*SPAM* BASEADO EM AUTENTICAÇÃO E REPUTAÇÃO

Danilo Michalczuk Taveira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. Otto Carlos Muniz Bandeira Duarte, Dr. Ing.

Prof. Virgílio Augusto Fernandes de Almeida , Ph.D.

Prof. Luís Henrique Maciel Kosmalski Costa, Dr.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2008

TAVEIRA, DANILO MICHALCZUK

Mecanismo Anti-*Spam* Baseado em Autenticação e Reputação[Rio de Janeiro] 2008

XI, 86 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Mensagens não Solicitadas

2. *Spam*

3. Reputação

I. COPPE/UFRJ II. Título (série)

À minha família.

Agradecimentos

Um agradecimento especial aos meus pais Vera Lucia Michalczuk Taveira e Antônio Florindo Taveira e aos demais familiares pelo carinho, atenção, compreensão, confiança e incentivo ao longo de toda a minha vida.

Ao professor Otto, meu orientador e responsável por uma importante parte da minha formação acadêmica e profissional, por sua amizade, conselhos e orientação. Aos professores do GTA, Luís Henrique, Rezende, Aloysio e Rubi, pela amizade, ensinamentos e conselhos.

A todos do GTA, em especial aos amigos Reinaldo, Carina, Natalia, Marcelo Duffles, Igor, Miguel, Kleber, Carlos Henrique, Sávio, Natanael, André, Raphael, Marcel, Vinícius e ao pessoal da Iniciação Científica, pelas risadas e pelo incentivo durante o trabalho.

Agradeço em particular aos professores Virgílio Augusto Fernandes de Almeida e Luís Henrique Costa pela participação na banca examinadora.

Aos funcionários do Programa de Engenharia Elétrica da COPPE/UFRJ pela presteza no atendimento na secretaria do Programa. A todos que me incentivaram, contribuindo de forma direta ou indireta, para a minha formação acadêmica e profissional. Ao CNPq, à Capes, ao UOL, à FINEP, à RNP e ao FUNTTEL pelo financiamento da pesquisa.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MECANISMO ANTI-*SPAM* BASEADO EM AUTENTICAÇÃO E REPUTAÇÃO

Danilo Michalczuk Taveira

Março/2008

Orientador: Otto Carlos Muniz Bandeira Duarte

Programa: Engenharia Elétrica

As mensagens eletrônicas não solicitadas, também conhecidas como *spams*, causam grandes prejuízos para usuários e provedores de serviço. Para filtrar os *spams* surgiram os sistemas anti-*spams*. No entanto, esses sistemas de combate aos *spams* acabam também filtrando mensagens legítimas, causando falsos positivos. Os falsos positivos de um sistema anti-*spam* têm um impacto muito grande para os usuários, já que uma mensagem legítima é identificada como *spam* e filtrada, podendo causar grandes prejuízos financeiros e atrasos no processo de comunicação. Neste trabalho é proposto um mecanismo anti-*spam* que tem como foco principal a redução dos falsos positivos. A idéia chave é decidir se a mensagem é legítima ou não a partir do histórico de comportamento dos usuários, utilizando um mecanismo de autenticação e reputação. Usuários que possuem um histórico de enviar apenas mensagens legítimas têm uma boa reputação e uma probabilidade menor de terem suas mensagens filtradas incorretamente. São desenvolvidos um modelo analítico do modelo proposto e um simulador que implementa o mecanismo proposto e outro mecanismo da literatura para comparação. Análises de desempenho são realizadas variando os parâmetros do simulador e analisando vários cenários diferentes. Em comparação com outro mecanismo similar, o mecanismo proposto sempre possui uma taxa de falsos positivos menor em todos os cenários analisados.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ANTI-*SPAM* MECHANISM BASED ON AUTHENTICATION AND REPUTATION

Danilo Michalczuk Taveira

March/2008

Advisor: Otto Carlos Muniz Bandeira Duarte

Department: Electrical Engineering

Unsolicited messages, also known as spams, cause huge losses to users and service providers. Anti-spam systems aim to filter the unsolicited messages. These systems, however, may also incorrectly filter legitimate messages, causing false positives. False positives of an anti-spam mechanism can have a great impact for users when a legitimate message is classified as spam, causing financial losses and delays in communication. This work proposes a mechanism that aims to reduce false positive rates. The key idea is to take into account the history of the user behavior on the message filtering process, using an authentication and reputation mechanism. A user that has a history of sending only legitimate messages has a high reputation, reducing the likelihood of false positives. Analytical models of the proposed mechanism and another mechanism proposed in the literature are derived and a simulator, which implements both mechanisms, is also developed. Performance evaluation of the both mechanism in different scenarios are shown, varying different parameters. The proposed mechanism always improve the false positive rate on all the scenarios.

Sumário

Lista de Figuras

Lista de Acrônimos

ADES :	Análise DE Spam;
BGP :	*Border Gateway Protocol*;
CDF :	*Cumulative Distribution Function*;
DNS :	*Domain Name System*;
DNSBL :	*Domain Name System Blacklist*;
FP :	Falso Positivo;
FN :	Falso Negativo;
HTML :	*HyperText Markup Language*;
IP :	*Internet Protocol*;
MX :	*Mail Exchange*;
ORBS :	*Open Relay Behavior-modification System* ;
POP3 :	*Post Office Protocol version 3*;
PTR :	*Pointer*;
RBL :	*Real-time Blackhole List*;
RFC :	*Request For Comments*;
SMTP :	*Simple Mail Transfer Protocol* ;
SPF :	*Sender Policy Framework*;
TCP :	*Transmission Control Protocol*;
TOPAS :	*Trust Overlay Protocol for Anti Spam*;
VoIP:	*Voice over IP*.

Capítulo 1

Introdução

ATUALMENTE , o combate aos *spams*, ou mensagens eletrônicas não solicitadas, é um dos grandes desafios na Internet. O *spam*, de forma simplificada, é toda mensagem eletrônica não desejada pelo destinatário. Devido à simplicidade do protocolo SMTP (*Simple Mail Transfer Protocol*), o correio eletrônico é a aplicação mais afetada pelos *spams*. As estatísticas mostram que os *spams* já correspondem a pelo menos dois terços de todo o tráfego de correio eletrônico transportado pelos provedores de serviço, causando prejuízos anuais da ordem de bilhões de dólares [1]. Algumas previsões mais pessimistas estimam que, em poucos anos, as mensagens não solicitadas serão responsáveis por 95% do tráfego de correio eletrônico na Internet [2]. No Brasil esse problema também é bastante grave.

1.1 Motivação

Além de causarem enormes prejuízos aos provedores de serviço, devido ao consumo de recursos tais como banda passante, memória e processamento, os *spams* também consomem inutilmente o tempo dos destinatários e reduzem a credibilidade dos usuários na Internet. A insatisfação entre os usuários é cada vez maior tanto pela perda de tempo na recepção e leitura das mensagens quanto pela possibilidade de disseminação de vírus e de outros programas que causam a perda de dados e o comprometimento da segurança de

seus computadores.

Os sistemas anti-*spam* são as principais contramedidas ao envio de mensagens não solicitadas. O objetivo principal desses sistemas é reduzir o número de *spams* recebidos por um usuário. Para reduzir o número de *spams* recebidos, os sistemas anti-*spam* utilizam técnicas e procedimentos que atuam de forma preventiva e coibitiva em todas as etapas do processo de envio de mensagens. Esses sistemas tentam evitar a coleta de endereços de correio eletrônico para construção das listas de destinatários, coibir o envio, caracterizar e filtrar as mensagens. No entanto, estes sistemas precisam estar em constante evolução, pois para cada novo sistema desenvolvido, surgem novas técnicas para tentar burlá-lo.

As principais propriedades de um mecanismo anti-*spam* são as suas taxas de falsos positivos e de falsos negativos. Entendem-se como falsos positivos todas as mensagens legítimas que são classificadas como *spams* e, como falsos negativos, todos os *spams* que são classificados como mensagens legítimas. Os falsos negativos têm um impacto menor, pois o usuário receberá o *spam* e poderá apagá-lo. Dessa forma, os custos com os falsos negativos estão relacionados com a perda de produtividade e a perda de foco na realização de atividades. Já a taxa de falsos positivos tem um impacto muito maior, pois uma mensagem legítima acaba sendo perdida, gerando grandes transtornos e atrasos no processo de comunicação. Sendo assim, os custos com os falsos positivos tendem a ser altos, uma vez que informações estratégicas e oportunidades podem ser perdidas, gerando graves conseqüências profissionais e pessoais.

A filtragem de mensagens de correio eletrônico pelo conteúdo é uma técnica anti-spam muito usada atualmente que se baseia em analisar e identificar determinadas características de texto usadas pelos spammers nas mensagens. Porém, a cada dia esta técnica se torna mais complexa e menos eficaz. Ao tomar conhecimento da forma como funcionam os filtros de mensagens, os spammers facilmente modificam o conteúdo das mensagens para tentar burlar os sistemas anti-spam. Além disso, essas técnicas tratam cada mensagem individualmente, sem levar em consideração o contexto no qual a mensagem foi recebida. Um usuário que envia várias mensagens legítimas tem cada uma das suas mensagens analisada de forma independente, sem levar em consideração o seu histórico. Essa característica é o principal fator para a ocorrência de falsos positivos. Outro problema do

SMTP que é utilizado para enviar mensagens de correios eletrônico é a falta de um mecanismo de autenticação para assegurar a verdadeira identidade do remetente. Isso faz com que *spammers* possam facilmente fazer uma mensagem parecer que tenha sido originada de qualquer endereço de correio eletrônico. Dessa forma, pode-se utilizar um mecanismo que leve em conta o histórico de comportamento do usuário remetente, diminuindo a probabilidade de ocorrer um falso positivo para uma mensagem enviada por um usuário que sempre envia mensagens legítimas.

1.2 Objetivo

O mecanismo anti-*spam* proposto nesse trabalho tem como principal objetivo a redução dos falsos positivos. A idéia chave é considerar o histórico de comportamento dos usuários para decidir se a mensagem é legítima ou não. Portanto, um usuário que já enviou várias mensagens legítimas passa a ter uma probabilidade muito menor de ter suas mensagens classificadas como *spam*. Em primeiro lugar, é necessário identificar de forma precisa o remetente das mensagens, já que o protocolo SMTP não possui nenhum mecanismo de autenticação dos remetentes. Para identificar os usuários é utilizado um mecanismo de autenticação, porém sem utilizar informações pessoais, por questões de privacidade. Procura-se uma forma de autenticar o remetente da mensagem ao mesmo tempo em que se permite o anonimato do remetente. Isto é obtido através de uma autenticação por pseudônimo.

Após a identificação do usuário, o seu histórico de comportamento pode ser monitorado. Para uma maior eficácia do mecanismo proposto, o histórico de comportamento de um usuário é determinado por diversos servidores de correio eletrônico, que trocam informações sobre esse histórico de comportamento. Na troca de informações é utilizado um mecanismo de reputação, uma vez que se deve levar em consideração a reputação de cada servidor para avaliar qual será o grau de confiança dessa informação.

Para avaliar o mecanismo proposto foram desenvolvidos um modelo matemático do funcionamento do mecanismo e um simulador de eventos discretos que implementa o mecanismo. O modelo matemático permite avaliar o desempenho do mecanismo em estado

estacionário, enquanto que o simulador permite avaliar o período transiente do mecanismo. Também foi desenvolvido um modelo matemático e foi implementado no simulador outro mecanismo proposto na literatura baseado em reputação, chamado TOPAS [3], para compará-lo com o mecanismo proposto. Os resultados mostram que o mecanismo proposto é eficiente na redução dos falsos positivos sem prejudicar a taxa de falsos negativos e também é robusto a várias contramedidas que podem ser adotadas pelos *spammers* para tentar burlar o mecanismo.

1.3 Organização

Este trabalho está organizado da seguinte forma. No Capítulo 2 são apresentadas características das mensagens não solicitadas, denominadas *spams*. São discutidas ainda as técnicas utilizadas por *spammers* para enviar as mensagens e as técnicas anti-*spams* convencionalmente utilizadas. Em seguida, uma ferramenta que foi desenvolvida para avaliar as atuais técnicas anti-*spams* é apresentada. Resultados de desempenho dos mecanismos analisados através da ferramenta também são apresentados. O mecanismo baseado no histórico de comportamento dos usuários proposto nesse trabalho é introduzido no Capítulo 3. No Capítulo 4 tanto o modelo analítico do mecanismo proposto quanto o de outro mecanismo existente na literatura são apresentados. Baseado no modelo analítico derivado, vários cenários são analisados. No Capítulo 5, o desempenho do mecanismo proposto é avaliado e comparado com o desempenho de outro mecanismo baseado em reputação. A avaliação é feita a partir de simulações realizadas através de um simulador que foi implementado. Por fim, no Capítulo 6, são apresentadas as conclusões sobre este trabalho e as considerações sobre trabalhos futuros.

Capítulo 2

Mensagens não Solicitadas *

A MAIOR motivação para o envio dos *spams* é o retorno financeiro que é obtido pelos *spammers*. Em primeiro lugar, o custo de divulgação é muito menor do que uma divulgação feita através de meios de comunicação convencionais como rádio, TV e meios impressos como jornais ou revistas. Outra vantagem dos *spams* é que podem atingir um número muito grande de usuários de todas as partes do mundo, diferentemente dos outros meios de comunicação convencionais. Nesse cenário, os *spammers* usam os *spams* como uma forma altamente lucrativa de anunciar produtos ou serviços. Estima-se que um *spammer* gaste cerca de US$250 para enviar um milhão de *spams* [1]. Em uma pesquisa com usuários a respeito das suas atitudes em relação aos *spams*, 39% dos usuários domésticos e 13% dos usuários corporativos disseram que já compraram produtos ou serviços oferecidos através de *spams* [5]. Outra motivação para o envio de *spams* é o enriquecimento ilícito, através de golpes ou tentativas de estelionato. É evidente que o *spam* possui um "modelo de negócio" altamente eficiente e lucrativo.

A vantagem para os *spammers* da divulgação com um custo reduzido causa uma série de prejuízos, tanto para usuários como para os provedores de serviço e operadoras de telecomunicações [6]. Para os usuários o prejuízo maior está na perda de tempo com a leitura dos *spams*. No meio corporativo, estima-se que cada empresa tenha um prejuízo anual da ordem de US$1300 por cada empregado, devido a perda de produtividade, gastos com tec-

*Este capítulo é baseado no minicurso "Técnicas de Defesa Contra Spam" [4] apresentado no VI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG'06).

nologia de combate aos *spams* e gastos com infra-estrutura. Para os provedores de serviço e operadoras de telecomunicações os gastos são principalmente devido ao desperdício de recursos da rede com a transmissão dos *spams*, além dos gastos com o processamento e armazenamento das mensagens, soluções anti-*spam*, treinamento de pessoal, chamadas e atendimentos de suporte técnico e outros gastos indiretos. Estima-se que o prejuízo para os provedores seja da ordem de US$2800 para cada milhão de *spams* encaminhados [7].

2.1 Técnicas de Envio de *Spams*

Nesta seção são discutidas as técnicas utilizadas pelos *spammers* para enviarem os *spams*. Essas técnicas têm como principais objetivos enganar os mecanismos anti-*spam* e atingir o maior número de destinatários possível. O processo para enviar os *spams* é composto de três etapas: coleta de dados, formatação das mensagens e envio. A etapa de coleta de dados tem como objetivo criar uma lista com os destinatários dos *spams*. A etapa de formatação da mensagem se refere à criação do conteúdo das mensagens. Por fim, tem-se o envio das mensagens.

2.1.1 Coleta de Dados

Nessa fase os *spammers* procuram obter o maior número possível de endereços eletrônicos, para posteriormente servirem de destinatários dos *spams* enviados. O procedimento mais utilizado para a coleta de dados é a varredura de páginas da Internet, de grupos de discussão, de arquivos de listas e de outros meios que possuam diversos endereços eletrônicos. Para que a coleta seja eficiente e de custo muito baixo, o procedimento é totalmente automatizado através de um software chamado robô. O robô acessa uma determinada página, busca em seu conteúdo endereços eletrônicos e segue para uma nova página repetindo o procedimento. A busca é extremamente simples, pois todo endereço eletrônico deve possuir o caractere @ seguido do domínio. Para tentar dificultar a tarefa dos robôs de busca de endereços eletrônicos, é comum a substituição do símbolo @ do endereço eletrônico por palavras como `at`, `[at]` e outras variações. A idéia básica é

dificultar a tarefa de busca do robô mas, ao mesmo tempo, manter simples a identificação do endereço eletrônico por um ser humano. Para uma pessoa fica razoavelmente claro que se trata de um endereço eletrônico e que deve ser adicionado o símbolo @. Já para os mecanismos robôs essa alteração torna o processo mais complicado, já que agora ao invés de buscar apenas o símbolo @ nas páginas, existem várias combinações diferentes que podem ser usadas. Um mecanismo ainda mais eficaz para enganar os robôs é criar uma figura contendo o endereço eletrônico, o que exige do robô a execução de um processo de reconhecimento de caracteres em imagem para descobrir o endereço eletrônico. Um aspecto importante a ser observado é que, mesmo se o usuário não divulgar seu endereço eletrônico em sua página ou utilizar técnicas como as descritas anteriormente, ainda existem sítios que divulgam estes mesmos endereços eletrônicos sem nenhuma preocupação com os robôs de coleta de endereços eletrônicos.

Outro meio de coleta de dados é através da invasão de sítios ou computadores através de vírus[1], cavalos de tróia[2] ou outras pragas digitais [8]. A invasão de sítios tem a vantagem de obter informações sobre o perfil dos usuários, além do endereço eletrônico. Assim, quanto mais informações do usuário o *spammer* possuir, a propaganda poderá ser mais seletiva e direcionada. As listas construídas também podem ser vendidas para outros *spammers* e quanto mais informações estiverem disponíveis maior o valor comercial dessas listas. Novos vírus também são desenvolvidos com o objetivo de capturar toda a lista de contatos pessoais dos usuários, disponíveis nos computadores para determinados aplicativos, e enviar para os *spammers*.

Existem também ataques de dicionário onde as listas são construídas utilizando nomes próprios e palavras comuns [9]. Essa técnica tem uma grande efetividade em domínios com muitos usuários, já que existe uma grande quantidade de usuários que possuem endereços com apenas o seu primeiro nome. Outra técnica similar é a de ataque de força bruta, em que são enviadas mensagens para todos os endereços eletrônicos dentro de uma determinada faixa de endereços, como por exemplo, entre aaaaa e zzzzz. Nesse exemplo todas as pessoas que possuem um endereço eletrônico de cinco caracteres irão receber os *spams*. Apesar de essa técnica possuir uma baixa eficiência para domínios com poucos

[1] Trecho de código com o objetivo de se anexar a outros programas e propagar-se.

[2] Programas que permitem acessos não autorizados aos recursos do computador.

usuários, para domínios grandes muitos usuários irão receber os *spams*.

Uma etapa importante do processo de coleta também é a verificação dos endereços eletrônicos, já que uma lista com um número muito grande de endereços inexistentes causa uma perda de tempo para os *spammers*. Um dos meios de verificação de endereço utilizados pelos *spammers* era o uso do comando VRFY do protocolo SMTP. Entretanto, esse comando é atualmente desabilitado pela maioria dos administradores, com o objetivo de dificultar a verificação dos endereços. Atualmente a verificação dos endereços é feita incluindo iscas nas mensagens, como um link para a remoção do usuário na lista. O usuário, achando que não vai mais receber os *spams*, ao clicar nesse link acaba apenas confirmando para o *spammer* que aquele endereço existe e, além disso, o *spammer* identifica que é um usuário que lê as mensagens de *spams*. Dessa forma, o *spammer* passa a enviar ainda mais *spams* para esse usuário. Outra técnica utilizada para realizar a verificação é a inclusão na mensagem de imagens que contêm um determinado número ou texto de referência relacionado com o endereço eletrônico. Assim, quando o usuário visualizar a mensagem, o seu programa de correio eletrônico irá acessar o servidor onde a imagem está hospedada para buscá-la e enviar esse texto ou número de referência juntamente com a requisição da imagem. Verificando e associando os números ou textos de referência, o *spammer* pode então saber quais são os endereços válidos e ativos além de poder tipificar a preferência deste usuário. Para combater esse processo, os programas de correio eletrônico mais novos não abrem imagens externas por padrão, sendo necessário que o usuário confirme a abertura de imagens externas caso deseje. Entretanto, a falta de informação dos usuários pode fazer com que os usuários sempre confirmem a abertura das imagens.

2.1.2 Formato das Mensagens

No início, a construção das mensagens de *spam* era feita sem nenhuma preocupação com o seu conteúdo. *Spams* eram facilmente identificados pela análise de conteúdo ao se verificar determinadas palavras, como por exemplo: Viagra, Cialis etc. Assim, para contornar os mecanismos anti-*spam* baseados em análise de conteúdo, os *spammers* passaram a se preocupar com o formato das mensagens, de forma a evitar ou camuflar determinadas palavras.

Uma decisão importante dessa fase é a opção por criar mensagens no formato texto ou mensagens no formato HTML. As mensagens de texto geralmente são mais difíceis de serem classificadas como *spam* pelos mecanismos de análise de conteúdo, já que são geralmente bem curtas. Isso dificulta a identificação de características de que a mensagem se trata de um *spam*. Esse formato, no entanto, tem o inconveniente de ser menos atrativo para o usuário, já que não permite o uso de cores, imagens, animações e outros efeitos que chamam a atenção. O formato HTML fornece uma grande variedade de opções para formatação do conteúdo e permite o uso de figuras. Portanto, o formato HTML permite a criação de uma mensagem muito mais atrativa para o usuário que uma mensagem no formato texto. Os *spammers*, para atrair a atenção do usuário, se servem do formato HTML empregando textos com caracteres coloridos e de diferentes tamanhos além de usarem imagens. A desvantagem para os *spammers*, nesse caso, é que estas mensagens HTML geralmente possuem um tamanho maior e características adicionais que podem ser usadas para facilitar a classificação destas mensagens como *spam* [10]. Por outro lado, a versatilidade do formato HTML também é usada pelos *spammers* para tornar mais difícil a análise do conteúdo do texto por parte dos mecanismos anti-*spam*. Uma técnica muito utilizada é a disposição de palavras em células de uma tabela, onde cada uma das letras da palavra é colocada em uma célula de uma linha ou coluna da tabela. O mecanismo anti-*spam* para poder analisar o conteúdo da mensagem, tem então que interpretá-la, e descobrir que as letras que estavam separadas na verdade formam apenas uma palavra. Esse processo, no entanto, aumenta a complexidade da análise de conteúdo já que as variações tornam-se agora muito grandes. Uma técnica similar é a inserção de caracteres aleatórios no meio das palavras com uma cor igual à cor do fundo da mensagem. Para os usuários, esses caracteres não são notados. Já para os mecanismos anti-*spam* esses caracteres vão parecer que fazem parte da palavra, dificultando a análise do conteúdo. Uma evolução dessa técnica é o uso de figuras, onde a propaganda está contida dentro de uma figura. Para esse tipo de mensagem a análise por conteúdo se torna complexa, já que a análise do texto presente na imagem é difícil e custosa computacionalmente.

Atualmente, a maioria das mensagens usa também a técnica de adicionar várias palavras ou frases retiradas de textos legítimos, fazendo com que os mecanismos de análise por conteúdo analisem esse texto inserido na mensagem [11]. Como o texto aleatório não

tem nenhuma ligação com o anúncio do *spammer*, o mecanismo de análise por conteúdo acaba não classificando a mensagem como *spam*.

2.1.3 Envio das Mensagens

Depois de criada a lista de destinatários e de ter produzido o conteúdo do *spam*, resta ao *spammer* o envio da mensagem de *spam*. Para enviar as mensagens para um grande número de usuários, é necessário que haja uma grande banda disponível e que o *spammer* não possa ser rastreado. Para isso, servidores ou máquinas alheias são invadidos e utilizados, dificultando o rastreamento do verdadeiro responsável pelo envio das mensagens.

Um método muito utilizado pelos *spammers* é o abuso de servidores de mensagens ou servidores *proxy* mal configurados ou com vulnerabilidades. Os servidores de mensagens mal configurados, também chamados de servidores *open-relay*, permitem o envio de mensagens originadas de qualquer endereço. Com isso, os *spammers* podem enviar as mensagens para esses servidores e esses servidores se encarregarão de enviar as mensagens para os destinatários. O uso desses servidores mal configurados traz enorme benefício para os *spammers*, devido à dificuldade de identificar o *spammer*. Além disso, se a mensagem contiver múltiplos destinatários o *spammer* só precisa enviá-la uma vez para o servidor. O servidor irá enviar uma cópia da mensagem para cada um dos destinatários. Dessa forma, o *spammer* pode enviar uma quantidade muito grande de mensagens enviando apenas algumas mensagens para o servidor. Os servidores *proxy* também são utilizados de forma similar, só que nesse caso o *spammer* precisa enviar uma cópia da mensagem para cada um dos destinatários. Mesmo sendo menos eficiente, ainda existe a vantagem de que o *spammer* será dificilmente rastreado, pois a conexão para os servidores é feita a partir do *proxy* e não da máquina do *spammer*. A desvantagem do uso de servidores de terceiros é a rápida inclusão dos servidores em listas negras, caso muitos *spammers* utilizem um mesmo servidor.

Atualmente, grande parte dos *spams* é enviada a partir de redes de máquinas zumbis, também chamadas de *botnets* [12]. Nessas redes, o *spammer* infecta uma grande quantidade de computadores com um vírus ou cavalo de tróia e a máquina passa a ser controlada

pelo *spammer*. Os zumbis se conectam a um servidor central que distribui as ordens para os zumbis. Assim, o *spammer* pode usar as máquinas dos clientes para enviar os *spams*. A vantagem nesse caso é que, como são utilizadas muitas máquinas zumbis, cada uma envia uma quantidade reduzida de mensagens, tornando o processo de inclusão das máquinas nas listas negras mais demorado.

Uma técnica mais avançada que está começando a ser utilizada atualmente é o seqüestro de prefixos BGP (*Border Gateway Protocol*) [13]. O protocolo de roteamento BGP é o protocolo utilizado na Internet para realizar o roteamento entre os diferentes sistemas autônomos. Os roteadores BGP recebem anúncios de rota para determinados prefixos de endereços e repassam essa informação para outros roteadores. No entanto, um roteador BGP mal configurado pode acabar aceitando anúncios de prefixo originados de qualquer máquina. Explorando essa falha, os *spammers* fazem o anúncio de uma grande faixa de endereços IP para um roteador BGP mal configurado. Feito isso, eles passam então a enviar as mensagens de endereços IP dentro dessa faixa de endereços IP que foi seqüestrada. A grande vantagem nesse caso é que como uma grande faixa é seqüestrada, pode-se distribuir o envio das mensagens entre todos os endereços disponíveis, tornando mais lento o processo de inclusão dos endereços IP nas listas negras. Os anúncios geralmente também são de curta duração, para dificultar o rastreamento do *spammer*.

2.2 Técnicas de Combate aos *Spams*

O objetivo de um sistema anti-*spam* é reduzir o número de *spams* recebidos por um usuário, classificando as mensagens para, então, filtrá-las. Esses sistemas estão em constante evolução já que para cada novo sistema, tenta-se criar técnicas para enganá-lo e permitir a passagem dos *spams*. As principais propriedades de um sistema anti-*spam* são a sua taxa de falsos positivos e de falsos negativos, ou seja, a taxa de mensagens legítimas classificadas como *spams* e vice-versa. Em geral, a taxa de falsos positivos tem um valor mais importante, já que uma mensagem legítima acaba sendo filtrada, o que pode gerar grandes transtornos e atrasos no processo de comunicação. Os falsos negativos têm um impacto menor, pois o usuário irá receber o *spam*, mas provavelmente acabará apagando-

o. Outro aspecto importante de um sistema anti-*spam* é a sua interferência com o usuário, seja ela por necessidade de configuração, manutenção, atualização ou desafios que são feitos ao usuário e que devem ser respondidos. Quanto maior o nível de interação com o usuário, mais complexo e menos amigável o sistema se torna, dificultando sua adoção em grande escala. Neste capítulo, os sistemas anti-*spam* atuais e novos mecanismos propostos na literatura são classificados e analisados.

2.2.1 Listas Brancas e Negras

Os primeiros sistemas anti-*spam* a surgirem baseavam-se na utilização de listas negras, que são listas com endereços de origem ou endereços IP de remetentes que reconhecidamente são fontes de *spam*. Nas listas brancas, são colocados endereços de pessoas ou servidores confiáveis. Dessa forma, quando uma mensagem é recebida, as duas listas são consultadas. Se o endereço de origem constar da lista negra, a mensagem é diretamente classificada como *spam* e não é mais analisada por nenhum outro mecanismo anti-*spam*. Por outro lado, caso o endereço esteja na lista branca, a mensagem é aceita diretamente. Caso o endereço não esteja em nenhuma das duas listas, a mensagem pode ser aceita, ou então, são utilizados outros mecanismos anti-*spam* que estejam disponíveis, para tentar classificar a mensagem. Inicialmente, essas listas eram feitas pelos próprios usuários e cada um ficava responsável por adicionar e remover os endereços das duas listas. Isso demandava um grande esforço do usuário, pois ele tinha que separar as mensagens e determinar se o endereço deveria ser colocado na lista branca na negra ou, em caso de dúvidas, em nenhuma das duas listas. A evolução natural foi tornar o sistema centralizado, onde entidades centrais controlam a adição e remoção de endereços das listas. A primeira implementação desse tipo de sistema distribuído foi chamada de *Real-time Blackhole List* (RBL) e foi criada por Paul Vixie. Na RBL eram listados os endereços IP de servidores utilizados para enviar *spam*. Contudo, esses endereços eram adicionados manualmente. Em seguida a RBL, surgiu outra proposta chamada ORBS (*Open Relay Behavior-modification System*) cuja principal diferença para a RBL é a realização automática de testes para identificar servidores que permitem o envio de mensagens sem nenhum controle. Os servidores com essa característica são adicionados automaticamente na lista

negra, bloqueando as mensagens originadas por eles. O processo de remoção dessa lista, no entanto, é realizado de forma manual, através do contato do administrador do servidor listado com a entidade responsável pela manutenção da lista [14]. A consulta a estas listas distribuídas é geralmente feita através do protocolo DNS (*Domain Name System*), fazendo com que elas sejam chamadas genericamente de DNSBLs (*Domain Name System Black Lists*). Para realizar a consulta a estas listas, o servidor verifica se o endereço IP do cliente ou de outro servidor que se conectou a ele está presente na lista DNSBL anteriormente configurada. A verificação utilizando o protocolo DNS é realizada fazendo uma consulta DNS ao endereço formado invertendo-se os bytes do endereço IP do cliente e adicionando o nome do domínio da entidade responsável pela DNSBL. Um exemplo desse processo é mostrado na Figura 2.1, onde o servidor A quer enviar uma mensagem eletrônica para o Servidor B. Nesse caso, o Servidor B verifica se o Servidor A está presente ou não na lista negra. Essa verificação é realizada através do envio de um pedido DNS para um servidor DNS que fará a consulta à lista DNSBL. Caso a resposta do pedido de DNS seja positiva, isso significa que o endereço testado está presente na lista negra. Essas listas são também um alvo freqüente de ataques de negação de serviço por parte de *spammers* para tentar neutralizar esse mecanismo de proteção anti-*spam*.

As listas negras podem ser efetivas no bloqueio de servidores utilizados por *spammers*. Estudos mostram que até 80% dos *spams* podem ser evitados por meio do uso desses mecanismos [15]. Essas listas, no entanto, sofrem o problema de falsos positivos, quando um endereço é incorretamente adicionado à lista. O processo de retirada da lista pode demorar, causando perdas de mensagens legítimas. Máquinas contaminadas por vírus também podem ser afetadas por esse sistema. O vírus pode aproveitar-se dos recursos da máquina e instalar um servidor de correio eletrônico para ser usado por *spammers*. Os endereços destas máquinas podem acabar nas listas negras e as mensagens legítimas do usuário da máquina contaminada são recusadas.

2.2.2 Uso de Pesos e Regras

Esses sistemas utilizam regras lógicas para a classificação das mensagens como *spam* ou não. Cada uma das regras define um teste que deve ser realizado na mensagem. Caso

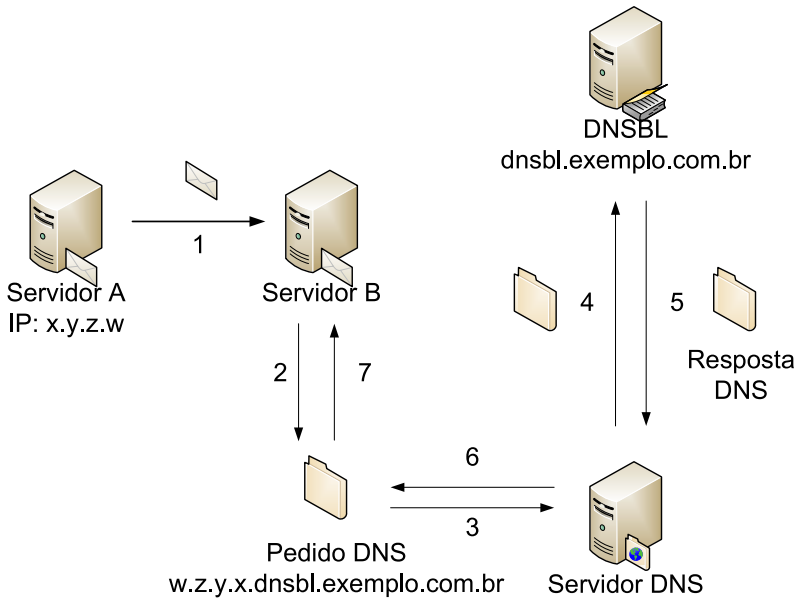


Figura 2.1: Verificação de listas negras.

o resultado de um teste seja positivo, esse resultado é usado de forma ponderada para determinar a probabilidade da mensagem ser ou não *spam*. Os pesos de cada um dos testes podem ser positivos ou negativos, indicando uma maior ou uma menor probabilidade da mensagem ser *spam*. Quando uma mensagem é recebida, o mecanismo anti-*spam* realiza todos os testes previamente definidos, somando os pesos de todos os testes cujo resultado foi positivo. Com base no valor final da soma de todos os pesos, são tomadas ações, podendo a mensagem ser encaminhada para o destinatário, marcada como um provável *spam* ou então descartada, o que geralmente acontece se o valor da soma de todos os pesos for muito alto.

Uma das principais e mais delicadas etapas desse mecanismo é a construção de regras, que ao mesmo tempo tem que balancear a generalidade, para a adaptação a novos tipos de *spam*, e a especificidade, para não classificar mensagens legítimas como *spam*. O processo de criação de regras é feito extraindo-se frases ou palavras características de mensagens de *spam*. Em seguida é feita uma análise utilizando uma grande base de mensagens manualmente classificadas como *spam* ou não, para avaliar a taxa de falsos positivos e negativos. Se alguma das taxas for alta, a regra é refinada, alterando-se os testes que são realizados, até que as duas taxas sejam baixas o suficiente. A determinação do peso da regra é baseada na percentagem das mensagens de *spam* onde a regra apre-

sentou um resultado positivo, tendo maior peso quanto maior for essa percentagem. Um dos principais sistemas anti-*spam* baseados em pesos e regras é o *SpamAssassin* [16] que atualmente é um dos mecanismos mais utilizados. O *SpamAssassin*, no entanto, também utiliza outros métodos como listas negras, filtros bayesianos, verificação do DNS reverso e verificação SPF (*Sender Policy Framework*), criando regras especiais que representam esses outros tipos de teste e não são regras estáticas.

Nas soluções baseadas apenas em regras estáticas, para construir as regras, deve-se procurar características nas mensagens que correspondam à características de mensagens *spams* e legítimas para criar as regras. As regras também devem ser atualizadas freqüentemente, já que os *spams* estão em constante evolução. O problema da construção manual de regras levou ao surgimento de sistemas adaptativos, que são capazes de se adaptar a mudanças nas características dos *spams* ao longo do tempo. Como a maioria dos usuários não tem conhecimento e/ou tempo suficiente para criar suas próprias regras, geralmente um grupo de pessoas fica responsável por criar as novas regras. Esse processo, no entanto, pode causar problemas em situações específicas, principalmente para pessoas que trabalham com assuntos que estão presentes em mensagens de *spam*, fazendo com que as mensagens legítimas acabem sendo classificadas como *spam*. Uma empresa que vende produtos farmacêuticos, por exemplo, poderia ser afetada por regras que verificam se o nome de alguns medicamentos estão presentes na mensagem, como Viagra. Analisando as regras [17] utilizadas pelo sistema anti-*spam* SpamAssassin, se uma mensagem contiver nomes de certos medicamentos, a probabilidade da mensagem ser classificada como *spam* é bem alta, o que pode causar grandes transtornos para pessoas que atuam no segmento de farmácias *on-line*.

2.2.3 Filtros Bayesianos

Os sistemas que utilizam filtros bayesianos funcionam com base em métodos estatísticos que levam em conta a freqüência de ocorrência de determinadas palavras ou frases em mensagens classificadas ou não como *spam*. Um usuário de correio eletrônico quando recebe uma nova mensagem *spam* faz a sua leitura e acaba identificando e memorizando algumas palavras que provocaram a decisão da classificação da mensagem como

um *spam*. Da próxima vez que este usuário ler e identificar essas mesmas palavras em outra mensagem já terá uma suspeita maior de que a nova mensagem recebida também seja uma mensagem *spam*. Dessa forma, algumas palavras-chave acabam se constituindo em características dos *spams*. Os filtros bayesianos têm por finalidade repetir este mesmo procedimento humano de identificação de *spam* realizado pelo usuário, só que de forma automatizada. Para a identificação automatizada de *spam* um filtro bayesiano deve ser construído. O primeiro passo é um processo de aprendizagem, onde o usuário identifica manualmente mensagens legítimas e mensagens *spam*, para a construção de um filtro que permita classificar mensagens futuras como legítimas ou *spams*. Uma característica importante desses filtros é que, como eles são geralmente feitos com base nas informações de identificação de mensagens *spams* passadas pelo usuário, eles se adaptam ao padrão de *spams* e de mensagens legítimas recebidas pelo próprio usuário.

Para construir os classificadores usados para filtrar os *spam*, são utilizadas redes bayesianas. Uma rede bayesiana é um grafo acíclico direcionado que representa uma distribuição de probabilidade [18]. Nesse grafo, cada variável aleatória X_i é representada por um nó. Uma aresta entre dois nós indica a probabilidade de influência do nó pai para o nó filho. Além disso, cada nó X_i da rede é associado a uma tabela de probabilidade condicional, que determina a distribuição de X_i dados os valores dos seus pais. Um classificador bayesiano utiliza uma rede bayesiana onde existe um nó C representando uma das possíveis classes c_k que representam as possíveis classificações que o filtro pode realizar e vários filhos X_i, para cada uma das características testadas. Na Figura 2.2 é mostrado um exemplo de um grafo de um classificador bayesiano. Dessa forma, dado um conjunto de valores de X_i, pode-se calcular a probabilidade de cada classe c_k de acordo com a Equação 2.1, onde o termo $P(X = x|C = c_k)$ é dado pela Equação 2.2.

$$P(C = c_k|X = x) = \frac{P(X = x|C = c_k)P(C = c_k)}{P(X = x)} \tag{2.1}$$

$$P(X = x|C = c_k) = \prod_i P(X_i = x_i|C = c_k) \tag{2.2}$$

Para a utilização dos filtros bayesianos na classificação de mensagens, é necessário

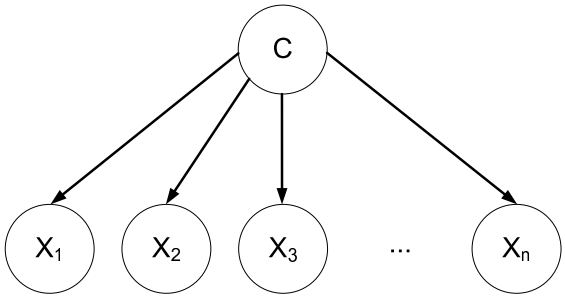


Figura 2.2: Grafo de um classificador bayesiano.

que, em primeiro lugar, as mensagens sejam representadas como vetores de características X_i a serem testadas, chamadas de símbolos. Esses vetores são construídos com base na separação das palavras da mensagem em vários símbolos. Além de considerar as palavras para a classificação das mensagens, são utilizadas outras características que estão presentes em *spams*, como o uso exagerado de exclamações e outras informações características, como o uso de *tags* HTML. Após a separação da mensagem em vários símbolos, cada um deles é comparado com sua freqüência de ocorrência em mensagens *spams* e não *spams* anteriores. Se o símbolo apareceu predominantemente em mensagens *spams*, é atribuída uma alta probabilidade deste símbolo representar uma característica de *spam*. A probabilidade do símbolo é geralmente calculada como sendo a proporção entre o percentual de aparição do símbolo em mensagens *spams* e legítimas. Para economizar espaço no filtro, somente os símbolos com probabilidade muito alta e muito baixa são armazenados, já que eles determinam as principais características.

O processo de divisão da mensagem em vários símbolos é de extrema importância no processo de classificação. Os primeiros mecanismos separavam as palavras utilizando-se sinais de pontuação, facilitando a exploração dessa vulnerabilidade pelos *spammers*, criando frases como *C/A/L/L/ N-O-W - I/T/S F_R_E_E*, onde as letras são separadas por caracteres que geralmente são utilizados para separar frases ou palavras. Se for utilizado um método convencional de separação em palavras, a frase acima produzirá apenas símbolos que são compostos de uma letra e não os símbolos que correspondem às palavras utilizadas no *spam*, dificultando a determinação se o símbolo corresponde a uma característica de *spam* ou não. Atualmente, no procedimento de separação de símbolos das mensagens são utilizados processos mais complexos, como a junção de símbolos forma-

dos por apenas uma letra, a agregação de símbolos e a tentativa de reduzir novos símbolos a símbolos que tenham uma similaridade com os já conhecidos, com o objetivo de reduzir o número de símbolos diferentes que são armazenados e tratar da mesma forma símbolos que tenham grande similaridade.

2.2.4 Verificação do Endereço DNS Reverso

A verificação do endereço de DNS reverso foi um dos primeiros mecanismos anti-*spam* baseados na verificação da origem que surgiram. O objetivo da verificação do endereço de DNS reverso é tornar mais difícil a falsificação do endereço de origem, que pode ser facilmente forjado no protocolo SMTP. O sistema DNS tem dois tipos principais de registro, os registros A e os registros PTR. Os registros A são utilizados para fazer o mapeamento entre nomes de domínio e endereços IP. Quando é feita uma consulta a um servidor de DNS para descobrir o endereço IP de um determinado domínio, o registro A é consultado. Por outro lado, pode ser feita uma requisição do registro PTR de um determinado endereço IP, para descobrir o nome de domínio registrado para ele. Esse tipo de consulta é chamado de consulta reversa, pois funciona de forma contrária ao mecanismo normal de resolução de nomes para endereços IP.

Uma maneira de verificar a autenticidade da origem é fazer uma consulta reversa do endereço IP do servidor que está tentando enviar uma mensagem. Quando o servidor do destinatário recebe uma conexão para receber uma mensagem, ele faz uma consulta DNS reversa do endereço IP do servidor que se conectou a ele. Se o endereço IP não possuir um nome associado, a mensagem é então descartada. Caso exista um nome registrado, é feita uma consulta DNS desse nome, para verificar se o endereço IP desse nome realmente corresponde ao endereço IP original. Caso os endereços IP sejam correspondentes, diz-se que o endereço de DNS reverso é válido. Em seguida, o servidor espera pelos comandos `HELO` e `MAIL FROM` do protocolo SMTP e então compara se os domínios informados nesses dois tipos de mensagem estão de acordo com o domínio que foi obtido pela verificação reversa do endereço IP. Caso os domínios sejam diferentes, a mensagem é recusada. Com essa medida, apenas os servidores cujo endereço IP tem como endereço reverso um nome que pertença ao domínio podem enviar mensagens do domínio. Na verificação do

DNS reverso, muitas vezes esse segundo passo não é executado, pois podem acontecer situações onde o servidor de correio eletrônico está em um domínio diferente do que está enviando as mensagens, como é o caso de servidores que apenas encaminham mensagens de outros domínios.

A idéia básica da verificação de DNS reverso reside no fato dos *spammers* não configurarem o endereço reverso de seus servidores, já que se forem configurados, pode-se obter o nome do domínio a que o endereço IP pertence. A partir do nome do domínio, podem-se descobrir informações sobre a pessoa que registrou o domínio, aumentando as chances de rastreamento. Muitos provedores de serviço da Internet, em geral, também não registram o endereço reverso de seus clientes. Como a maioria dos zumbis são máquinas de usuários de provedores de serviço, isto acaba fazendo com que muitas máquinas funcionando como zumbis também não passem no teste de verificação do DNS reverso.

O teste de DNS reverso tem uma baixa taxa de falsos negativos, já que ele elimina grande parte dos *spams* gerados por máquinas zumbis sem DNS reverso e por servidores que não possuem registro de DNS reverso de forma proposital para dificultar o rastreamento. Em contrapartida, sua taxa de falsos positivos é alta, acima da média dos outros sistemas anti-*spam*, pois muitos provedores legítimos de correio eletrônico não configuram corretamente seus DNSs reversos, fazendo com que as mensagens de todos os seus usuários sejam descartadas por servidores que utilizam a verificação do DNS reverso.

2.2.5 *Sender Policy Framework* (SPF)

Este mecanismo também tem como objetivo dificultar a falsificação do endereço de origem das mensagens. Seu funcionamento se baseia na publicação de informações sobre quais servidores têm permissão de enviar mensagens de um determinado domínio. Dessa forma, cada domínio fica sendo responsável por determinar quais máquinas podem enviar mensagens utilizando o domínio no endereço do remetente. Para determinar as máquinas autorizadas a enviar mensagens, o domínio define uma série de testes que devem ser realizados por outros servidores que recebam uma mensagem com o domínio do remetente igual ao domínio em questão.

As informações sobre as máquinas autorizadas a enviar mensagens são publicadas em registros no servidor DNS do domínio, utilizando um registro de DNS de modo texto também chamado TXT. O registro do SPF publicado é composto de uma parte inicial, identificada pela seqüência "v=" que especifica a versão utilizada do SPF. Atualmente somente a versão 1 está definida e tem spf1 como identificador. Em seguida à informação de versão, são definidos os mecanismos que são conjuntos de testes que podem retornar um resultado positivo ou negativo. Para cada um dos mecanismos, podem ser atribuídos modificadores, que irão determinar a ação a ser tomada caso o teste feito pelo mecanismo forneça um resultado positivo. Os mecanismos são listados de forma ordenada e são avaliados da esquerda para a direita. Caso um deles forneça um resultado positivo é tomada a ação definida pelo modificador e os outros mecanismos não são testados.

Os mecanismos definidos na RFC4408 [19] para o SPF são apresentados na Tabela 2.1 e os modificadores utilizados com esses mecanismos são apresentados na Tabela 2.2.

A seguir é mostrado um exemplo de registro SPF permitindo que apenas o servidor cujo endereço IP está associado no nome do domínio e os servidores de correio eletrônico do domínio enviem mensagens. Além disso, todas as regras utilizadas no domínio dominio.com.br também serão verificadas, permitindo que as máquinas autorizadas por ele também sejam aceitas. Todas as outras máquinas que não atendam às características anteriores são impedidas de enviar mensagens como sendo do domínio, pois o modificador − é utilizado no mecanismo all que representa qualquer outro caso.

```
v=spf1 a mx include:dominio.com.br -all
```

O mecanismo SPF, embora não seja utilizado diretamente para filtrar *spams*, pode ajudar a reduzí-los, pois torna mais difícil o envio de mensagens com endereços falsos, caso os provedores utilizem o SPF para determinar de forma precisa as máquinas autorizadas a enviar mensagens. O uso do SPF, entretanto, não impede que um *spammer* crie vários domínios e publique registros SPF permitindo que qualquer máquina envie mensagens utilizando como remetente esses domínios. Nessa situação os servidores que utilizem o mecanismo SPF vão concluir que o servidor que está tentando enviar mensagem tem autorização, aceitando a mensagem. Os *spammers*, no entanto, não conseguirão falsifi-

Tabela 2.1: Mecanismos do SPF.

Mecanismo	Descrição
all	Retorna sempre verdadeiro. É utilizado como o último mecanismo a ser executado, para definir uma ação padrão caso nenhum dos mecanismos anteriores tenha retornado um resultado positivo.
include	Utilizado para incluir os mecanismos SPF definidos em outro domínio especificado no parâmetro nome de domínio. Esse mecanismo é utilizado quando um servidor de um domínio também aceita que as suas mensagens sejam enviadas através de outros domínios. A sintaxe desse mecanismo é `include:<nome domínio>`.
a	Realiza uma consulta DNS ao nome do domínio para verificar se o endereço IP do servidor que está enviando a mensagem é um dos endereços IP associados ao nome do domínio indicado.
mx	Consulta através do protocolo DNS se o endereço IP de origem corresponde a um dos servidores de correio eletrônico do domínio, se servindo dos registros MX do DNS.
ptr	Realiza o teste do DNS reverso, apresentado na Seção 2.2.4.
ip4	Define uma faixa de endereços IPv4 que estão autorizados a enviar mensagens. A sintaxe desse mecanismo é `ip4:<endereço de rede>/<máscara de sub-rede>`.
ip6	Similar ao mecanismo anterior só que utilizado para testar faixas de endereços IPv6. Sua sintaxe é a mesma do mecanismo ip4.
exists	Permite a utilização de macros para criar um determinado nome de domínio baseado em informações da mensagem. Com base no nome de domínio que foi criado, verifica se o domínio possui um registro de DNS válido. Caso o registro DNS seja válido, o mecanismo retorna um resultado positivo. A principal utilização desse mecanismo é em conjunto com as listas negras DNSBL.

car endereços de provedores legítimos que usem o SPF, já que os mesmos podem definir políticas permitindo que somente seus servidores enviem mensagens como sendo do seu domínio. Embora um registro SPF devidamente configurado garanta que apenas os servidores de correio eletrônico do domínio especificados na política do SPF sejam aceitos como remetentes de mensagens do domínio, deve-se utilizar métodos para a autenticação dos clientes que enviam mensagens para os servidores do domínio.

Cada um dos mecanismos apresentados possui vantagens e desvantagens. Para se verificar a eficiência, em termos de falsos positivos e falsos negativos, e também o de-

Tabela 2.2: Modificadores do SPF.

Modificador	Descrição
+	A mensagem é classificada como em conformidade com a política definida e o remetente é autorizado pelo domínio a enviar mensagens. Esse é o modificador padrão, tornando opcional sua definição explícita.
-	A mensagem não está de acordo com a política e o remetente não está autorizado a enviar mensagens como sendo do domínio.
?	O resultado da análise é neutro.
$\sim$	Define que o remetente provavelmente não está autorizado a enviar mensagens como sendo do domínio, mas não é feita nenhuma afirmação da sua autenticidade, permitindo assim que o servidor trate esse caso de forma diferente dos casos em que o servidor garante que o remetente está autorizado, ou não, a enviar mensagens.

sempenho, em termos de recursos requeridos, de cada um destes mecanismos de forma isolada é necessário o desenvolvimento de uma ferramenta especial que é apresentada na próxima seção.

2.3 O Sistema de Análise de *Spam* - ADES

Uma avaliação das técnicas anti-*spam* atuais é importante para verificar a taxa de falsos positivos e falsos negativos e também o custo computacional. Grande parte dos sistemas anti-*spam* existentes avaliam as mensagens, as classificam e as bloqueiam quando são consideradas *spams*. Dessa forma, por causa do bloqueio da mensagem, não é possível avaliar a eficiência de cada um dos sistemas individualmente para cada mensagem recebida. Também não se encontrou nenhum sistema que permitisse esta avaliação. Assim, o sistema ADES (**A**nálise **DE** **S**pam), descrito neste capítulo, foi desenvolvido para realizar a análise da eficiência dos mecanismos anti-*spam* [20]. No ADES, todos os mecanismos anti-*spam* implementados analisam cada mensagem, verificando a ação que seria tomada por cada mecanismo, mas não descartam a mensagem. A Figura 2.3 apresenta a arquitetura do sistema ADES, que é composta por três módulos detalhados a seguir.

O sistema ADES possui um Pote de Mel que foi desenvolvido para divulgar endereços

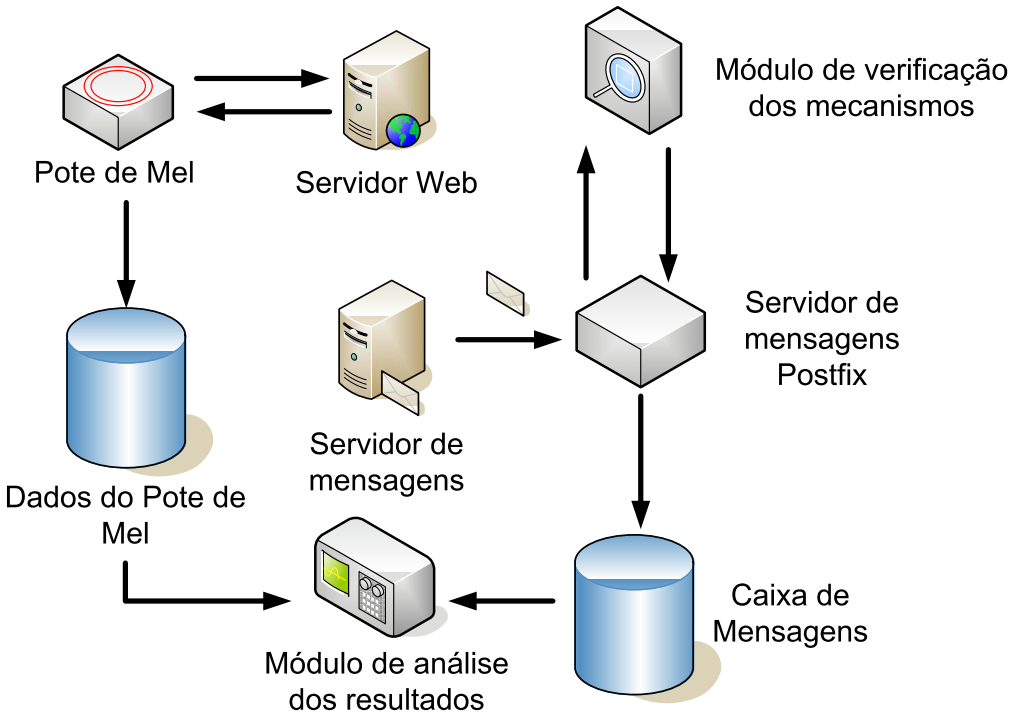


Figura 2.3: Arquitetura do sistema ADES.

eletrônicos que não pertencem a usuários reais. O objetivo é fazer com que os *spammers* capturem estes endereços e os incluam em suas listas de destinatários [21, 22]. O Pote de Mel permite uma identificação trivial de mensagens *spams*, uma vez que todas as mensagens enviadas para os endereços divulgados como iscas são mensagens *spams*, já que não são endereços eletrônicos utilizados por usuários legítimos. Na implementação do Pote de Mel são gerados aleatoriamente endereços eletrônicos que são divulgados na página web principal do Grupo de Teleinformática e Automação (GTA). A cada acesso a pagina são gerados e divulgados cinco endereços eletrônicos diferentes. Para cada endereço eletrônico divulgado, são armazenadas em um banco de dados as informações sobre a data de divulgação e o endereço IP da máquina que acessou a página. Assim, como cada endereço é divulgado apenas uma vez, as informações sobre o processo de coleta de endereços pelos *spammers* podem ser obtidas através do registro da base de dados da divulgação do endereço quando uma mensagem é enviada para um dos endereços divulgados no Pote de Mel [20].

O módulo de verificação dos mecanismos realiza os testes de desempenho dos mecanismos anti-*spam* analisados. O módulo foi implementado na linguagem Perl e funciona

como um servidor de políticas [23] do servidor de mensagens Postfix. A arquitetura de servidores de política do Postfix permite a criação de módulos que recebem as informações das mensagens que chegam ao servidor de correio eletrônico e que definem a ação a ser tomada com a mensagem. Todas as mensagens recebidas são analisadas através dos mecanismos de listas negras, DNS reverso e SPF. Para o mecanismo de listas negras são consultadas as cinco listas negras[3] mais utilizadas atualmente [12]. Após a realização dos testes de cada um dos mecanismos, os resultados são armazenados adicionando-se uma linha ao cabeçalho da mensagem.

O servidor de correio eletrônico utilizado está configurado para utilizar o mecanismo de pesos e regras *SpamAssassin* com a configuração padrão e a utilização da base de dados distribuída de *spams* chamada Razor [24]. Dessa forma, as informações deste mecanismo são adicionadas automaticamente ao cabeçalho da mensagem pelo servidor de correio eletrônico. O mecanismo de filtros bayesianos não foi implementado no sistema ADES, uma vez que existem implementações desse mecanismo que possibilitam a análise das mensagens sem bloqueá-las. Dessa forma, todas as mensagens podem ser avaliadas por todos os mecanismos testados.

Para a avaliação de desempenho do mecanismo de filtros bayesianos é utilizado o mecanismo de filtros bayesianos do programa de correio eletrônico Mozilla Thunderbird. A metade inicial das mensagens legítimas e a metade inicial dos *spams* são separadas para treinar o filtro. Na outra metade aplica-se o filtro. Assim, as mensagens mais antigas são utilizadas para treinar o filtro, simulando o cenário em que o filtro é treinado com as mensagens recebidas e utiliza-se o filtro para classificar novas mensagens.

O módulo de análise dos resultados tem como objetivo realizar a análise das características dos processos de coleta de endereços e envio de *spams*, além de análises dos resultados dos mecanismos anti-*spam*. Antes de calcular os índices de falsos positivos e de falsos negativos de cada um dos mecanismos, é necessário saber se a mensagem é legítima ou *spam*, para comparar com o resultado do mecanismo. Para isso, todas as mensagens recebidas por dezoito usuários diferentes foram classificadas manualmente como

[3] As listas DNSBL consultadas foram: sbl-xbl.spamhaus.org, cbl.abuseat.org, dnsbl.sorbs.net, list.dsbl.org e bl.spamcop.net.

legítimas ou *spams*. As mensagens enviadas para endereços divulgados no Pote de Mel foram automaticamente consideradas como *spams*. O número total de mensagens utilizadas na análise foi de 63.325 mensagens legítimas e 3.392.931 *spams*, recebidas pelos dezoito usuários legítimos e pelo Pote de Mel durante um ano e seis meses.

2.3.1 Resultados

A Figura 2.4 mostra os percentuais de falsos positivos e falsos negativos dos mecanismos. Muitos domínios não publicam os registros SPF e, portanto, o universo de mensagens usadas no teste para este mecanismo foi de 53,6% das mensagens legítimas e 19,9% dos *spams*. A taxa de falsos positivos para os mecanismos de SPF, listas negras e DNS reverso foi consideravelmente alta, chegando a 17,1% das mensagens para o mecanismo SPF. A alta taxa de falsos positivos para o mecanismo de consulta ao DNS reverso mostra que muitas mensagens legítimas provêm de servidores legítimos mal configurados quanto ao DNS reverso. O elevado número de falsos positivos das listas negras se justifica em máquinas infectadas sem que o usuário perceba, atraso na retirada do endereço IP de uma máquina que já foi desinfectada e, principalmente, devido à inclusão de endereços IP de provedores de serviço. Dessa forma, a utilização individual desses mecanismos como forma de classificar uma mensagem como *spam* pode gerar uma alta taxa de falsos positivos, causando uma insatisfação dos usuários. Os mecanismos de pesos e regras e filtros bayesianos obtiveram as menores taxas de falsos positivos, embora ainda sejam altas ao considerar-se o impacto negativo que podem causar aos usuários. Em todos os mecanismos a taxa de falsos negativos é alta, chegando a 67,4% para o mecanismo de DNS reverso, mostrando a ineficiência desse mecanismo. A alta taxa de falsos negativos para o DNS reverso é causada por *spammers* que invadem máquinas de terceiros com o DNS reverso configurado corretamente e as utilizam para enviar *spams*.

A Figura 2.5 apresenta o gráfico da função de distribuição cumulativa (CDF) do percentual de mensagens em função do tempo necessário para a verificação de cada mecanismo. Esse tempo é afetado por fatores como a disponibilidade da rede e a configuração do computador. Como todas as verificações são realizadas utilizando o mesmo computador e a mesma conexão de rede à Internet, os resultados podem ser comparados.

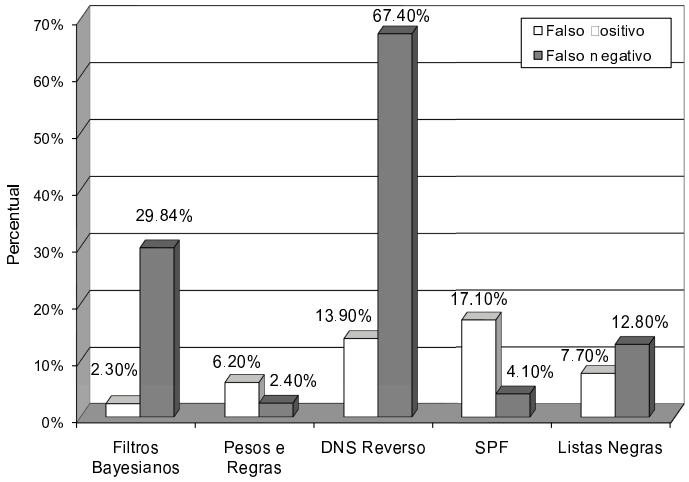


Figura 2.4: Percentual de falsos positivos e falsos negativos.

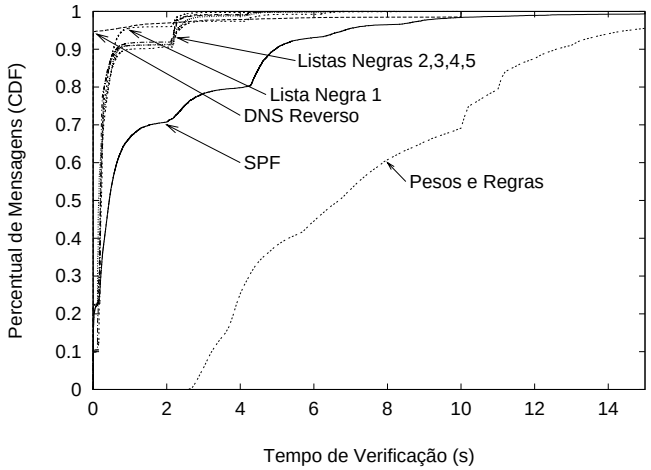


Figura 2.5: Percentual acumulado de mensagens em função do tempo de verificação.

Apesar dos mecanismos de SPF, listas negras e DNS reverso utilizarem apenas consultas DNS, o DNS reverso apresenta melhores resultados. Isto se deve ao fato de que a implementação do servidor de correio eletrônico já realiza a consulta do DNS reverso a partir do endereço IP do remetente da mensagem. Dessa forma, quando o mecanismo de DNS reverso realiza novamente a consulta, o resultado já se encontra no *cache*, reduzindo o tempo de consulta. Dentre as listas negras consultadas, a lista negra 1 obteve um resultado melhor em relação às outras listas e corresponde à lista negra mais utilizada. Já o mecanismo de pesos e regras obteve o pior resultado quando comparado com os outros, levando até 12 segundos em 90% dos casos. O menor desempenho desse mecanismo pode

ser explicado pela realização da análise de conteúdo, o que requer um processamento significativamente maior. O mecanismo de filtros bayesianos não foi analisado em relação ao tempo de verificação, pois essa informação não é disponibilizada pelo programa que implementa o mecanismo.

Os resultados mostraram uma taxa de falsos negativos alta, entre 2,3% e 67,4%. Também foi observada uma taxa de falsos positivos acima de 2,3% para todos os mecanismos, o que é alto considerando-se o impacto negativo que um falso positivo pode causar para os usuários. O mecanismo de filtros bayesianos obteve o melhor resultado, com 2,3% de falsos positivos. Mesmo assim, a taxa de falsos positivos dos mecanismos atuais é consideravelmente alta, considerando o impacto que os falsos positivos causam para os usuários.

Capítulo 3

O Mecanismo Proposto

OS falsos positivos causam um grande prejuízo e transtorno para os usuários. Conforme a análise feita na Seção 2.3 para os mecanismos anti-*spam* utilizados atualmente, mesmo o mecanismo com melhor desempenho ainda possui uma taxa de falsos positivos de 2,3%. Nos experimentos realizados com o sistema ADES, os usuários recebem em média 3.000 mensagens por ano e, portanto, 69 mensagens legítimas seriam classificadas como spam e apenas uma destas pode causar um prejuízo financeiro enorme ao usuário.Para reduzir os falsos positivos, o mecanismo anti-*spam* proposto utiliza um mecanismo de autenticação dos remetentes e um mecanismo de reputação para a troca de informações entre os servidores sobre os usuários autenticados [25]. A reputação dos usuários é determinada a partir do histórico de mensagens enviadas e, quanto melhor for a reputação, menor é a probabilidade das mensagens serem classificadas como *spam*, diminuindo os falsos positivos.

3.1 Trabalhos Relacionados

Atualmente, vários mecanismos anti-*spam* são utilizados, tais como as listas negras, filtros bayesianos e mecanismos baseados em pesos e regras [4] que foram avaliados pela ferramenta apresentada na Seção 2.3. No entanto, esses mecanismos geralmente apresentam taxas de falsos positivos relativamente altas [26]. Assim, novas técnicas que se

servem de reputação ou redes sociais vêm sendo adotadas [27–42] em conjugado com os mecanismos convencionais, para tomar a decisão final, aumentando a precisão da classificação.

Golbeck e Hendler propõem um mecanismo de reputação baseado em redes sociais [27]. As redes sociais correspondem ao grafo que representa a comunicação entre usuários [28–31]. Cada usuário representa um nó do grafo e uma aresta entre dois nós significa que os usuários já trocaram mensagens. Nesse mecanismo, cada usuário define a reputação dos usuários com os quais ele troca mensagens. Quando uma mensagem de um usuário desconhecido é recebida, procura-se no grafo da rede social se existe algum caminho entre o usuário que recebeu a mensagem e o usuário que a enviou. Caso exista o caminho, a reputação de cada nó é levada em consideração para determinar a reputação do usuário que era desconhecido. Para um nó i calcular a reputação de um nó j, a Equação 3.1 é utilizada, onde k são os vizinhos do nó i, n é o número de vizinhos e R_{ij} é a reputação do nó j calculada pelo nó i.

$$R_{ij} = \frac{\sum_{j=0}^{n} \left\{ \begin{array}{ll} (R_{kj}R_{ik}) & \text{se } R_{ik} \geq R_{kj} \\ R_{ik}^2 & \text{se } R_{ik} < R_{kj} \end{array} \right\}}{n} \tag{3.1}$$

O inconveniente deste mecanismo é que cada usuário deve atribuir manualmente a reputação dos outros usuários. Além disso, esse mecanismo causa a perda da privacidade dos usuários, já que a informação de quais usuários se comunicam com quais usuários se torna pública para todos os usuários do sistema. Como o mecanismo não utiliza um sistema de autenticação dos remetentes, um *spammer* pode simplesmente enviar as mensagens se passando por um usuário legítimo, aumentando a taxa de falsos negativos.

Outro tipo de *spam* que está surgindo é o *spam* através de VoIP [32]. Balasubramaniyan *et al.* propõem um mecanismo baseado em redes sociais para combater *spams* em VoIP [33]. Diferente do mecanismo descrito acima, de Golbeck e Hendler, em que os usuários definem manualmente a reputação dos usuários que conhecem, no mecanismo de Balasubramaniyan a reputação de cada usuário é determinada através do tempo das ligações VoIP entre os usuários. A proposta se baseia no fato de que quando um usuário

receber um *spam* através de uma ligação VoIP, ele irá desligar rapidamente. Já as ligações entre usuários legítimos tendem a ser mais longas e, quanto mais tempo duas pessoas se falam, maior é a probabilidade da pessoa não ser um *spammer*. Dessa forma, toda vez que um usuário A encerra uma ligação com um usuário B, o usuário B envia para A uma credencial assinada com a chave privada de B atestando que A realizou uma ligação durante um determinado intervalo de tempo com B. Na Figura 3.1 é mostrado de forma esquemática o processo de envio de credenciais. Supondo que um usuário C realize uma ligação para B, ele também receberá uma credencial de B com o tempo de ligação. Agora, caso o usuário C realize uma ligação para o usuário A, o usuário C pode apresentar a credencial emitida por B. Como A já realizou uma chamada com B, o usuário A já possui uma determinada confiança no nó B e a credencial que B emitiu para C, prova que B confia em C, então o usuário A também confia no usuário C. O valor da reputação entre os usuários, calculado a partir do tempo de ligação, é utilizado para criar a rede social entre os usuários. Dessa forma, apenas usuários que recebam uma credencial de um usuário que já participa da rede social conseguem fazer parte da rede social. Um *spammer* que tente enviar um *spam* através de uma ligação VoIP não terá sucesso, uma vez que não pertence a uma rede social. Entretanto, um novo usuário do sistema, que conseqüentemente não pertence a nenhuma rede social, terá todas as suas chamadas bloqueadas. Para poder enviar mensagens um novo usuário tem que passar a pertencer a rede social e isto pode ser conseguido quando um usuário qualquer, que já pertença à rede social, lhe fizer uma ligação e fornecer-lhe uma credencial. Para resolver este problema devem existir outros mecanismos para permitir que os novos usuários ganhem uma confiança inicial para participarem na rede social.

Seigneur *et al.* propõem um mecanismo de reputação baseado em redes sociais com um mecanismo de autenticação [34]. O mecanismo de autenticação envia para o destinatário a nova mensagem juntamente com o resumo das mensagens enviadas anteriormente, para identificar se o usuário que está enviando a nova mensagem é o mesmo usuário que já enviou mensagens passadas. No entanto, mensagens enviadas podem chegar atrasadas ou serem perdidas, prejudicando a comparação das mensagens enviadas pelo remetente e recebidas pelo destinatário e fazendo com que os usuários não sejam autenticados. Para avaliar a reputação de cada usuário, os usuários devem marcar cada uma das mensa-

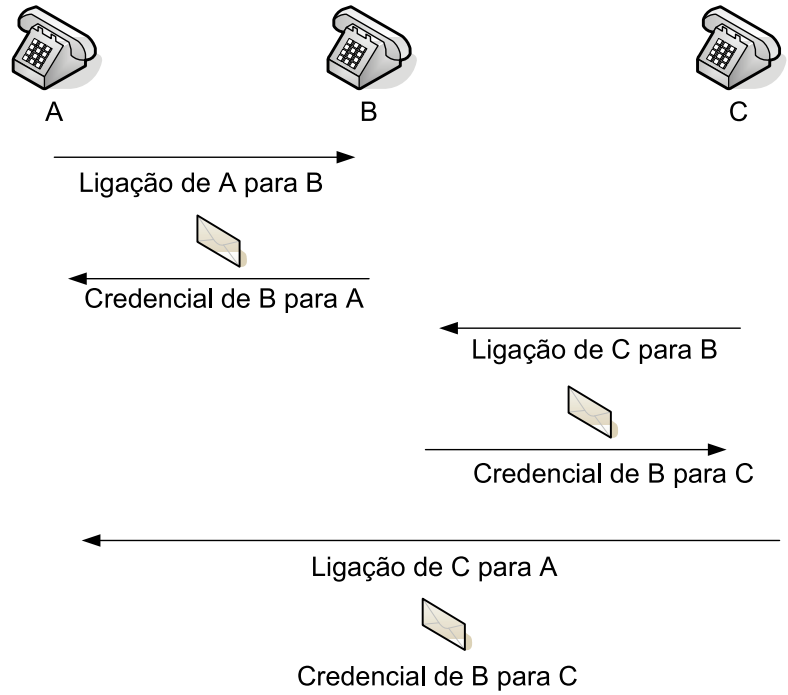


Figura 3.1: Troca de credenciais na proposta de Balasubramaniyan *et al.*

gens como *spam* ou legítima. O número de mensagens enviadas classificadas como *spam* e como legítimas é então utilizado para calcular a reputação do usuário. Quando uma mensagem é recebida de um usuário desconhecido, o usuário que recebeu a mensagem consulta alguns usuários pré-determinados para determinar se algum desses usuários tem informações a respeito do remetente da mensagem. Os usuários que são consultados são determinados manualmente pelo usuário que realiza a consulta e todas as recomendações recebidas desses usuários são consideradas confiáveis. Na consulta aos usuários sobre a reputação do remetente, os usuários que são consultados podem repassar a consulta para outros usuários. No entanto, os usuários que recebem uma consulta que já foi repassada não podem repassá-la novamente. Esse limite é imposto para diminuir os efeitos de um usuário adicionar incorretamente um *spammer* à lista de pessoas confiáveis. Nesse caso o *spammer* poderia sempre responder que a reputação dos *spammers* é alta. No entanto, essa característica acaba limitando a disseminação da informação de reputação entre os usuários.

Os mecanismos baseados em redes sociais discutidos anteriormente consultam e verificam a rede social dos usuários de diferentes servidores para formar uma visão global

da rede social formada pelos usuários. Embora eficaz na classificação dos *spams*, estas propostas eliminam a privacidade dos usuários, pois a rede social revela com quem todos os usuários se comunicam e, além disso, quais são as pessoas com as quais cada usuário mais se comunica, o que é uma grande invasão de privacidade.

McGibney e Botvich definem um mecanismo chamado TOPAS que se baseia na reputação dos servidores [3, 35]. A reputação dos servidores, R, é determinada a partir da classificação das mensagens através de um mecanismo anti-*spam* auxiliar. O mecanismo auxiliar realiza uma classificação preliminar da mensagem e a reputação R que o servidor i possui em relação ao servidor j é calculada de acordo com a Equação 3.2, onde S é igual a um caso o mecanismo auxiliar tenha classificado a mensagem como legítima ou zero caso contrário. O parâmetro α define a relevância da última classificação em relação ao passado. Assim, quanto maior o número de mensagens classificadas como legítimas, maior é o valor da reputação do servidor. Devido ao valor do parâmetro S, a reputação dos servidores sempre estará no intervalo $[0, 1]$. O valor da reputação de servidores desconhecidos é considerado igual a 0,5, já que não existe nenhuma informação que indique que o servidor é legítimo ou que envia *spam*. Nesse mecanismo, no entanto, a reputação dos usuários não é avaliada, pois se considera que os servidores de usuários legítimos só enviam mensagens legítimas e os servidores utilizados pelos *spammers* só enviam *spam*.

$$R_{ij} = \alpha S + (1 - \alpha) R_{ij} \tag{3.2}$$

O valor da reputação é utilizado para ajustar o limiar de classificação do mecanismo auxiliar e tomar a decisão final se a mensagem será aceita ou não. O limiar final é calculado como sendo 10 vezes o valor da reputação. Para determinar a reputação dos servidores, são utilizadas informações locais calculadas conforme descrito anteriormente e recomendações recebidas de outros servidores. Cada servidor também pode requisitar as reputações observadas por outros servidores. Quando um servidor i recebe de um servidor k uma recomendação de um servidor j, o valor da reputação de j calculada por i é atualizado de acordo com a Equação 3.3.

$$R_{ij} = \beta R_{ik} R_{kj} + (1 - \beta R_{ik}) R_{ij} \tag{3.3}$$

O grupo de servidores com os quais o servidor troca informações deve ser definido manualmente, o que é uma desvantagem do mecanismo. Outra desvantagem desse mecanismo é que apenas a reputação dos servidores é considerada, permitindo que *spammers* se aproveitem da reputação de servidores legítimos.

Oliveira [36] propõe um mecanismo também baseado na reputação dos servidores. Nesse mecanismo, a confiança do servidor determina o número máximo de mensagens que o servidor pode enviar por dia. O valor do número máximo de mensagens que podem ser enviadas por dia é calculado multiplicando a reputação do servidor por uma constante que determina o número máximo de mensagens. Assim, servidores com baixa reputação podem enviar menos mensagens por dia. No entanto, mesmo servidores que possuem um alto valor de reputação podem exceder o número máximo de mensagens por dia, causando uma alta taxa de falsos positivos. A reputação local de cada servidor é determinada com base apenas nas mensagens que o servidor recebeu e a reputação global do servidor é calculada com a troca de informações entre servidores. Por fim, o valor final da reputação do servidor é calculado de acordo com a média aritmética da reputação local e a global. Este mecanismo se baseia apenas na reputação dos servidores, sem levar em conta a reputação dos usuários.

Taylor [37] propõe um sistema de reputação juntamente com um mecanismo de autenticação dos remetentes que tanto pode ser o SPF quanto o *DomainKeys* [43]. Esses dois mecanismos de autenticação não autenticam o usuário, mas sim o domínio a que o usuário pertence. Dessa forma, apenas usuários que pertencem ao domínio conseguem se autenticar. A reputação é calculada para cada domínio de acordo com as mensagens recebidas pelo servidor. Quando uma nova mensagem é recebida, o processo de autenticação do domínio é realizado e, caso a reputação do domínio seja alta, a mensagem é automaticamente classificada como legítima. Caso a reputação do domínio seja baixa ou desconhecida, outro mecanismo anti-*spam* é utilizado para classificar a mensagem. Os usuários podem participar do processo, indicando as mensagens classificadas incorretamente. Essa informação é então utilizada para atualizar o valor da reputação do domínio. O cálculo da reputação é apenas local, não sendo trocadas informações entre servidores. Embora não se utilize recursos para consultar outros servidores, esse mecanismo possui um desempenho menor para servidores que recebem poucas mensagens, já que a avalia-

ção da reputação dos domínios fica prejudicada pelo baixo número de usuários.

Kamvar *et al.* mostram um mecanismo de reputação genérico para redes P2P. Nesse modelo cada nó calcula a reputação localmente e um mecanismo distribuído determina a reputação global de cada nó [38]. O mecanismo realiza várias iterações consultando outros nós até a reputação do nó consultado convergir. A desvantagem deste mecanismo de reputação é que, apesar de existir uma maior precisão na determinação da reputação, o número de mensagens trocadas é grande devido ao número de iterações do mecanismo.

Em resumo, pode-se afirmar que os mecanismos convencionais avaliados no capítulo precedente não são suficientemente eficazes para reduzir os falsos positivos. Diversos trabalhos comprovam que o uso de mecanismos de autenticação e de relações sociais é uma estratégia válida para se acrescentar aos mecanismos convencionais, pois melhoram a eficácia dos sistemas anti-*spam*. No entanto, tanto a autenticação dos usuários quanto o uso de técnicas de relações sociais não devem comprometer a privacidade do usuário. Por outro lado, os mecanismos devem se preocupar com o número de mensagens que são necessárias para implementá-los. Portanto, o mecanismo proposto neste trabalho objetiva a identificação e determinação da reputação dos remetentes das mensagens sem violar sua privacidade e, também, a determinação da reputação dos servidores. A proposta também objetiva não permitir que os *spammers* se aproveitem de servidores de usuários legítimos com boa reputação, tornando a proposta robusta a ataques dos *spammers*.

3.2 O Mecanismo de Autenticação por Pseudônimos

Para avaliar o histórico de comportamento dos usuários, o mecanismo de reputação deve ser capaz de identificá-los. Para isso é necessário um mecanismo de autenticação dos remetentes, pois o protocolo SMTP não possui nenhum mecanismo de autenticação dos remetentes, que podem ser facilmente forjados [44]. A técnica de autenticação convencional utiliza certificados digitais emitidos por autoridades certificadoras. Nesse caso, antes da emissão do certificado, é realizado um processo de verificação se as informações pessoais do usuário, que farão parte do certificado, estão corretas. Esse processo é realizado para garantir a relação entre o certificado e a identidade da pessoa. No entanto,

a certificação é dispendiosa e elimina a privacidade dos usuários, pois os certificados emitidos pelas autoridades certificadoras possuem dados pessoais dos usuários [45]. Assim, este trabalho propõe um mecanismo de autenticação dos usuários por pseudônimos simples e eficaz que não utiliza informações pessoais do remetente, garantindo a sua privacidade. Para isso, o mecanismo de autenticação por pseudônimos proposto baseia-se em pseudônimos sem relação com a identidade do usuário. A utilização de pseudônimos ao invés da autenticação convencional é simples, mantém a privacidade do remetente e bem flexível. Todo usuário pode possuir um ou mais pseudônimos além de poder se servir de pseudônimos diferentes para se comunicar com grupos de pessoas diferentes. Caso o usuário possua vários endereços de correio eletrônico, pode também utilizar um mesmo pseudônimo para enviar mensagens através de todos os endereços eletrônicos, já que o pseudônimo não possui nenhuma relação com a identidade do usuário. Os pseudônimos utilizados no processo de autenticação são compostos por um certificado digital e uma chave privada. A diferença para os certificados emitidos por autoridades certificadoras é que os certificados dos pseudônimos não possuem informações pessoais do usuário. Somente o usuário que é o dono do pseudônimo possui a chave privada correspondente à chave pública do pseudônimo. O certificado pode ser trocado livremente entre os usuários e a sua chave pública é utilizada como identificador do pseudônimo após o processo de autenticação. O certificado do pseudônimo, no entanto, não possui informações pessoais do usuário.

A geração dos pseudônimos pode ser feita localmente pelo próprio usuário ou através de um serviço de geração de pseudônimos. Os pseudônimos podem ser gerados através de um conjunto de servidores responsáveis por gerá-los. Nesse caso, o servidor pode requisitar a resolução de um desafio computacional antes de gerar o pseudônimo para o usuário [46–48]. Os certificados dos pseudônimos emitidos por estes servidores são assinados com a chave privada do servidor. Dessa forma é possível verificar posteriormente se o pseudônimo foi emitido por algum dos servidores que são considerados como confiáveis para a geração de pseudônimos. A necessidade de realizar um desafio computacional antes de obter o pseudônimo é uma vantagem, pois restringe que os *spammers* obtenham um grande número de pseudônimos, já que irão necessitar de uma alta capacidade computacional para resolver os desafios. No entanto, a desvantagem desse método

é a necessidade da criação de uma infra-estrutura para a geração dos pseudônimos e também a definição de quais servidores de geração de pseudônimos são considerados como confiáveis ou não. A outra opção para a geração de pseudônimos é a geração local. Nesse caso, o próprio usuário gera um certificado auto-assinado e utiliza-o como seu pseudônimo. A vantagem deste método é não necessitar de uma infra-estrutura adicional para a geração dos pseudônimos. No entanto, *spammers* que desejem criar vários pseudônimos têm a vantagem de criá-los de forma fácil.

O processo de autenticação por pseudônimos é baseado no modelo desafio-resposta e não requer alterações no protocolo SMTP. Nesse modelo de autenticação, uma entidade A que deseja se autenticar com uma entidade B, faz uma requisição de desafio para a identidade B. A entidade B envia para A um desafio que apenas a entidade A pode resolver e outra identidade que tente se passar por A não consegue resolver. A entidade A responde o desafio e envia a resposta para B, que verifica a resposta. Caso a resposta esteja correta, a entidade B autentica com sucesso a entidade A. Uma forma de implementar este modelo de autenticação é através do uso de chaves assimétricas. O desafio enviado à entidade que deseja se autenticar é apenas uma seqüência de caracteres que deve ser assinada digitalmente pela entidade que deseja se autenticar. Junto com a assinatura digital do desafio é enviada a chave pública da entidade que está se autenticando. Dessa forma, pode-se verificar se a assinatura digital do desafio é válida e foi realizada pela entidade que possui a chave privada correspondente à chave pública que foi enviada. Se uma entidade deseja se passar por outra e possui apenas a chave pública da entidade pela qual quer se passar não obterá sucesso, uma vez que não possui a chave privada correspondente e, por isso, não poderá realizar a assinatura digital do desafio.

A Figura 3.2 mostra o esquema de funcionamento do mecanismo de autenticação. O processo de autenticação é realizado com todos os servidores de destino. Os servidores de destino são determinados através do domínio dos destinatários da mensagem. Caso uma mensagem seja endereçada a vários destinatários em diferentes domínios, o processo de autenticação será realizado com os servidores de cada um dos domínios. Para iniciar o processo de autenticação, o cliente antes de enviar a mensagem ao seu servidor de correio eletrônico, envia um pedido de autenticação ao servidor de correio eletrônico de cada destinatário da mensagem. Cada um dos servidores de destino envia para o cliente um desafio

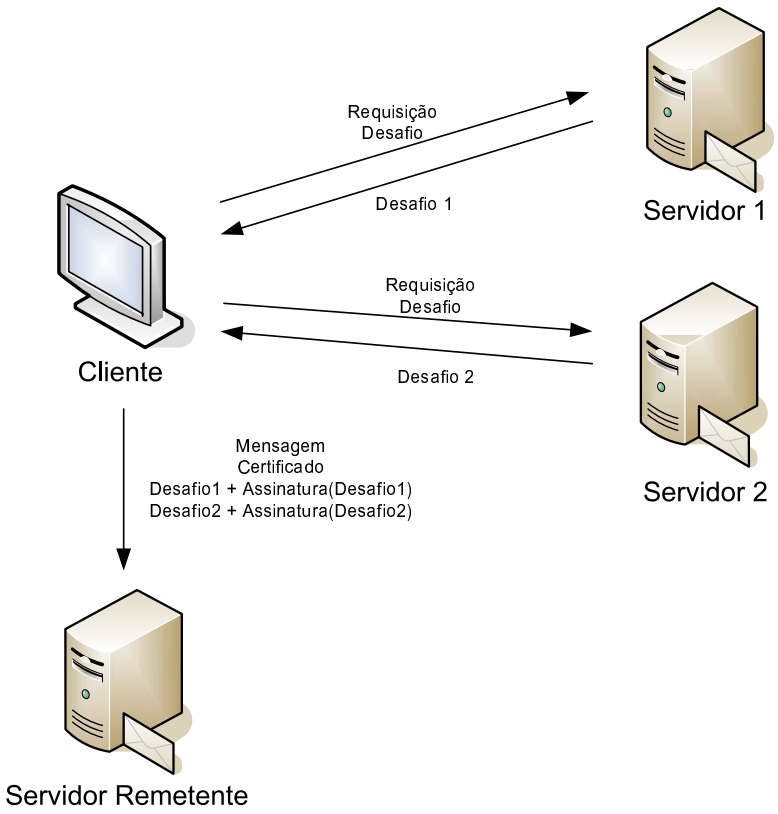


Figura 3.2: Esquema do processo de autenticação.

que é simplesmente uma seqüência de caracteres. O servidor armazena em uma tabela o desafio enviado e o instante em que o desafio é enviado para, posteriormente, verificar se a resposta corresponde a um desafio enviado pelo servidor. Quando uma resposta ao desafio é recebida, o desafio enviado é retirado da tabela. O objetivo de armazenar o desafio enviado em uma tabela é para impedir que um usuário malicioso obtenha a resposta de um desafio, que foi enviada por um usuário legítimo, e que utilize essa resposta em suas mensagens, se passando pelo usuário legítimo. Se um atacante (homem no meio) obtiver acesso a uma resposta de desafio e tentar utilizar novamente essa reposta, o servidor irá verificar que o desafio que está sendo respondido não se encontra mais na tabela de desafios, não aceitando a resposta. O campo de desafio possui um tamanho de 64 bits o que permite até 2^{64} desafios diferentes. Este valor de 64 bits inviabiliza o ataque de um usuário malicioso que tenha obtido a resposta de um desafio e que, por força bruta, requisite vários desafios até que receba um desafio igual ao que possui e foi respondido pelo usuário legítimo. A estampa de tempo na tabela de desafios é utilizada para remover entradas antigas da tabela, excluindo desafios antigos que não foram respondidos durante

um determinado intervalo de tempo. A Tabela 3.1 mostra de forma esquemática a tabela que armazena os desafios enviados e o tamanho em bits de cada campo.

Tabela 3.1: Tabela de desafios enviados.

Desafio	Estampa de Tempo
64 bits	32 bits

No processo de autenticação, o cliente responde aos desafios assinando digitalmente cada um dos desafios com a sua chave privada e adiciona no cabeçalho da mensagem a ser enviada uma linha contendo sua chave pública e uma linha para cada uma das respostas dos desafios. Utilizando o protocolo SMTP, a mensagem é enviada para o servidor de correio eletrônico do remetente, que a encaminha até os servidores dos destinatários da mensagem. A partir desse ponto, cada servidor dos destinatários verifica se alguma das respostas é de um desafio enviado por ele. Caso exista a resposta para o desafio, o pseudônimo só é autenticado se a assinatura digital estiver correta. Caso não ocorra sucesso nas verificações da resposta e da assinatura digital, o processo de autenticação falha. A Figura 3.3 mostra de forma esquemática o fluxo do processo de autenticação.

Depois do processo de autenticação do pseudônimo, a chave pública do pseudônimo pode ser utilizada como um identificador do pseudônimo. Utilizando esse identificador, o servidor pode buscar informações tanto localmente quanto remotamente através do mecanismo de reputação para determinar o histórico de comportamento do pseudônimo e determinar a reputação.

3.3 O Mecanismo de Reputação

O mecanismo de reputação proposto troca informações entre os servidores sobre os pseudônimos já autenticados. Para realizar essa troca de informações, cada servidor leva em conta a reputação dos outros servidores, o que vai determinar a confiança na informação vinda destes outros servidores. A reputação de um servidor é determinada a partir do total de mensagens legítimas e *spams* que foram enviadas por esse servidor. Nesta etapa, para classificar as mensagens, é utilizado um mecanismo anti-*spam* convencional

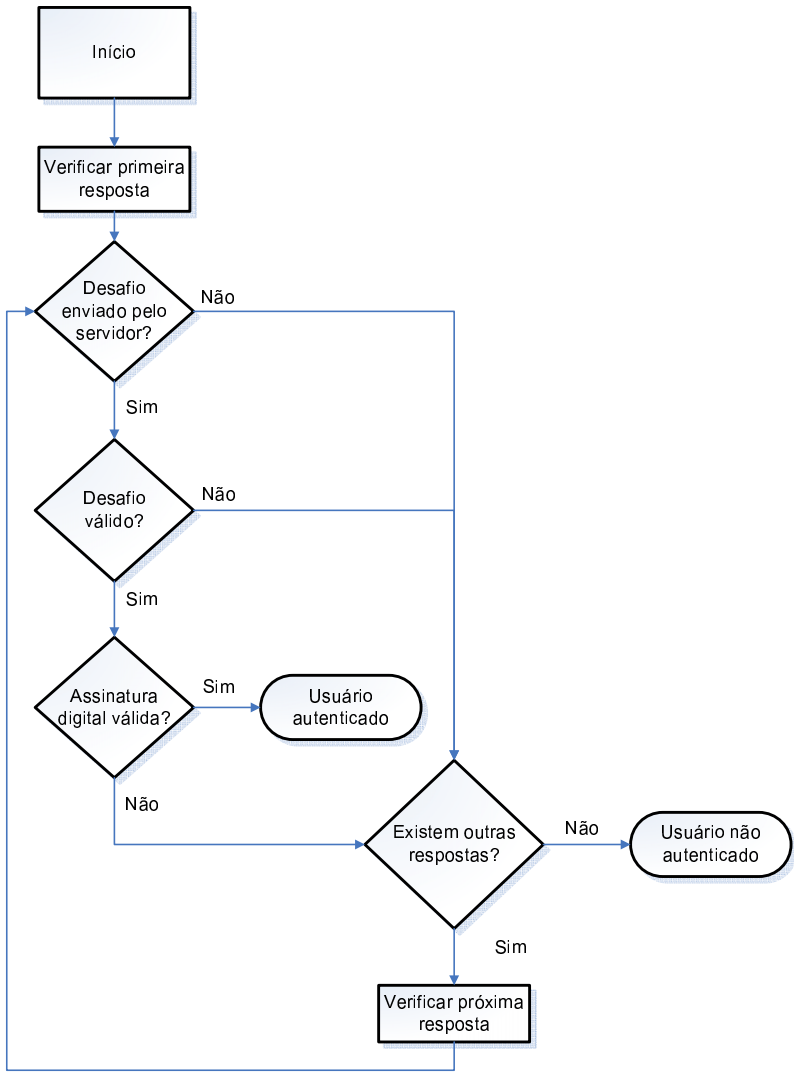


Figura 3.3: Fluxo do processo de autenticação.

(listas, filtros bayesianos, pesos e regras etc.) como mecanismo auxiliar. Caso a mensagem seja classificada como legítima, a média do servidor, que representa sua reputação, é atualizada com o valor 1. Caso a mensagem seja classificada como *spam*, a média é atualizada com o valor -1. É possível que servidores que apresentam uma boa reputação durante um longo período possam mudar de comportamento e passem a enviar *spams*. Assim, para que esta mudança de comportamento seja detectada rapidamente, o cálculo da reputação é realizado de acordo com a média móvel exponencial, que considera pesos maiores aos valores observados mais recentemente e os pesos dos elementos mais antigos decrescem exponencialmente. Assim, o peso na reputação calculada pelas mensagens mais recentes é muito mais significativo que a reputação obtida pelas mensagens

mais antigas, permitindo uma mudança mais rápida do valor da média, caso aconteça uma mudança de comportamento. A média móvel exponencial do valor da reputação R é atualizada segundo a equação

$$R = \frac{2}{Q+1} C_{aux} + \left(1 - \frac{2}{Q+1}\right) R, \tag{3.4}$$

onde C_{aux} é igual a -1 ou 1 dependendo da classificação do mecanismo auxiliar e o parâmetro Q é o período da média que representa o número de amostras mais significativas[1]. No cálculo da reputação também pode ser levada em conta a opinião do usuário. Caso uma mensagem seja classificada incorretamente, o usuário pode indicar que a mensagem foi classificada incorretamente e o valor da reputação do pseudônimo e do servidor que enviou a mensagem é atualizada. O valor de C_{aux} poderia ser diferente para as mensagens legítimas e para as mensagens *spams* como, por exemplo, $0,5$ para as mensagens legítimas e -1 para os *spams*. Nesse caso o peso de uma mensagem classificada como *spam* seria o dobro de uma mensagem classificada com legítima. No entanto, essa assimetria nos pesos pode fazer com que os efeitos dos falsos positivos diminuam muito o valor da reputação. Além disso, nesse caso a recuperação de um servidor que foi invadido e que teve o problema selecionado seria mais lenta, já que as mensagens legítimas teriam um peso menor.

A classificação do mecanismo auxiliar da mensagem ser legítima ou *spam* é utilizada apenas para atualizar a média usada no cálculo da reputação do servidor, a decisão final de classificar a mensagem como legítima ou *spam* é baseada na reputação calculada pelo mecanismo proposto. Como os valores utilizados para atualizar a média são sempre 1 e -1, a média sempre estará no intervalo $[-1, 1]$. Quanto mais perto de -1 maior é o número de *spams* enviados pelo servidor e, quanto mais perto de 1, maior é o número de mensagens legítimas enviadas e, neste caso, melhor a reputação do servidor.

Todo servidor também mantém uma informação de reputação de cada pseudônimo que já enviou mensagens para o servidor. A reputação dos pseudônimos é calculada da mesma forma que a reputação dos servidores, só que leva em conta apenas as mensagens enviadas pelo pseudônimo.

O mecanismo proposto utiliza duas tabelas para armazenar as reputações dos pseudô-

[1] Pode-se demonstrar que os Q últimos valores representam 86% dos pesos no cálculo da média.

nimos e dos servidores que são mostradas nas Tabelas 3.2(a) e 3.2(b), com o tamanho em bits de cada campo. A identificação dos pseudônimos é através da sua chave pública. Já os servidores são identificados através do endereço IP. Um servidor malicioso tentando falsificar o seu endereço IP não terá sucesso, uma vez que o protocolo SMTP utiliza o protocolo TCP, onde não é possível estabelecer uma conexão caso o iniciador da conexão utilize um IP falso. A estampa de tempo é utilizada para remover entradas que não foram atualizadas por um longo período. O limiar para remover as entradas antigas pode ser escolhido de acordo com a capacidade de armazenamento do servidor. O tamanho total das tabelas é dado por $576N_P + 96N_S$, onde N_P e N_S são o número de entradas na tabela de pseudônimos e servidores, respectivamente. Para armazenar a reputação de 10 milhões de usuários e 10 milhões de servidores são necessários apenas 840 Megabytes, que não é uma capacidade de armazenamento alta para um servidor. Em [49, 50] são realizadas análises do tráfego gerado por mensagens *spams* e o número de domínios de remetentes distintos é 27.338. Mesmo considerando que cada domínio possui diversos usuários e diversos servidores, ainda assim, o número de entradas nas tabelas de reputação de pseudônimos e servidores não é grande.

Tabela 3.2: Tabela de reputação de pseudônimos e servidores.

(a) Reputação dos pseudônimos.

Chave pública do pseudônimo	Reputação	Estampa de tempo
512 bits	32 bits	32 bits

(b) Reputação dos servidores.

IP do servidor	Reputação	Estampa de tempo
32 bits	32 bits	32 bits

Para avaliar a reputação de um pseudônimo, o servidor consulta outros servidores que informam a reputação do pseudônimo observada por eles. A informação de cada servidor sobre a reputação do pseudônimo é multiplicada pela reputação do servidor que foi observada localmente. A única exceção é quando ambas as reputações, do servidor e do pseudônimo, são negativas. Nesse caso, a multiplicação desses dois valores resulta em um valor positivo, tendo que ser multiplicado por -1 para resultar novamente em um valor negativo. Assume-se que um servidor malicioso não é utilizado por usuários

legítimos para enviar mensagens legítimas, o que fará com que o servidor tenha uma baixa reputação. Dessa forma, um servidor malicioso, que responda que todos os pseudônimos possuem reputação máxima não é considerado, pois a sua reputação é baixa.

Na troca de informações, não é viável consultar todos os servidores da Internet devido ao grande número de mensagens que seriam necessárias. Assim, apenas um conjunto pequeno de servidores é consultado. O número de servidores consultados é definido pelo parâmetro N_c. A estratégia para escolher os servidores que são consultados pode variar. Podem ser escolhidos, por exemplo, os servidores com maior reputação ou então os últimos servidores que enviaram mensagens. Uma característica importante do mecanismo proposto é que a adoção pode ser incremental, o que é fundamental na Internet. Mesmo que apenas alguns servidores adotem o mecanismo, já é possível a troca de informações entre eles.

A Figura 3.4 mostra esquematicamente os servidores consultados e os valores de reputação dos servidores e da reputação do pseudônimo observada por cada um dos servidores. Após consultar os servidores, o servidor i calcula a reputação final R_f do pseudônimo j através da Equação 3.5, onde V_c é o conjunto de servidores consultados, $R_S(a, b)$ representa a reputação que o servidor a observou do servidor b, $R_P(c, d)$ representa a reputação que o servidor c observou do pseudônimo d e a função $f_a(x, y)$ é definida de acordo com a Equação 3.6. A reputação é calculada dessa forma para considerar tanto a reputação do pseudônimo quanto a reputação dos servidores que são consultados. Caso nenhum dos servidores consultados possua informações sobre o pseudônimo, a reputação será zero.

$$R_f(j) = R_P(i, j) + \sum_{l \in V_c} R_S(i, l) \cdot R_P(l, j) \cdot f_a(R_S(i, l), R_P(l, j)) \tag{3.5}$$

$$f_a(x, y) = \begin{cases} -1 & \text{se x<0 e y<0} \\ 1 & \text{caso contrário} \end{cases} \tag{3.6}$$

Caso o remetente envie a mensagem sem utilizar um pseudônimo, o valor da reputação do remetente da mensagem é calculado apenas com base na reputação do servidor do remetente que enviou a mensagem. A Figura 3.5 mostra de forma esquemática os valores de reputação que são considerados nesse caso. De forma similar são consultados N_c

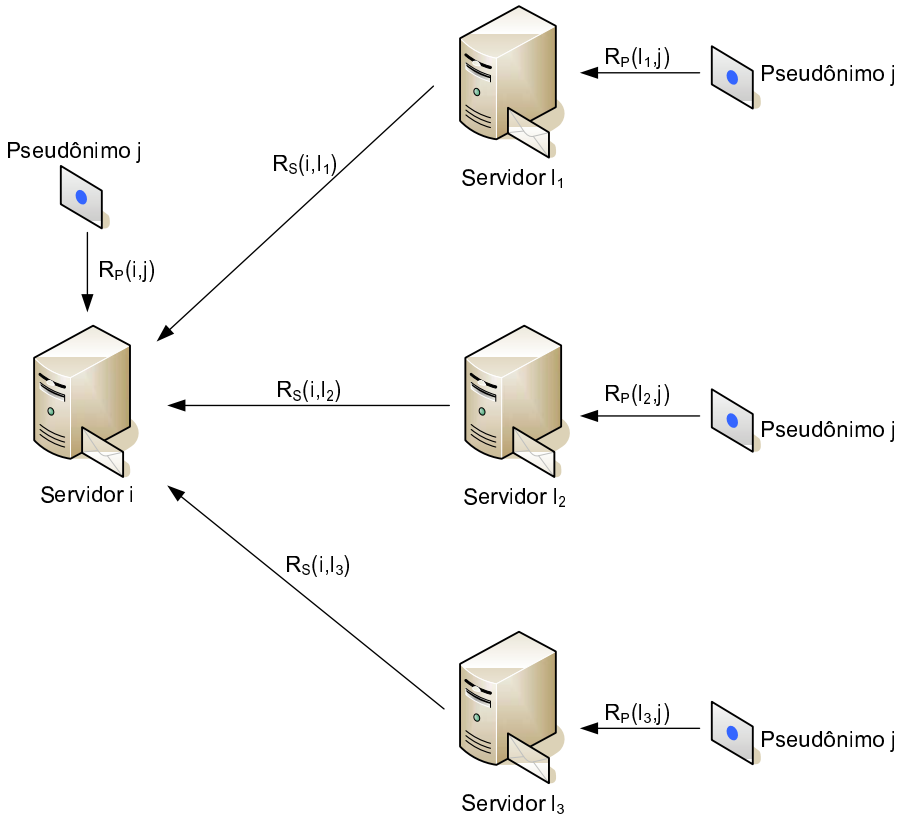


Figura 3.4: Esquema do mecanismo de confiança com o uso de pseudônimos.

servidores e o servidor i e a reputação final do servidor j, $R_f(j)$, é calculada por

$$R_f(j) = max\left(0, R_S(i,j) + \sum_{l \in V_c} R_S(i,l) \cdot R_S(l,j) \cdot f_a(R_S(i,l), R_S(l,j))\right), \quad (3.7)$$

onde V_c é o conjunto de servidores consultados, $R_S(i,j)$ representa a reputação que o servidor i observou do servidor j. Devido à utilização da função máximo com um dos argumentos igual a zero, o valor máximo da reputação considerada nesse caso é zero. Esse comportamento é adotado para que um *spammer*, que utilize um servidor legítimo com boa reputação e não se autentique utilizando pseudônimos, não possa usufruir da reputação do servidor legítimo, que seria maior do que zero. Por outro lado, um *spammer* que utilize servidores com baixa reputação sem se autenticar utilizando pseudônimos terá um valor de reputação negativa.

Uma medida que pode ser adotada pelos *spammers* para contornar o mecanismo anti-*spam* proposto é o ataque *Sybil* [51], onde o *spammer* troca constantemente a identidade utilizada para se passar por vários usuários. Um *spammer* pode utilizar um pseudônimo

distinto para cada mensagem que enviar. Nesse caso, ele não conseguirá obter um valor de reputação positivo, mas também não possuirá um valor de reputação negativo, uma vez que a reputação dos novos pseudônimos inicia em zero. Dessa forma, esse segundo esquema de autenticação, quando não são utilizados pseudônimos, somente traz benefícios se a geração de pseudônimos pelos *spammers* for computacionalmente difícil, inviabilizando o uso de um pseudônimo diferente a cada mensagem [52]. Outra medida que pode ser adotada pelos *spammers* é realizar um conluio. Nesse ataque um *spammer* cria vários usuários falsos em servidores legítimos e envia mensagens legítimas para esses endereços, fazendo com que a reputação do pseudônimo que envia as mensagens seja alta. Depois disso, o *spammer* utiliza o pseudônimo que possui uma alta reputação para enviar os *spams*. Esse ataque, no entanto, requer que sejam criados vários usuários em servidores diferentes e que várias mensagens sejam enviadas para esses usuários, tornando o ataque difícil de ser realizado.

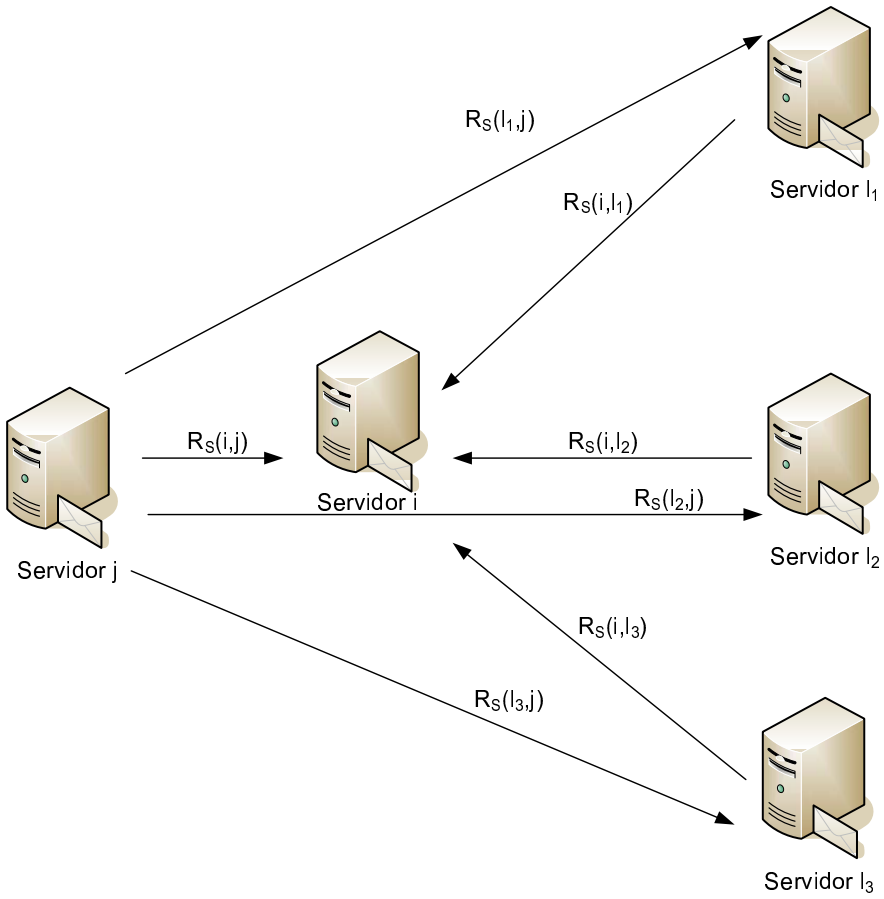


Figura 3.5: Esquema do mecanismo de confiança.

O valor $R_f(j)$, que representa a reputação do remetente da mensagem, é então utilizado para determinar se a mensagem deve ser classificada como *spam* ou não. Uma maneira de utilizar essa reputação para filtrar as mensagens é ajustando o limiar padrão (ρ) de um mecanismo anti-*spam* que se baseia em pesos e regras de acordo com o valor da reputação. Quanto maior for a reputação, maior será o limiar, reduzindo a probabilidade de ocorrência de falsos positivos. O valor do limiar utilizado é alterado para $(R_f(j)+1)\rho$. O valor um é somado à reputação, pois quando nenhum servidor observou nenhum comportamento do pseudônimo, o valor da reputação é zero. Nesse caso, é adotado o valor do limiar igual ao limiar padrão.

O mecanismo anti-*spam* auxiliar que é utilizado para determinar a reputação pode ser o mesmo mecanismo de pesos e regras só que utilizando o limiar padrão. Depois da reputação do remetente da mensagem ser calculada, avalia-se novamente se a mensagem deve ser classificada como *spam* ou não, de acordo com o novo limiar calculado com base na reputação. Dessa forma, um *spammer* que não utiliza um pseudônimo, não poderá se aproveitar da reputação dos servidores legítimos. Se o *spammer* utilizar um pseudônimo p diferente para cada mensagem, $R_P(j, p)$ será sempre zero para todos os servidores j e a reputação de acordo com a Equação 3.5 também será zero. Assim, a decisão tomada é a mesma do mecanismo auxiliar. Se fosse utilizada a reputação do servidor como acontece no TOPAS, o *spammer* conseguiria se aproveitar da reputação dos servidores legítimos.

A proposta também é robusta ao roubo de pseudônimos de usuários legítimos. Caso a máquina de um usuário seja invadida por alguma praga digital [8], o pseudônimo do usuário pode ser utilizado por *spammers*, que se aproveitam da boa reputação do pseudônimo para enviar *spams*. Porém, a reputação do pseudônimo irá diminuir com o tempo, tornando-se inútil para os *spammers* e também para o usuário legítimo, que deverá realizar a troca do pseudônimo. Para realizar essa troca de forma automática, é utilizado um mecanismo de proteção. Esse mecanismo consulta a reputação do pseudônimo no servidor para o qual o usuário está enviando uma mensagem. O servidor realiza todos os procedimentos para o cálculo da reputação do pseudônimo e retorna esse valor para o cliente. Caso o valor da reputação seja baixo, o usuário troca automaticamente seu pseudônimo. Mesmo se um *spammer* realizar um ataque de força bruta consultando vários pseudônimos até descobrir algum com boa reputação, ele não conseguirá a chave

privada do pseudônimo, tornando inútil esse tipo de ataque.

No processo de autenticação são trocadas ao todo 2 mensagens para cada destinatário. O cliente envia uma mensagem requisitando um desafio e o servidor responde enviando um desafio. Após o processo de autenticação, o mecanismo de reputação troca duas mensagens para cada servidor consultado. O servidor que realiza a consulta envia um pedido de informações sobre a reputação do pseudônimo e o outro servidor responde com a reputação observada por ele. Como são consultados N_c servidores, o número total de mensagens trocadas pelo mecanismo de reputação é $2N_c$. Antes de enviar as mensagens, o mecanismo de proteção consulta a reputação do pseudônimo do próprio usuário, utilizando mais 2 mensagens. Desta forma, o processo de consulta da reputação do pseudônimo acabará sendo realizado duas vezes e o número total de mensagens trocadas pelo mecanismo de reputação será $4N_c$. O número total de mensagens trocadas pelo mecanismo proposto por destinatário é $4N_c + 4$. O custo computacional do cálculo da reputação é pequeno, uma vez que são realizadas apenas operações simples de soma e multiplicação. Apesar de utilizar mais recursos dos servidores, o mecanismo proposto pode tornar mais precisa a detecção das mensagens, o que justifica os recursos extras que são utilizados pelo mecanismo.

Capítulo 4

Modelo do Mecanismo

DIVERSOS parâmetros influenciam o desempenho do mecanismo proposto. A taxa de falsos positivos e falsos negativos do mecanismo auxiliar, o percentual de usuários que utilizam pseudônimos e o percentual de *spammers* que utilizam servidores legítimos são alguns dos fatores que influenciam a taxa de falsos positivos e falsos negativos do mecanismo proposto. Para avaliar analiticamente a eficiência do mecanismo proposto, este capítulo propõe um modelo simplificado do mecanismo proposto e do mecanismo TOPAS. A análise matemática considera que o sistema está em estado estacionário, ou seja, todos os servidores já receberam um número de mensagens suficiente para avaliar a reputação dos outros servidores e dos pseudônimos.

O mecanismo anti-*spam* auxiliar considerado na análise é um mecanismo baseado em pesos e regras, similar ao mecanismo anti-*spam* mais utilizado na Internet, chamado *SpamAssassin* [16]. O somatório dos pesos das regras, que correspondem às características da mensagem, define o grau da mensagem ser *spam* ou não. Um limiar, ρ, é definido para determinar que as mensagens com grau acima do limiar sejam classificadas como *spam*. Neste trabalho, o grau das mensagens legítimas e *spams* é modelado de acordo com uma distribuição normal com desvio padrão σ. Desta forma, fixando-se o limiar e alterando a média dos graus das mensagens legítimas e *spams*, pode-se alterar a taxa de falsos positivos (FP) e falsos negativos (FN) do mecanismo auxiliar. Assim, variando a média do grau das mensagens legítimas (μ_l) e *spams* (μ_s), diferentes taxas de falsos positivos e falsos negativos podem ser obtidas. Na prática, a variação da média é provocada pela uti-

lização de regras com maior ou menor precisão. A probabilidade de ocorrência de falsos positivos é igual à probabilidade do grau da mensagem avaliada ser maior do que o limiar e pode ser expressa pela Equação 4.1, onde $erf(x)$ é a função erro da distribuição normal padrão definida pela Equação 4.3. De forma similar, a taxa de falsos negativos pode ser expressa pela Equação 4.2. A Figura 4.1 mostra graficamente as distribuições utilizadas para modelar o grau das mensagens legítimas e *spams*.

$$FP(\mu_l, \rho, \sigma) = \frac{1}{2} - \frac{1}{2} erf\left(\frac{\rho - \mu_l}{\sigma\sqrt{2}}\right) \tag{4.1}$$

$$FN(\mu_s, \rho, \sigma) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\rho - \mu_s}{\sigma\sqrt{2}}\right) \tag{4.2}$$

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{4.3}$$

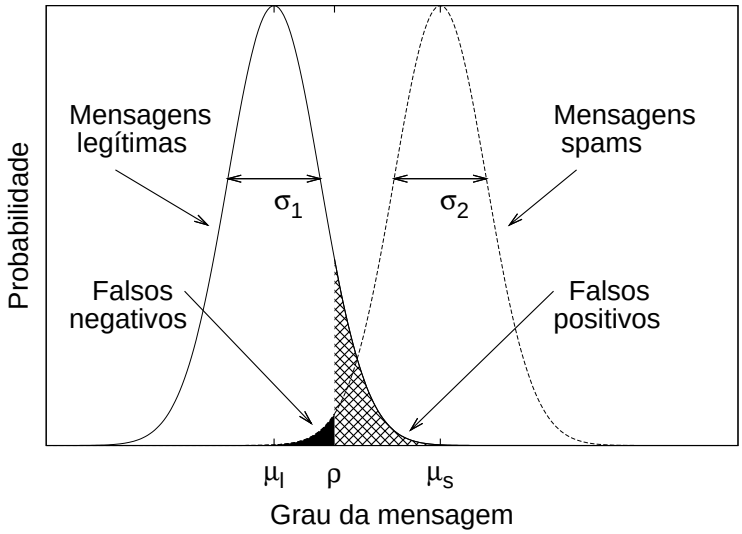


Figura 4.1: Distribuição do grau das mensagens.

Considerando as probabilidades de ocorrência de falsos positivos e falsos negativos do mecanismo auxiliar p_{fp} e p_{fn}, a probabilidade de uma mensagem enviada por um usuário com um pseudônimo P_i ser classificada como legítima pelo mecanismo auxiliar é dada por

$$p_{l_{P_i}} = (1 - p_{fp}) e_{l_{P_i}} + p_{fn} e_{s_{P_i}}, \tag{4.4}$$

onde $e_{l_{P_i}}$ é a probabilidade do usuário que possui o pseudônimo P_i enviar uma mensagem legítima e, $e_{s_{P_i}}$ é a probabilidade deste usuário enviar um *spam*. De forma análoga, a probabilidade de uma mensagem enviada pelo usuário que possui o pseudônimo P_i ser classificada como *spam* é dada por

$$p_{s_{P_i}} = p_{fp} e_{l_{P_i}} + (1 - p_{fn}) e_{s_{P_i}}. \tag{4.5}$$

Portanto, a reputação do pseudônimo P_i é dada por

$$R_{P_i} = p_{l_{P_i}} - p_{s_{P_i}}. \tag{4.6}$$

As probabilidades de um servidor S_i enviar mensagens legítimas, $e_{l_{S_i}}$, e *spams*, $e_{s_{S_i}}$, são expressas por

$$e_{l_{S_i}} = \frac{1}{N_{V_{S_i}} \sum_{k \in V_{S_i}} \lambda_k} \sum_{j \in V_{S_i}} \lambda_j e_{l_{P_j}}, \tag{4.7}$$

$$e_{s_{S_i}} = \frac{1}{N_{V_{S_i}} \sum_{k \in V_{S_i}} \lambda_k} \sum_{j \in V_{S_i}} \lambda_j e_{s_{P_j}}, \tag{4.8}$$

onde V_{S_i} é o conjunto de usuários que utilizam o servidor S_i para enviar mensagens, λ_i é a taxa de mensagens enviadas pelo usuário i e $N_{V_{S_i}}$ é o número de elementos do conjunto V_{S_i}.

As probabilidades das mensagens do servidor S_i serem classificadas como legítimas, $p_{l_{S_i}}$, e *spams*, $p_{s_{S_i}}$, são

$$p_{l_{S_i}} = (1 - p_{fp}) e_{l_{S_i}} + p_{fn} e_{s_{S_i}}, \tag{4.9}$$

$$p_{s_{S_i}} = p_{fp} e_{l_{S_i}} + (1 - p_{fn}) e_{s_{S_i}}. \tag{4.10}$$

Portanto, a reputação do servidor S_i, R_{S_i}, é então

$$R_{S_i} = p_{l_{S_i}} - p_{s_{S_i}}. \tag{4.11}$$

No mecanismo TOPAS, a reputação do servidor i é

$$R_{S_i} = p_{l_{S_i}}. \tag{4.12}$$

Para avaliar o desempenho do mecanismo proposto, é considerado que um percentual dado por p_a de usuários legítimos adotam o mecanismo proposto e um percentual de *spammers* dado por p_c utilizam contramedidas para não possuírem uma má reputação no mecanismo proposto, utilizando um pseudônimo diferente a cada mensagem.

Para avaliar o impacto dos *spammers* que tentam se beneficiar da reputação dos servidores legítimos, é considerado que um percentual de *spammers* dado por p_{s_l} envia mensagens através de servidores legítimos. O número de usuários legítimos é considerado igual ao dobro do número de *spammers*, mas a taxa que os usuários legítimos enviam mensagens é considerada como quatro vezes menor do que a taxa que os *spammers* enviam *spam*. Esses valores foram escolhidos de tal forma que a proporção final das mensagens *spams* seja igual a $2/3$, que é a proporção de *spams* observada na prática [1]. Nesse caso, as Equações 4.7 e 4.8 podem ser simplificadas e resultam, respectivamente, em

$$e_{l_S} = \frac{2p_{s_l}}{1 + 2p_{s_l}}, \tag{4.13}$$

$$e_{s_S} = \frac{1}{1 + 2p_{s_l}}. \tag{4.14}$$

Como está sendo considerado que todos os servidores já receberam mensagens suficientes para avaliar os pseudônimos e os outros servidores, a reputação de um dado pseudônimo e um dado servidor é a mesma em todos os servidores. A reputação dos servidores legítimos e dos pseudônimos de usuários legítimos é dada por

$$R_{S_l} = \left((1 - p_{fp}) \cdot el_S + p_{fn} \cdot es_S\right) - \left(p_{fp} \cdot el_S + (1 - p_{fn}) \cdot es_S\right), \tag{4.15}$$

$$R_P = 1 - 2p_{fp}. \tag{4.16}$$

A reputação final dos pseudônimos dos usuários legítimos é dada de acordo com a Equação 3.5. O valor final da reputação dos pseudônimos de usuários legítimos que utilizam pseudônimos é dado por

$$R_{f_{U_P}} = R_P + N_c R_{S_l} R_P f_a(R_{S_l}, R_P). \tag{4.17}$$

Como somente um percentual p_a de usuários utilizam pseudônimos, para os outros usuários a reputação considerada será apenas a reputação do servidor que enviou a mensagem. Considerando que todos os usuários legítimos enviam mensagens através de servidores legítimos, a reputação nesse caso é dada por

$$R_{f_{U_{\bar{P}}}} = max(0, R_{S_l} + N_c R_{S_l}^2 f_a(R_{S_l}, R_{S_l})). \tag{4.18}$$

A probabilidade da ocorrência de falsos positivos para os usuários legítimos que utilizam pseudônimos é dada pela Equação 4.1. Para os usuários que utilizam pseudônimos, o valor do limiar de classificação das mensagens é dado por

$$\rho_{U_P} = (R_{f_{U_P}} + 1)\rho_p. \tag{4.19}$$

Já para os usuários que não utilizam pseudônimos, o limiar utilizado é

$$\rho_{U_{\bar{P}}} = (R_{f_{U_{\bar{P}}}} + 1)\rho_p. \tag{4.20}$$

Assim, a taxa final de falsos positivos é dada por

$$FP(\mu_l, \rho_{U_P}, \sigma)p_a + FP(\mu_l, \rho_{U_{\bar{P}}}, \sigma)(1 - p_a). \tag{4.21}$$

Os *spammers* podem utilizar servidores legítimos ou servidores que enviam apenas *spam* e também podem utilizar ou não contramedidas como a troca do pseudônimo a cada vez que a mensagem é enviada. Sempre que o *spammer* utilizar a contramedida de trocar o pseudônimo a cada mensagem, sua reputação será zero, já que nenhum servidor

consultado terá informações sobre o pseudônimo. Nesse caso o limiar adotado na classificação será o limiar padrão ρ. Caso o *spammer* utilize um servidor legítimo para enviar as mensagens, porém não utilize um pseudônimo, a reputação será dada de acordo com

$$R_{f_{S_l}} = max(0, R_{S_l} + N_c R_{S_l}^2 f_a(R_{S_l}, R_{S_l})), \tag{4.22}$$

e o limiar adotado será dado por

$$\rho_{_{S_l}} = (R_{f_{S_l}} + 1)\rho_p. \tag{4.23}$$

A reputação dos servidores que só enviam *spam* é dada por

$$R_{S_s} = 2p_{fn} - 1, \tag{4.24}$$

e os *spammers* que utilizam esses servidores e não utilizam contramedidas têm a reputação dada por

$$R_{f_{S_s}} = max(0, R_{S_s} + N_c R_{S_s}^2 f_a(R_{S_s}, R_{S_s})). \tag{4.25}$$

Nesse caso, o limiar adotado será dado por

$$\rho_{_{S_s}} = (R_{f_{S_s}} + 1)\rho_p. \tag{4.26}$$

Considerando as técnicas que os *spammers* podem utilizar para enviar as mensagens, a taxa final de falsos negativos para o mecanismo proposto será dada por

$$FN(\mu_s, \rho, \sigma)p_c + \Big(FN(\mu_s, \rho_{_{S_l}}, \sigma)p_{sl} + FN(\mu_s, \rho_{_{S_s}}, \sigma)(1 - p_{sl})\Big)\,(1 - p_c). \tag{4.27}$$

O mecanismo proposto foi comparado com o TOPAS. No TOPAS é considerada apenas a reputação do servidor que envia as mensagens. A reputação do servidor legítimo é dada por

$$R_{S_l} = (1 - p_{fp})el_{_S} + p_{fn} \cdot es_{_S}. \tag{4.28}$$

Assim, o limiar para o mecanismo TOPAS é dado por

$$\rho_{_{S_l}} = 10\left((1 - p_{fp}) \cdot el_{_S} + p_{fn} \cdot es_{_S}\right), \tag{4.29}$$

de acordo com a descrição feita na Seção 3.1. A taxa de falsos positivos para o mecanismo TOPAS é dada por

$$FP(\mu_l, \rho_{_{S_l}}, \sigma). \tag{4.30}$$

Para o mecanismo TOPAS, o limiar adotado para os *spams* enviados através de servidores legítimos é dado pela Equação 4.29. Quando as mensagens são enviadas através de servidores que enviam apenas *spam*, a reputação desses servidores é dada por

$$R_{S_s} = p_{fn}. \tag{4.31}$$

O limiar adotado nesse caso é dado por

$$\rho_{_{S_s}} = 10p_{fn}. \tag{4.32}$$

A taxa de falsos negativos do mecanismo TOPAS depende se o *spam* é enviado através de um servidor legítimo ou através de um servidor que envia apenas *spams*. Portanto, se a mensagem é enviada através de um servidor legítimo, a taxa de falsos negativos é dada pela Equação 4.2, utilizando o limiar da Equação 4.29. Caso o *spam* seja enviado através de um servidor que envia apenas *spam*, a taxa de falsos negativos é dada pela Equação 4.2, utilizando o limiar da Equação 4.32. A taxa final de falsos negativos é dada por

$$FN(\mu_s, \rho_{_{S_l}}, \sigma)p_{sl} + FN(\mu_s, \rho_{_{S_s}}, \sigma)(1 - p_{sl}). \tag{4.33}$$

4.1 Resultados Analíticos

Diversos parâmetros influenciam o desempenho do mecanismo proposto. A taxa de falsos positivos e falsos negativos do mecanismo auxiliar, o percentual de usuários que utilizam pseudônimos e o percentual de *spammers* que utilizam servidores legítimos são alguns dos fatores que influenciam a taxa de falsos positivos e falsos negativos do mecanismo proposto. Nesta seção, o desempenho do mecanismo proposto é avaliado e comparado com outro mecanismo proposto na literatura chamado TOPAS.

O mecanismo auxiliar usado na avaliação é o mesmo descrito na Seção 4. O valor do limiar ρ utilizado para a decisão da classificação da mensagem como legítima ou *spam* é igual a 5, que é o limiar padrão do mecanismo *SpamAssassin*, e o desvio padrão σ da distribuição normal utilizado é igual a 4. Considera-se ainda na avaliação que são consultados três servidores para determinar a reputação do pseudônimo.

A Figura 4.2 mostra a variação da reputação dos servidores legítimos utilizando o mecanismo proposto e o mecanismo TOPAS e também a reputação dos pseudônimos de usuários legítimos em função do percentual de *spammers* que utilizam servidores legítimos para enviar as mensagens. Nesta análise foi considerado que a taxa de falsos positivos e falsos negativos do mecanismo auxiliar é de 10%. A reputação dos pseudônimos não se altera, uma vez que ela é influenciada apenas pelas mensagens enviadas pelo usuário que possui o pseudônimo. O valor da reputação dos pseudônimos é igual a 0,8 devido aos falsos positivos do mecanismo auxiliar, que acabam por fazer com que a reputação mesmo dos usuários que enviam apenas mensagens legítimas não seja igual a 1.Já a reputação dos servidores legítimos no mecanismo proposto diminui com o aumento da utilização de servidores legítimos pelos *spammers*, já que mais mensagens *spam* são enviadas através dos servidores legítimos. Para o mecanismo TOPAS a reputação do servidor também diminui devido ao mesmo motivo. A reputação neutra do mecanismo proposto é zero, enquanto que a reputação neutra no mecanismo TOPAS é 0,5. Para os dois mecanismos, a reputação do servidor legítimo fica menor do que o valor da reputação neutra quando mais de 50% dos *spammers* utilizam os servidores legítimos para enviar *spams*, ou seja, quando o número de mensagens legítimas enviadas é igual ao número de *spams*.

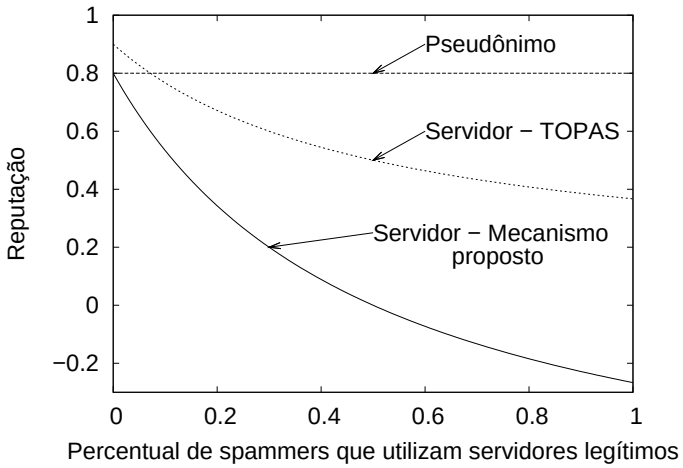


Figura 4.2: Reputação em função do percentual de utilização de servidores legítimos por *spammers*.

A Figura 4.3 mostra o valor do limiar que é utilizado na classificação das mensagens dos usuários legítimos que utilizam pseudônimos e dos usuários que não utilizam pseudônimos para o mecanismo proposto. Deve ser ressaltado que os usuários que não usam pseudônimos não podem ser autenticados pelo mecanismo proposto e, portanto, a sua reputação máxima de ser deve ser a reputação neutra de valor zero. Para os usuários que não utilizam pseudônimos o limiar inicia no valor do limiar padrão, uma vez que a reputação máxima para os usuários legítimos sem pseudônimos é zero. O limiar só passa a ser menor do que limiar padrão quando a reputação do servidor é menor do que zero. De acordo com a Figura 4.2 a reputação do servidor legítimo passa a ser menor do que zero quando mais de 50% dos *spammers* utilizam servidores legítimos. Para os usuários legítimos que utilizam pseudônimos, o limiar decresce devido à diminuição da reputação dos servidores legítimos. Mesmo se os servidores legítimos reportarem uma reputação alta para o pseudônimo, a reputação final será baixa, pois a reputação do pseudônimo informada pelo servidor é multiplicada pela reputação desse servidor, que diminui com o aumento da utilização de servidores legítimos por *spammers*. Mesmo assim, o valor da reputação dos usuários também leva em conta a reputação observada localmente pelo servidor, que continua alta. Dessa forma, mesmo com todos os *spammers* utilizando os servidores legítimos, o limiar adotado é sempre maior do que o limiar padrão, garantindo a eficácia do mecanismo proposto uma vez que a taxa de falsos positivos será menor que a do mecanismo auxiliar.

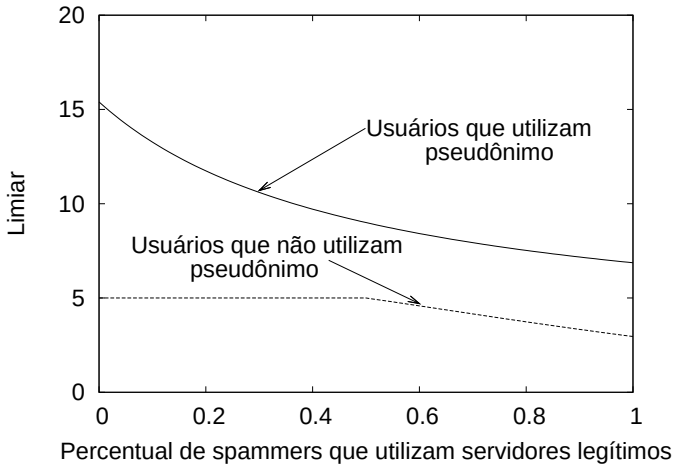


Figura 4.3: Limiar utilizado pelo mecanismo proposto para os usuários legítimos.

A Figura 4.4 mostra o valor do limiar que é utilizado pelo mecanismo proposto na classificação dos *spams* enviados através de servidores legítimos e servidores que enviam apenas *spam*. Para estes resultados apresentados, assume-se que os *spammers* não adotam a contramedida de enviar as mensagens com pseudônimos diferentes. Como não estão sendo utilizados pseudônimos, a reputação máxima que os *spammers* podem alcançar é zero, de acordo com a Equação 3.7. Para os *spammers* que utilizam servidores legítimos, sempre que a reputação do servidor legítimo for maior do que zero, a reputação para o *spammer* será zero e é adotado o limiar padrão. Quando a reputação do servidor é menor do que zero, os *spammers* começam a se prejudicar, já que a reputação passa a ser negativa e o limiar passa a ser menor do que o limiar padrão, o que significa um maior rigor na classificação das mensagens como *spam*. Já para os *spammers* que não utilizam servidores legítimos, a reputação considerada é a reputação do servidor que envia *spam* somada à reputação informada por outros servidores legítimos que são consultados. A reputação do servidor que envia *spam* é dada pela Equação 4.24 e é sempre menor do que zero o que corresponde a um maior rigor na classificação das mensagens como *spam*. No cálculo da reputação final do servidor que envia apenas *spam* são consultados N_c servidores legítimos e cada um desses servidores irá informar a reputação do servidor que envia apenas *spam* observada por eles. No cálculo da reputação final do servidor que envia apenas *spam*, as recomendações informadas pelos servidores legítimos são multiplicadas pela reputação de cada um dos servidores. No entanto, conforme mostrado na Figura 4.2, a reputação dos servidores legítimos diminui com o aumento do percentual de *spammers*

que utilizam servidores legítimos. Com essa diminuição da reputação dos servidores legítimos, a informação sobre a reputação do servidor que envia apenas *spam* que é informada pelos servidores legítimos passa a ter menos importância. Assim, a informação da reputação negativa do servidor que envia apenas *spam* informada pelos servidores legítimos, passa a ter menos peso ou importância e o valor da reputação final acaba aumentando, causando um aumento do limiar que é observado na Figura 4.4 quando o percentual de *spammers* que utilizam servidores legítimos está entre 0 e 50%. Acima de 50%, conforme observado na Figura 4.2 a reputação dos servidores legítimos passa a ser menor do que zero, fazendo com que a informação da reputação negativa informada pelos servidores legítimos passe a ser considerada novamente. Como conseqüência, a reputação final dos servidores que enviam apenas *spam* passa a reduzir novamente e com isso o limiar também diminui. É importante ressaltar que o limiar utilizado para *spammers* é sempre menor que 5 (cinco) que corresponde ao limiar usado pelo mecanismo auxiliar. Isto significa um maior rigor aplicado na classificação como *spam* às mensagens enviadas pelos *spammers*.

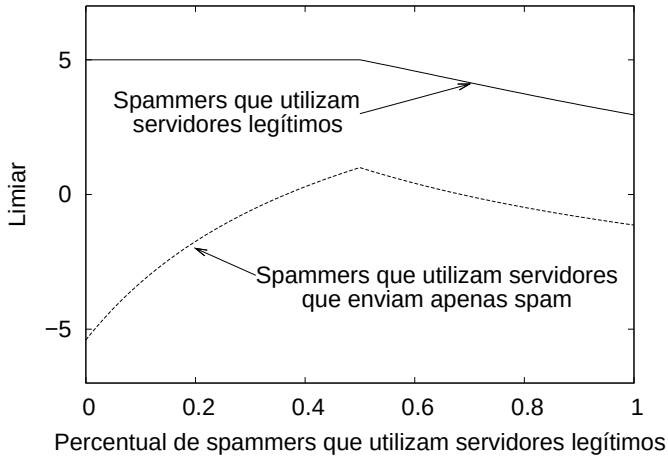


Figura 4.4: Limiar utilizado pelo mecanismo proposto para os *spammers*.

A Figura 4.5 mostra o valor do limiar que é utilizado pelo mecanismo TOPAS quando uma mensagem é enviada através de um servidor legítimo ou através de um servidor que envia apenas *spam*. Quando a mensagem é enviada através de servidores que enviam apenas *spam* o limiar adotado é sempre um, que é menor do que o limiar padrão do mecanismo auxiliar, reduzindo a taxa de falsos negativos para os *spams* que são enviados através de servidores que enviam apenas *spam*. Já para as mensagens enviadas através de servidores legítimos, o limiar decresce da mesma forma que foi observado na Figura 4.2.

A diferença do mecanismo TOPAS em relação ao mecanismo proposto é que os *spams* enviados através de servidores legítimos são sempre classificados com um limiar acima do limiar padrão, o que corresponde a uma taxa de falsos negativos para o mecanismo TOPAS maior que a obtida pelo mecanismo proposto.

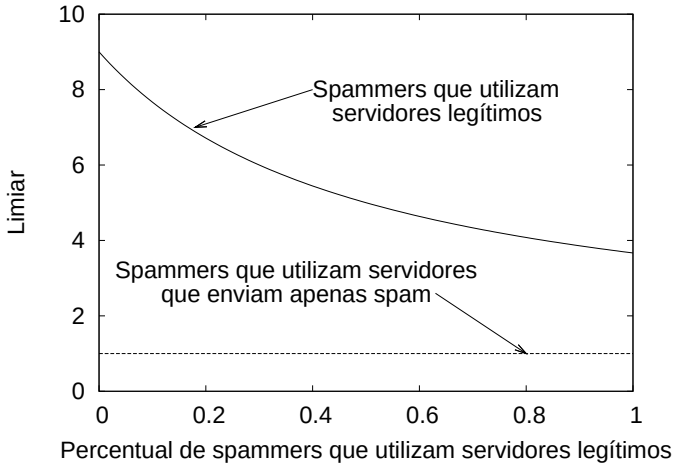


Figura 4.5: Limiar utilizado no mecanismo TOPAS.

A Figura 4.6 mostra a relação entre os falsos positivos e o percentual de *spammers* que utiliza servidores legítimos para o mecanismo proposto e o mecanismo TOPAS. Para o mecanismo TOPAS o resultado é o mesmo com ou sem a utilização de pseudônimos, já que ele não utiliza pseudônimos. A curva relativa ao mecanismo proposto sem a utilização de pseudônimos corresponde a utilização apenas da informação sobre a reputação dos servidores que, no máximo, tem o valor zero. Assim, quando nenhum dos usuários utiliza pseudônimos, a taxa de falsos positivos é inicialmente igual a 10%, já que é adotado o limiar padrão e a taxa de falsos positivos do mecanismo auxiliar é de 10%. Com o aumento da utilização de servidores legítimos por *spammers*, a taxa de falsos positivos aumenta, já que a reputação dos servidores legítimos passa a ser menor do que zero e o limiar diminui, conforme observado nas Figuras 4.3 e 4.2. Nessa situação tanto o mecanismo proposto quanto o mecanismo TOPAS possuem uma taxa de falsos positivos maior do que 10%, que é a taxa de falsos positivos do mecanismo auxiliar, tornando-se ineficientes. Por outro lado, quando todos os usuários utilizam pseudônimos, a eficácia do mecanismo proposto em relação aos falsos positivos é evidenciada. Neste caso, a taxa de falsos positivos do mecanismo proposto é sempre menor do que a taxa de falsos positivos do mecanismo auxiliar já que, conforme mostra a Figura 4.3, o limiar adotado é sempre

maior do que o limiar padrão. É importante ressaltar que no mecanismo TOPAS, quando o percentual de *spammers* que utilizam servidores legítimos é maior do que 50%, a taxa de falsos positivos fica pior do que a do mecanismo auxiliar que é igual a 10%. O mesmo não ocorre para o mecanismo proposto quando se usa pseudônimos uma vez que a taxa de falsos positivos é sempre muito menor que a do mecanismo auxiliar.

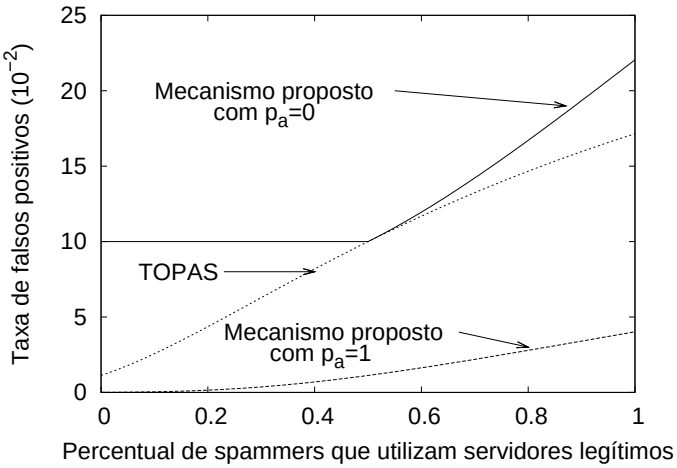


Figura 4.6: Taxa de falsos positivos em função da utilização de servidores legítimos por *spammers* para o mecanismo TOPAS e para o mecanismo proposto com ($p_a = 1$) e sem ($p_a = 0$) pseudônimos.

.

A Figura 4.7 mostra a relação entre os falsos negativos e o percentual de *spammers* que utilizam servidores legítimos para o mecanismo proposto e o mecanismo TOPAS. Quando os *spammers* não utilizam um pseudônimo diferente em cada mensagem ($p_c = 0$), a taxa de falsos negativos inicialmente aumenta e depois diminui, devido ao comportamento do limiar que foi observado na Figura 4.4. Comparado com o mecanismo TOPAS, o mecanismo proposto é sempre mais eficiente nesse caso, possuindo uma taxa de falsos negativos menor. Já quando todos os *spammers* adotam a contramedida de enviar cada mensagem com um pseudônimo diferente ($p_c = 1$), a reputação desse pseudônimo será sempre zero e, dessa forma, será sempre adotado o limiar padrão. Ao adotar o limiar padrão na classificação das mensagens, a taxa de falsos negativos acabará sempre sendo igual a 10%, que é a taxa de falsos negativos do mecanismo auxiliar. Nesse caso o mecanismo TOPAS possui uma taxa de falsos negativos menor, embora a taxa de falsos positivos seja maior que a do mecanismo proposto.

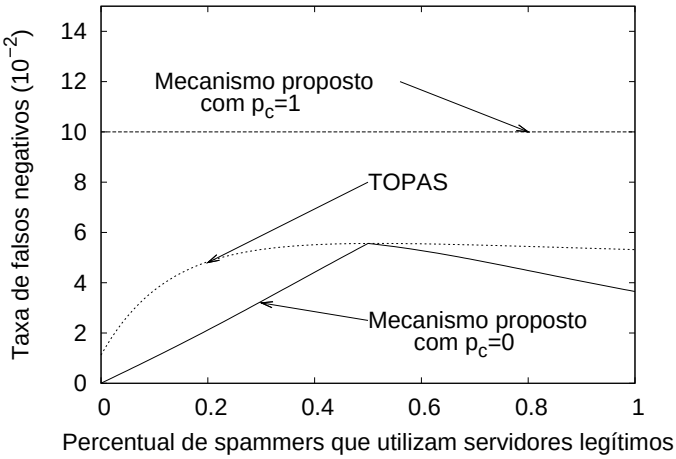


Figura 4.7: Taxa de falsos negativos em função da utilização de servidores legítimos por *spammers* para o mecanismo TOPAS e para o mecanismo proposto com ($p_c = 1$) e sem ($p_c = 0$) a utilização de contramedidas.

.

Para mostrar a influência da utilização de servidores legítimos por *spammers* e do percentual de usuários que utilizam pseudônimos, a Figura 4.8 mostra a relação entre esses dois percentuais e a taxa de falsos positivos. Os resultados confirmam que, quanto maior o percentual de usuários que utilizam pseudônimos, menor é a influência da utilização de servidores legítimos por *spammers*, uma vez que o valor do limiar adotado para os usuários que utilizam pseudônimos é sempre maior do que o adotado para os usuários que não utilizam pseudônimos, conforme indicado na Figura 4.3. Quando todos os usuários utilizam pseudônimos, a taxa de falsos positivos aumenta de 0,0051% para apenas 4,02% quando todos os *spammers* utilizam servidores legítimos, que ainda é menor do que a taxa de falsos positivos do mecanismo auxiliar. Quando 30% dos usuários legítimos utilizam pseudônimos a taxa global de falsos positivo já é reduzida de 10% para 7%. Já quando o percentual de usuários legítimos que utilizam é 70%, a taxa de falsos positivos passa a ser de 3%.

A Figura 4.9 mostra a influência do percentual de *spammers* que utilizam servidores legítimos e o percentual de *spammers* que adotam a contramedida de enviar cada mensagem com um pseudônimo diferente na taxa de falsos negativos. A estratégia mais eficaz para os *spammers*, que corresponde ao pior caso para o mecanismo proposto, é quando

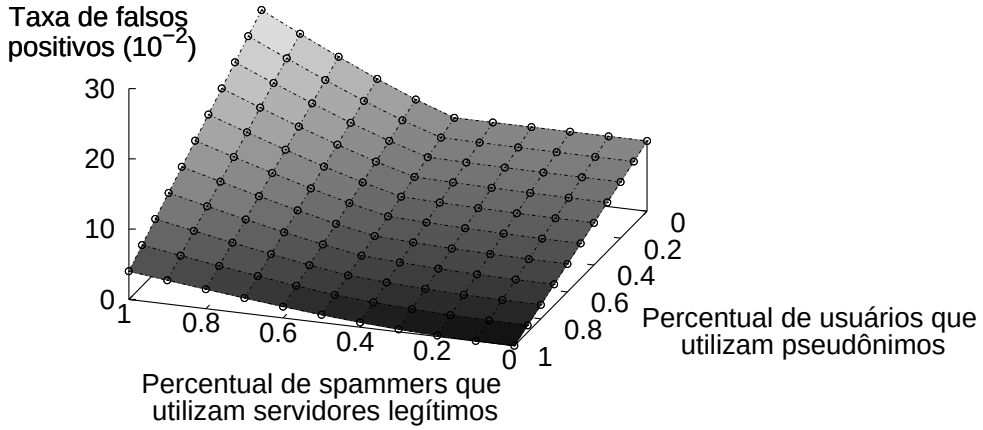


Figura 4.8: Influência da utilização de servidores legítimos por *spammers* e do percentual de usuários que utilizam pseudônimos na taxa de falsos positivos.

o *spammer* envia cada mensagem com um pseudônimo diferente. Para o *spammer* esta estratégia é mais eficaz que utilizar servidores legítimos para enviar a mensagem. Ao utilizar pseudônimos diferentes para cada mensagem, a taxa de falsos positivos é sempre igual a 10%, que corresponde à adoção do limiar padrão. Já a utilização de servidores legítimos não tem a mesma eficácia, não sendo vantajosa para os *spammers*. É também importante ressaltar que embora não haja redução na taxa de falsos negativos do mecanismo auxiliar, o mecanismo proposto acaba tornando menos vantajosa a utilização de servidores legítimos por *spammers*. Utilizando os servidores legítimos, a taxa máxima de falsos negativos que os *spammers* conseguem atingir é de 5,56%. Já utilizando a contramedida de usar um pseudônimo diferente para cada mensagem, a taxa de falsos negativos é de 10%. Essas características têm como conseqüência a menor utilização de recursos de servidores legítimos, que pode resultar em benefícios econômicos e também a redução da taxa de falsos negativos, conforme mostra a Figura 4.8, já que o percentual de *spammers* que utilizam servidores legítimos tende a ser pequeno.

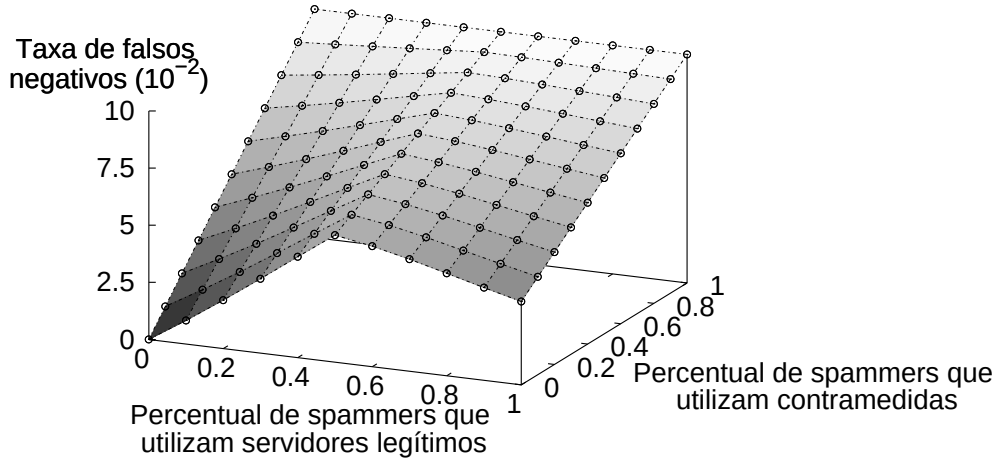


Figura 4.9: Influência do percentual de *spammers* que utilizam servidores legítimos e o percentual de *spammers* que adotam contramedidas na taxa de falsos negativos.

Capítulo 5

Avaliação de Desempenho por Simulação

PARA avaliar o mecanismo proposto foi desenvolvido um simulador de eventos discretos na linguagem C++. O simulador foi desenvolvido uma vez que outros simuladores de rede já existentes como o NS-2 [53] simulam redes com um nível de detalhamento que não é relevante para avaliar o mecanismo proposto. No NS-2, por exemplo, o tempo de transmissão dos pacotes, o controle de acesso ao meio e várias outras características das redes são simuladas, acabando por aumentar o tempo de simulação e não trazendo nenhum benefício para a análise do mecanismo proposto. Foram implementados o mecanismo anti-*spam* proposto e, para realizar a comparação com mecanismos que levam em conta apenas a reputação no servidor, o mecanismo TOPAS descrito na Seção 3.1.

Nos testes foram usados 100 usuários legítimos e 50 *spammers* e os usuários legítimos enviam mensagens a uma taxa quatro vezes menor do que os *spammers*. Com isso, o percentual médio de *spams* é de $2/3$ das mensagens, que é aproximadamente o observado na prática [1]. O intervalo entre o envio de cada mensagem pelos usuários legítimos possui uma distribuição exponencial com média de 2 unidades de tempo e para os *spammers* a média é de 0,5 unidades de tempo. O tempo total de simulação é de 100.000 unidades de tempo. A distribuição e a taxa de mensagens que os usuários geram não têm um impacto grande nos resultados da simulação, uma vez que o mecanismo proposto é baseado no

percentual de mensagens legítimas e *spams* e não no volume ou taxa das mensagens. O valor das taxas de mensagem e o tempo de simulação foram escolhidos apenas de tal forma que o número de mensagens recebidas por cada servidor seja suficiente para que a reputação seja avaliada de forma mais precisa.

Nas simulações são utilizados 50 servidores de usuários legítimos e 50 servidores de usuários que enviam *spam*. Os servidores são separados em dois tipos, pois os servidores legítimos geralmente possuem medidas para evitar que *spammers* os utilizem [54–57]. Na prática, a distribuição dos usuários legítimos não é igual para cada servidor, pois poucos servidores possuem muitos usuários e muitos servidores possuem poucos usuários. Dessa forma, para modelar essa distribuição dos usuários nos servidores, foi utilizada uma distribuição Zipf. Com essa distribuição, a probabilidade de um usuário estar em um servidor i é igual a

$$\frac{1}{\sum_{k=1}^{M} \frac{1}{k^v}} \frac{1}{i^v}, \tag{5.1}$$

onde M é o número total de servidores e v é um parâmetro da distribuição. Quanto maior o valor de v, maior será a probabilidade dos usuários se concentrarem em poucos servidores. A Figura 5.1 mostra a densidade de probabilidade da distribuição Zipf para os valores do parâmetro v iguais a zero, um e dois. Dessa forma, a concentração dos usuários nos primeiros servidores será maior, simulando a distribuição dos usuários na prática. Nas simulações é utilizada uma distribuição Zipf com parâmetro $v = 1$. A maioria dos *spammers* utiliza servidores diferentes, que podem ser servidores invadidos, mal-configurados ou máquinas zumbis [12]. Devido a essa diferença, os *spammers* são distribuídos nos servidores que enviam *spam* através de uma distribuição uniforme. Como os *spammers* também utilizam servidores que possuem usuários legítimos, foi definido um parâmetro que determina o percentual de *spammers* que utilizam os servidores legítimos. Os *spammers* que utilizam os servidores legítimos são distribuídos de acordo com a mesma distribuição Zipf utilizada para a distribuição dos usuários legítimos.

A reputação local dos pseudônimos calculada pelo servidor é determinada conforme descrito na Seção 3, utilizando a média móvel exponencial com parâmetro $Q = 50$. Já para a reputação dos servidores, é utilizado o parâmetro $Q = 500$. O valor é diferente,

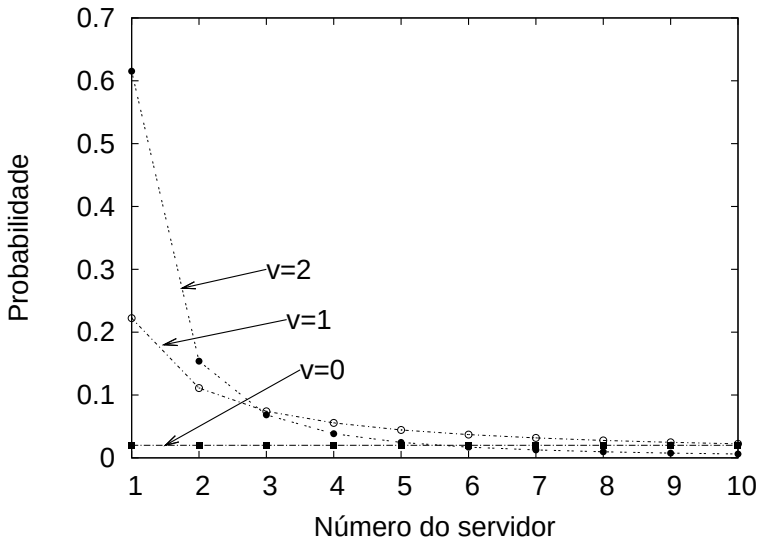


Figura 5.1: Função de densidade de probabilidade da função zipf (M=50).

pois a média do servidor é influenciada pelas mensagens enviadas por todos os usuários do servidor. Dessa forma, adota-se um período maior para a média do servidor, para considerar mais mensagens no cálculo da média. Caso um servidor legítimo seja invadido e seja utilizado para enviar *spams* a sua reputação irá reduzir rapidamente. No entanto, quando o problema for corrigido, a reputação irá subir novamente de forma rápida, devido à utilização da média móvel exponencial.

Na simulação, os *spammers* não utilizam pseudônimos próprios para enviar os *spams*, pois não possuem nenhuma vantagem em utilizá-los. No entanto, um *spammer* pode se beneficiar roubando o pseudônimo de um usuário legítimo e enviar mensagens por um curto período, se aproveitando da boa reputação do pseudônimo. O percentual de pseudônimos legítimos que são roubados pelos *spammers* é determinado por um dos parâmetros do simulador. O mecanismo de proteção dos pseudônimos é utilizado tanto pelos usuários legítimos quanto pelos *spammers*, pois não é do interesse de nenhum dos dois utilizar pseudônimos com má reputação. A cada mensagem enviada para um servidor, o mecanismo de proteção dos pseudônimos consulta a reputação do seu pseudônimo. Caso o valor da reputação seja menor ou igual a zero, o usuário muda automaticamente o seu pseudônimo, passando a utilizar um novo pseudônimo sem histórico.

O mecanismo auxiliar usado na simulação é o mesmo descrito na Seção 4. O valor

do limiar ρ utilizado é igual a 5, que é o limiar padrão do mecanismo *SpamAssassin*, e o desvio padrão σ da distribuição normal utilizado é igual a 4.

5.1 Resultados de Simulação

Nesta seção são apresentados os resultados obtidos através do simulador desenvolvido. Em todos os casos, onde não for especificado o contrário, o mecanismo auxiliar possui uma taxa de falsos positivos e falsos negativos igual a 10%. Todos os resultados das simulações são mostrados com um intervalo de confiança de 95%.

5.1.1 Testes de Sanidade

Para validar o modelo analítico e a implementação do simulador foram realizados testes de sanidade e comparados os resultados do simulador com os resultados esperados através do modelo matemático. Para validar o cálculo da reputação dos pseudônimos e dos servidores, foi definida uma métrica de erro que calcula a diferença entre a reputação calculada através do simulador e a reputação calculada através do modelo. A reputação esperada através do modelo para os pseudônimos legítimos R_P é dada pela Equação 4.16. Dessa forma, a métrica para o cálculo do erro da reputação dos usuários legítimos é calculada por

$$E_P = \sum_{j \in V_{S_l}} \sum_{i \in V_{U_l}} \left(R_P - R_P(i,j) \right), \tag{5.2}$$

onde $R_P(i,j)$ é a reputação calculada pelo servidor j do pseudônimo i no simulador, V_{S_l} é o conjunto dos servidores legítimos e V_{U_l} é o conjunto de usuários legítimos. De forma análoga, a métrica que calcula o erro da reputação dos servidores é calculada com base na reputação esperada para os servidores legítimos R_{s_l} definida na Equação 4.15 e na reputação esperada para os servidores que enviam apenas *spam*, dada por R_{S_s} definida na Equação 4.24.

A métrica do erro da reputação dos servidores é dada por

$$E_S = \sum_{j \in V_{S_l}} \left(\sum_{i \in V_{S_l}, i \neq j} (R_{S_l}(i,j) - R_{S_l}) + \sum_{i \in V_{S_s}} (R_{S_s}(i,j) - R_{S_s}) \right), \qquad (5.3)$$

onde $R_{S_l}(i,j)$ é a reputação do servidor legítimo j calculada pelo servidor i no simulador, $R_{S_s}(i,j)$ é a reputação do servidor que envia *spams* j calculada pelo servidor i no simulador, V_{S_l} é o conjunto dos servidores legítimos e V_{S_s} é o conjunto dos servidores que enviam *spam*.

A Figura 5.2 mostra o gráfico do valor das duas métricas de erro da reputação em função do tempo de simulação. Inicialmente a diferença é grande devido ao cálculo da média utilizado na reputação que inicia em zero. Dessa forma, são necessárias várias atualizações no cálculo da média até a convergência para o valor correto. O resultado deste teste de sanidade mostra que o valor das duas métricas de erro converge para zero, o que mostra que o simulador calcula a reputação conforme esperado.

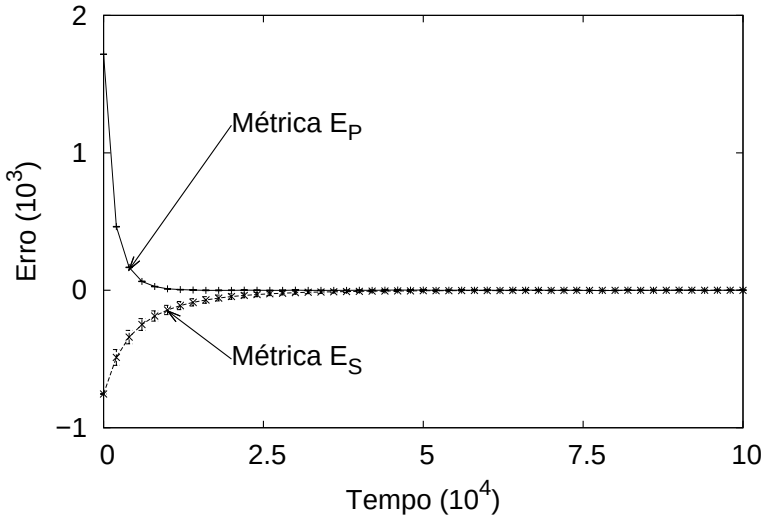


Figura 5.2: Teste de sanidade para o cálculo da reputação.

O mecanismo proposto consulta outros servidores para determinar o valor da reputação final que é utilizada para determinar o limiar. Para avaliar o processo de consulta a outros servidores, também foi calculada a métrica de erro entre o valor do limiar esperado para as mensagens enviadas pelos usuários legítimos e pelos usuários que enviam *spam*. Para os usuários legítimos, o limiar calculado é comparado com o valor do limiar ρ_{U_P} dado pela Equação 4.19. Já para os usuários que enviam *spam*, o limiar adotado é sempre

o limiar padrão, uma vez que assume-se o pior caso, onde todos os *spammers* utilizam a contramedida de enviar cada mensagem com um pseudônimo diferente. Dessa forma, as métricas de erro do limiar para os usuários legítimos e usuários que enviam *spam* são

$$E_{\rho_l} = \sum_{j \in V_{S_l}} \sum_{i \in V_{U_l}} \rho_{U_P} - \rho_{U_P}(i,j), \tag{5.4}$$

$$E_{\rho_s} = \sum_{j \in V_{S_l}} \sum_{i \in V_{S_s}} \rho_s - \rho_s(i,j), \tag{5.5}$$

onde $\rho_{U_P}(i,j)$ é o valor do limiar para o usuário do pseudônimo i calculado pelo servidor j no simulador, $\rho_s(i,j)$ é o limiar utilizado para o *spammer* que envia *spams* através do servidor i calculado pelo servidor j,V_{S_l} é o conjunto dos servidores legítimos, V_{S_l} é o conjunto dos servidores que enviam *spam*,V_{U_l} é o conjunto de usuários legítimos e V_{U_s} é o conjunto de *spammers*.

No cálculo do limiar podem ser adotadas várias estratégias para determinar os servidores que são consultados. O simulador desenvolvido implementa a estratégia de consultar os servidores com maior reputação e também a estratégia de consulta aos últimos servidores que enviaram mensagens. Nesta última estratégia, somente são considerados os últimos servidores que possuem uma reputação maior do que zero, uma vez que os servidores com reputação menor do que zero não são confiáveis e devem ser evitados. A Figura 5.3 mostra as duas métricas de erro do limiar que, conforme o esperado converge para zero, validando a implementação do mecanismo proposto realizada no simulador em relação ao modelo matemático. A métrica de erro para o limiar adotado para as mensagens enviadas por usuários legítimos converge mais rapidamente para zero utilizando a estratégia de consultar os servidores com maior reputação. Esse fato ocorre uma vez que, no cálculo da reputação estão sendo considerados os servidores com maiores reputações disponíveis, diferente da outra estratégia, que nem sempre irá selecionar os servidores com melhor reputação. A métrica de erro para o limiar adotado para as mensagens enviadas por *spammers* é sempre zero, já que os *spammers* sempre utilizam um pseudônimo diferente a cada mensagem. Como a reputação nesse caso é sempre zero, é adotado o limiar padrão e a métrica de erro é sempre zero.

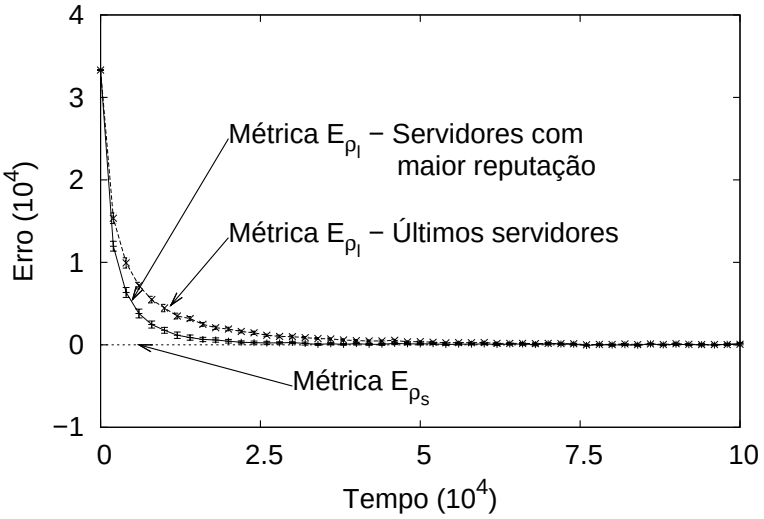


Figura 5.3: Teste de sanidade para o cálculo do limiar.

Para validar a implementação do mecanismo TOPAS foram realizados os mesmos testes de sanidade descritos anteriormente. Segundo o modelo desenvolvido, a reputação esperada no mecanismo para os servidores legítimos é dada por R_{S_l} definida na Equação 4.28 e a reputação dos servidores que enviam apenas *spam* é dada por R_{S_s} definida na Equação 4.31. A métrica que define o erro no cálculo da reputação dos servidores é

$$E_{S_t} = \sum_{j \in V_{S_l}} \left(\sum_{i \in V_{S_l}, i \neq j} (R_{S_l}(i,j) - R_{S_l}) + \sum_{i \in V_{S_s}} (R_{S_s}(i,j) - R_{S_s}) \right), \qquad (5.6)$$

onde $R_{S_l}(i,j)$ é a reputação do servidor legítimo j calculada pelo servidor i no simulador, $R_{S_s}(i,j)$ é a reputação do servidor que envia *spams* j calculada pelo servidor i no simulador, V_{S_l} é o conjunto dos servidores legítimos e V_{S_l} é o conjunto dos servidores que enviam *spam*.

A Figura 5.4 mostra a métrica de erro da reputação calculada pelo mecanismo TOPAS, que converge para zero, validando a implementação do mecanismo TOPAS. No caso do mecanismo TOPAS, a métrica de erro converge mais rapidamente para zero devido às recomendações que são trocadas entre os servidores, acelerando o processo de convergência da reputação.

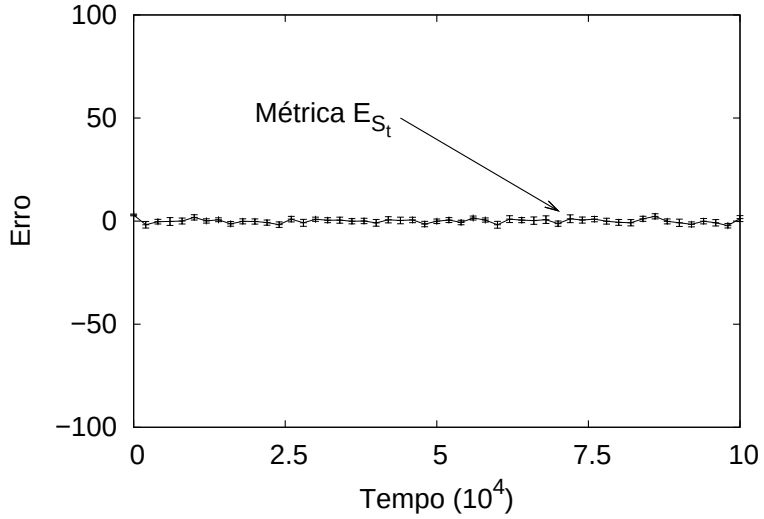


Figura 5.4: Teste de sanidade para o cálculo da reputação no mecanismo TOPAS.

5.1.2 Resultados

A Figura 5.5 mostra a taxa de falsos positivos do mecanismo proposto de acordo com o número de servidores que são consultados e de acordo com a estratégia de escolha dos servidores que são consultados. A estratégia de consulta apenas localmente possui o pior resultado, já que menos informações estão disponíveis. Nessa estratégia apenas a reputação observada localmente é levada em consideração e outros servidores não são consultados. Nessa curva, independente do parâmetro de número de servidores consultados, a consulta é apenas local. A estratégia que obteve o melhor resultado foi a estratégia de consultar os servidores com maior reputação, já que no cálculo da reputação são considerados os maiores valores de reputação dos servidores. A estratégia de consultar os últimos servidores obteve um resultado ligeiramente pior devido ao período transiente que conforme demonstrado na Seção 5.1.1 é maior para essa estratégia.

Conforme demonstrado na Seção 3, o número de mensagens trocadas pelo mecanismo de reputação aumenta linearmente com o número de consultas. No entanto, a taxa de falsos positivos não decresce linearmente com o número de consultas. Por isso, foi escolhido um valor de três consultas por vez para as simulações mostradas a seguir. Considerando a consulta a apenas três servidores e seguindo a estratégia de consultar os servidores com maior reputação, a taxa de falsos positivos diminuiu de 10%, obtida pelo mecanismo auxi-

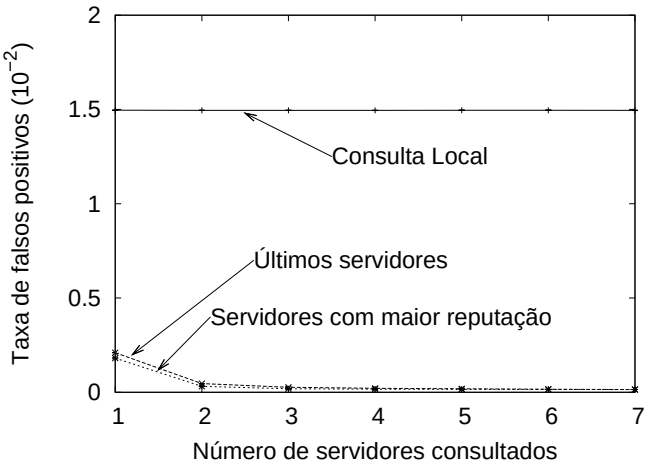


Figura 5.5: Relação entre falsos positivos e número de servidores consultados.

liar, para 0,018%. Este resultado comprova a significativa melhora de desempenho obtida pelo mecanismo proposto que torna o número de falsos positivos muito pequeno. Nas simulações do mecanismo proposto seguintes, a estratégia de consulta aos servidores com maior reputação sempre é adotada. Outra observação importante é que mesmo assumindo o pior caso, onde os *spammers* utilizam a contramedida de enviar cada mensagem com um pseudônimo diferente, a taxa de falsos negativos é sempre a mesma do mecanismo auxiliar. Assim, o mecanismo proposto no pior caso mantém os falsos negativos e diminui significativamente os falsos positivos. Caso o processo de utilizar um pseudônimo por mensagem seja tornado computacionalmente custoso para que não seja rentável para os *spammers* este procedimento, o mecanismo proposto também melhorará a taxa de falsos negativos.

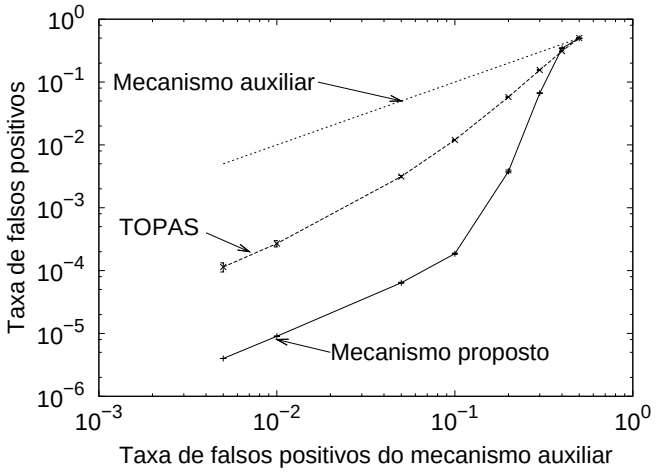
A Figura 5.6(a) mostra a taxa de falsos positivos do mecanismo proposto e do mecanismo TOPAS, em função da taxa de falsos positivos do mecanismo auxiliar. Para uma melhor visualização do desempenho a taxa de falsos positivos do mecanismo auxiliar também é plotada e o eixo y corresponde aos mesmos valores do eixo x. O mecanismo proposto possui uma taxa de falsos positivos bem menor do que a do mecanismo TOPAS, comprovando que a proposta de verificar também a reputação dos usuários é eficaz, uma vez que o mecanismo TOPAS verifica apenas a reputação dos servidores. Quando a taxa de falsos positivos do mecanismo auxiliar é de 0,5%, o mecanismo proposto reduz essa taxa para 0,0004%, ou seja, ocorre uma redução de 1.250 vezes na taxa de falsos positi-

vos. Já para a taxa de falsos positivos de 10%, a redução é de 537 vezes. Mesmo para uma taxa de falsos positivos de 20%, que é considerada alta, a redução é de 52 vezes. Para valores de falsos positivos do mecanismo auxiliar acima de 30%, a eficiência dos mecanismos TOPAS e proposto passa a ser pequena, uma vez que a avaliação do histórico de comportamento fica comprometida devido à alta taxa de falsos positivos.

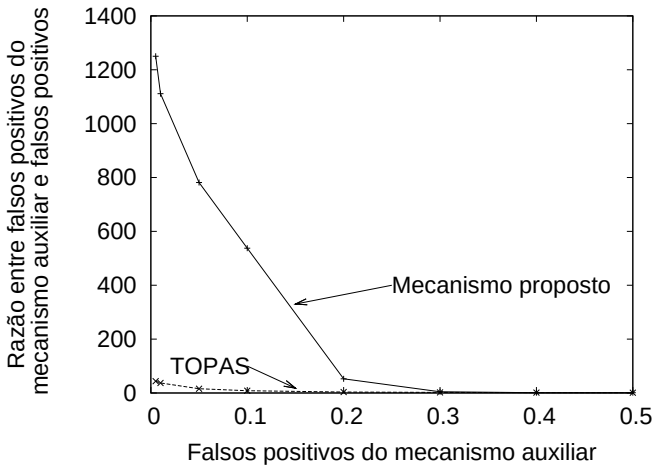
Já a Figura 5.6(b) mostra a razão entre as duas taxas. Quanto maior a razão significa que maior é a melhora na eficácia de detecção dos falsos positivos, ou seja, menor é a taxa de falsos positivos em relação à taxa de falsos positivos do mecanismo auxiliar. Portanto, o resultado da Figura 5.6(b) também mostra que o mecanismo proposto reduz substancialmente a taxa de falsos positivos e que a redução da taxa de falsos positivos no TOPAS é consideravelmente menor quando comparada ao mecanismo proposto.

A Figura 5.7 mostra a taxa de falsos negativos do mecanismo proposto e do mecanismo TOPAS, em função da taxa de falsos negativos do mecanismo auxiliar. Considerando o pior caso, onde todos os *spammers* utilizam a contramedida de enviar cada mensagem com um pseudônimo diferente, o mecanismo proposto não diminui a taxa de falsos negativos do mecanismo auxiliar. Já o TOPAS possui uma taxa de falsos negativos menor, pois considera a reputação dos servidores. Já a Figura 5.7(b) mostra a razão entre as duas taxas. Para o mecanismo TOPAS, a redução é de 8,6 vezes quando a taxa de falsos negativos do mecanismo auxiliar é de 10%.

A Figura 5.8 mostra a relação entre o percentual de pseudônimos roubados e as taxas de falsos positivos e falsos negativos. O roubo dos pseudônimos acontece na metade do tempo de simulação e os usuários legítimos continuam a usar os pseudônimos roubados até que o mecanismo de proteção troque-os. O mecanismo proposto obteve a menor taxa de falsos positivos, até mesmo quando todos os pseudônimos legítimos foram roubados. A taxa de falsos positivos do mecanismo proposto aumentou para apenas 0,74% quando todos os pseudônimos foram roubados. Já o mecanismo TOPAS teve um aumento da taxa de falsos positivos para 5,65%. Apesar do mecanismo TOPAS não utilizar pseudônimos, a taxa de falsos positivos aumenta, pois quando ocorre o roubo do pseudônimo, o *spammer* utiliza o servidor legítimo que estava sendo usado pelo usuário legítimo. Desta forma, os servidores legítimos terão sua reputação reduzida no mecanismo TOPAS. O



(a) Relação entre falsos positivos e falsos positivos do mecanismo auxiliar.

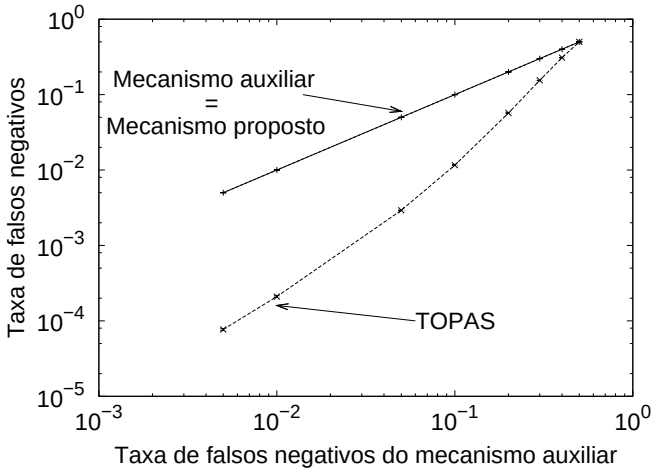


(b) Razão de falsos positivos do mecanismo auxiliar e de falsos positivos do TOPAS e do mecanismo proposto.

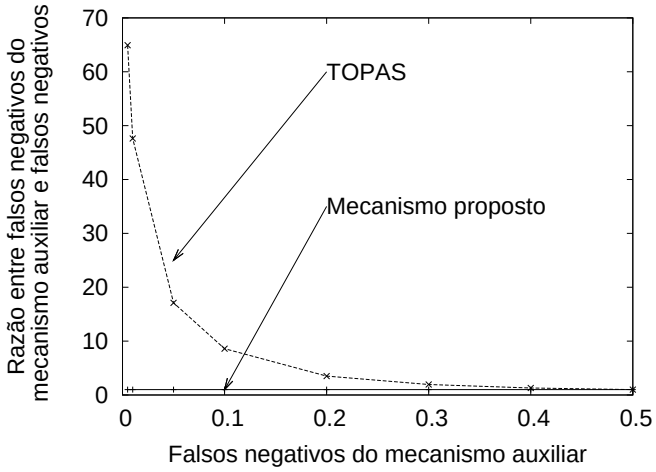
Figura 5.6: Avaliação dos falsos positivos.

mecanismo TOPAS alcançou uma taxa de falsos negativos menor do que o mecanismo proposto. No entanto, no mecanismo proposto a taxa de falsos negativos não aumenta consideravelmente, já que os *spammers* não conseguem se aproveitar da reputação dos pseudônimos roubados por um longo período.

A Figura 5.9 mostra a influência do percentual de *spammers* que utilizam os servidores legítimos nas taxas de falsos positivos e falsos negativos. O comportamento observado na simulação foi o mesmo observado através do modelo descrito na Seção 4. O mecanismo TOPAS possui uma taxa de falsos positivos maior do que o mecanismo proposto.



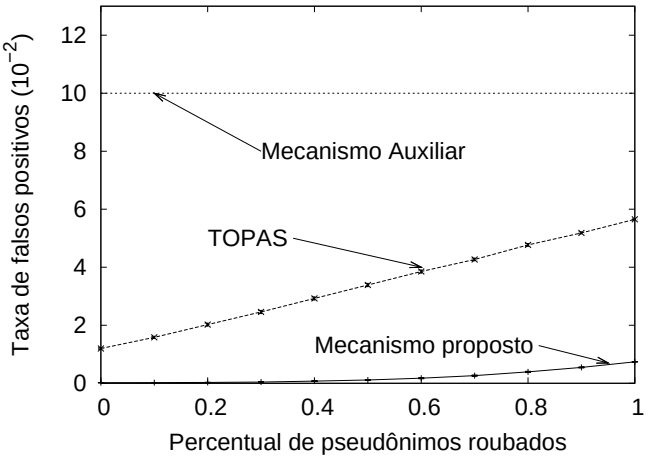
(a) Relação entre falsos negativos e falsos negativos do mecanismo auxiliar.



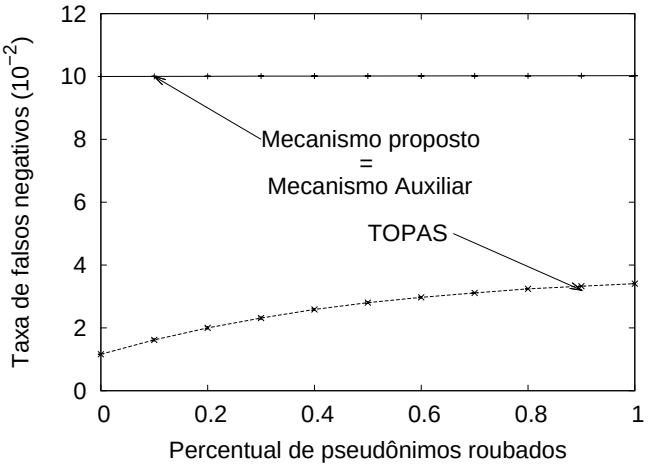
(b) Razão entre falsos negativos e falsos negativos do mecanismo auxiliar.

Figura 5.7: Avaliação dos falsos negativos.

É importante observar que o mecanismo TOPAS possui uma taxa de falsos positivos superior à do mecanismo auxiliar quando o percentual de *spammers* que utilizam servidores legítimos é maior do que 70%. Já o mecanismo proposto, além de possuir uma taxa de falsos positivos menor, não supera a taxa de falsos positivos do mecanismo auxiliar, mesmo quando 80% dos *spammers* utilizam os servidores legítimos, devido ao uso do mecanismo de proteção. A taxa de falsos negativos do mecanismo proposto é igual à taxa de falsos negativos do mecanismo auxiliar, porém maior do que a do mecanismo TOPAS. A taxa de falsos negativos do TOPAS possui o comportamento de aumentar inicialmente e depois diminuir uma vez que inicialmente os *spammers* se beneficiam da reputação dos servido-



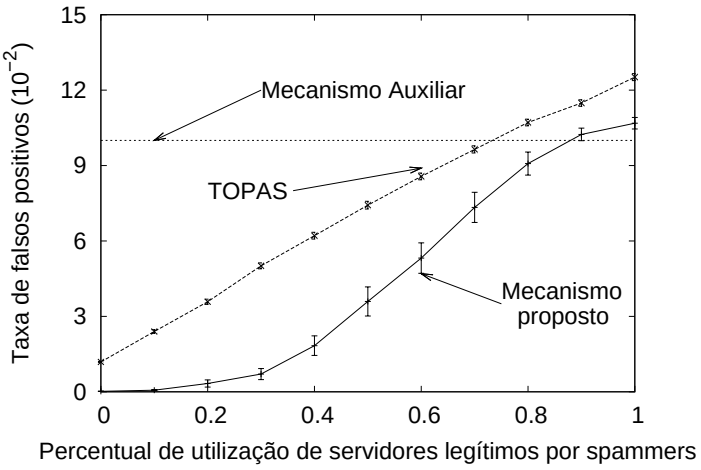
(a) Relação entre falsos positivos e pseudônimos roubados.



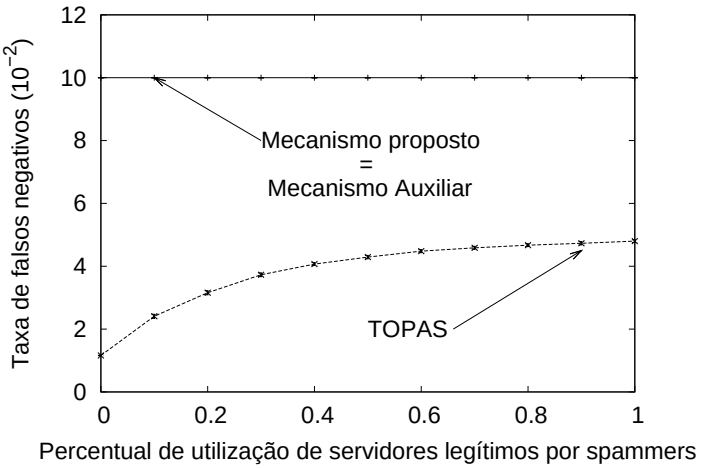
(b) Relação entre falsos negativos e pseudônimos roubados.

Figura 5.8: Influência dos pseudônimos roubados.

res legítimos, porém com o aumento da utilização dos servidores legítimos, a reputação desses servidores diminui, causando uma diminuição na taxa de falsos negativos para os *spammers* e o aumento dos falsos positivos para os usuários legítimos.



(a) Relação entre falsos positivos e utilização de servidores legítimos por *spammers*.



(b) Relação entre falsos negativos e utilização de servidores legítimos por *spammers*.

Figura 5.9: Influência da utilização de servidores legítimos por *spammers*.

Capítulo 6

Conclusões

MENSAGENS eletrônicas não solicitadas causam um grande prejuízo tanto para os usuários quanto para os provedores de serviço. Os prejuízos causados aos provedores de serviço são os relativos ao consumo de recursos tais como banda passante, memória e processamento, enquanto os relativos aos usuários são a perda de tempo e produtividade dos destinatários. Outro mal ainda maior é a redução da credibilidade dos usuários na Internet. A insatisfação entre os usuários é cada vez maior tanto pela perda de tempo na recepção e leitura das mensagens, quanto pela possibilidade de disseminação de vírus e de outros programas que causam a perda de dados e o comprometimento da segurança de seus computadores. As estatísticas mostram que os *spams* já correspondem a pelo menos dois terços de todo o tráfego de correio eletrônico transportado pelos provedores de serviço, causando prejuízos da ordem de milhões de dólares [1]. Algumas previsões mais pessimistas estimam que, em poucos anos, as mensagens não solicitadas serão responsáveis por 95% do tráfego de correio eletrônico na Internet [2].

A adoção de sistemas anti-spam é a principal contramedida às mensagens não solicitadas. No entanto, os sistemas anti-*spam* podem acabar filtrando incorretamente as mensagens legítimas, causando os falsos positivos. Esses falsos positivos têm um impacto muito grande para os usuários. As mensagens legítimas filtradas podem gerar grandes transtornos e atrasos no processo de comunicação. Sendo assim, os custos com os falsos positivos tendem a ser altos, uma vez que informações estratégicas e oportunidades podem ser perdidas, gerando graves conseqüências profissionais e pessoais. Para comparar

a importância das taxas de falsos positivos e falsos negativos, a ocorrência de um único falso negativo irá causar a perda de alguns segundos de atenção e produtividade do usuário, enquanto que apenas um falso positivo pode acarretar a perda de uma oportunidade extremamente lucrativa e vantajosa para o usuário.

Neste trabalho foi proposto um mecanismo anti-*spam* com o objetivo de reduzir a taxa de falsos positivos. Para tanto, o mecanismo proposto se serve de um mecanismo anti-*spam* convencional como mecanismo auxiliar e se serve de informações de reputação do servidor e também do usuário remetente. A reputação do remetente verifica o histórico de comportamento dos usuários para classificar se uma mensagem é *spam* ou não. Assim, usuários com um histórico de enviar mensagens legítimas passam a ter uma probabilidade menor de ter suas mensagens classificadas como *spam*. Para avaliar o histórico de comportamento dos usuários é necessário que exista um mecanismo de autenticação dos remetentes. A identificação dos usuários apenas pelo endereço eletrônico não é suficiente, já que o protocolo SMTP não possui nenhum mecanismo de autenticação dos remetentes, que podem ser facilmente forjados. Como solução para esse problema é proposto um mecanismo de autenticação baseado em pseudônimos. Cada usuário possui um ou mais pseudônimos sem nenhuma informação pessoal associada e o processo de autenticação passa a ser dos pseudônimos. Após realizar o processo de autenticação, o servidor pode verificar o histórico de comportamento do usuário que possui o pseudônimo que acabou de se autenticar. Entretanto, a observação do histórico apenas localmente não é suficiente, uma vez que o conhecimento sobre o comportamento dos usuários será apenas local e não contará com informações sobre o comportamento observado por outros servidores. Simplesmente trocar e aceitar informações originadas de outros servidores também não é eficiente, já que podem existir servidores maliciosos que informam valores errados para o seu próprio benefício ou para diminuir a eficiência do sistema. Por isso, deve-se levar em consideração na troca de informações entre os servidores a reputação de cada servidor, para avaliar se a informação deve ser considerada ou não e, caso seja considerada, qual será o grau de confiança dessa informação. Assim, junto com o mecanismo de autenticação dos remetentes que permite identificar os usuários, deve existir um mecanismo de reputação para que os servidores possam trocar informações sobre o histórico de comportamento dos usuários.

Neste trabalho foi inicialmente apresentado o sistema ADES (**A**nálise **DE** **S**pam), desenvolvido para avaliar a eficiência dos mecanismos anti-*spam* atuais. Os resultados mostraram uma taxa de falsos negativos alta, entre 2,3% e 67,4%. Também foi observada uma taxa de falsos positivos acima de 2,3% para todos os mecanismos, o que é alto considerando-se o impacto negativo que um falso positivo pode causar para os usuários. O mecanismo de filtros bayesianos obteve o melhor resultado, com 2,3% de falsos positivos. Mesmo assim, a taxa de falsos positivos dos mecanismos atuais é consideravelmente alta, considerando o impacto que os falsos positivos causam para os usuários. Assim, mecanismos que diminuam as taxas de falsos positivos dos mecanismos anti-*spam* utilizados atualmente são uma necessidade, que é justamente o objetivo do mecanismo proposto.

Para o mecanismo proposto, foi desenvolvido um modelo analítico do seu funcionamento. A partir do modelo, a proposta deste trabalho foi comparada com outro mecanismo proposto na literatura em relação à taxa de falsos positivos e falsos negativos para diferentes cenários. Os resultados da comparação com outro mecanismo, que considera apenas a reputação do servidor para classificar as mensagens, mostram que levar em consideração apenas a reputação do servidor não é suficiente, já que os *spammers* podem ser beneficiar dos servidores legítimos com alta reputação. Como o mecanismo proposto considera a reputação do usuário além da reputação do servidor, esse problema não ocorre. Assim, o mecanismo proposto sempre possui uma taxa de falsos positivos menor do que o outro mecanismo comparado no cenário em que se avalia a influência da utilização de servidores legítimos por *spammers*.

Para avaliar o mecanismo proposto também foi desenvolvido um simulador que implementa o mecanismo proposto e outro mecanismo que leva em conta apenas a reputação dos servidores para realizar a comparação com o mecanismo proposto. Vários testes de sanidade foram realizados para comparar os resultados obtidos através do simulador e o valor esperado através do modelo. Assim, os resultados do teste de sanidade validaram a implementação do simulador desenvolvido. As simulações mostram a eficiência do mecanismo proposto, que reduz em 537 vezes a taxa de falsos positivos quando a taxa de falsos positivos do mecanismo auxiliar é de 10%, ou seja, uma redução de 99,81%. Nas simulações também foram avaliadas contramedidas que podem ser utilizadas pelos *spammers* para tentarem burlar o mecanismo proposto. Primeiramente foi avaliado o impacto

do roubo de pseudônimos de usuários legítimos. Com o roubo do pseudônimo, o *spammer* passa a se beneficiar da reputação do usuário legítimo para enviar as mensagens. Os resultados da simulação mostram a robustez do mecanismo proposto em relação ao roubo de pseudônimos de usuários legítimos. Mesmo com altas taxas de roubo de pseudônimos, a taxa de falsos positivos não aumentou consideravelmente e a taxa de falsos negativos também não aumentou, já que a reputação do pseudônimo diminui muito rapidamente quando o *spammer* começa a enviar *spams* com o pseudônimo roubado, tornando-o inútil. Através da simulação também foi avaliada a influência do uso de servidores legítimos por *spammers*. O mecanismo que é comparado ao mecanismo proposto e que leva em consideração apenas a reputação dos servidores teve uma alta taxa de falsos positivos, já que os *spammers* se beneficiaram da reputação dos servidores legítimos. Já no mecanismo proposto, a taxa de falsos positivos somente teve um pequeno aumento devido à diminuição da reputação dos servidores legítimos, uma vez que um número maior de mensagens *spams* era enviado através dos servidores legítimos.

Dessa forma, o mecanismo proposto provê ganhos significativos na redução da taxa de falsos positivos, que é um problema extremamente relevante nos mecanismos anti-*spam*. Como trabalhos futuros, pretende-se avaliar outras formas de disseminação da informação de reputação entre os servidores, avaliar outras funções para o cálculo da reputação e a implementação do mecanismo proposto. A arquitetura do sistema ADES pode ser reutilizada para implementar o mecanismo proposto, já que o sistema ADES já implementa a interface com o servidor de mensagens, que permite avaliar as mensagens e decidir qual ação será tomada com a mensagem. Dessa forma, para implementar o mecanismo proposto é necessário apenas implementar o módulo de troca de informações entre servidores e o módulo de avaliação da reputação.

Referências Bibliográficas

[1] PFLEEGER, S. L., E BLOOM, G. Canning spam: Proposed solutions to unwanted email. *IEEE Security & Privacy Magazine 3*, 2 (março de 2005), 40–47.

[2] HOANCA, B. How good are our weapons in the spam wars? *IEEE Technology and Society Magazine 25*, 1 (abril de 2006), 22–30.

[3] MCGIBNEY, J., E BOTVICH, D. Establishing trust between mail servers to improve spam filtering. Em *ATC* (2007), vol. 4610 of *Lecture Notes in Computer Science*, Springer, pág. 146–155.

[4] TAVEIRA, D. M., MORAES, I. M., RUBINSTEIN, M. G., E DUARTE, O. C. M. B. Técnicas de defesa contra spam. Em *Livro Texto dos Mini-cursos do VI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais* (2006), Sociedade Brasileira de Computação, pág. 202–250.

[5] CUKIER, W. L., CODY, S., E NESSELROTH, E. J. Genres of spam: Expectations and deceptions. Em *Hawaii International Conference on System Sciences (HICSS)* (janeiro de 2006), pág. 1–10.

[6] WHITWORTH, B., E WHITWORTH, E. Spam and the social-technical gap. *IEEE Computer Magazine 37*, 10 (outubro de 2004), 38–45.

[7] EMERY, T. MIT conference takes aim at spam emails. *Associated Press* (jan de 2003).

[8] LAUFER, R. P., MORAES, I. M., VELLOSO, P. B., BICUDO, M. D. D., CAMPISTA, M. E. M., DE O. CUNHA, D., COSTA, L. H. M. K., E DUARTE, O. C.

M. B. Negação de serviço: Ataques e contramedidas. Em *Livro Texto dos Minicursos do V Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais* (2005), Sociedade Brasileira de Computação, pág. 1–63.

[9] WITTEL, G. L., E WU, S. F. On attacking statistical spam filters. Em *First Conference on Email and Anti-Spam (CEAS2004)* (julho de 2004).

[10] SPAMMER-X, POSLUNS, J., E SJOUWERMAN, S. *Inside the SPAM Cartel: Trade Secrets from the Dark Side*, 1^a ed. Syngress Publishing, 2004.

[11] LOWD, D., E MEEK, C. Goodword attacks on statistical spam filters. Em *Second Conference on Email and Anti-Spam (CEAS2005)* (julho de 2005).

[12] RAMACHANDRAN, A., E FEAMSTER, N. Understanding the network-level behavior of spammers. Em *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications* (2006), ACM Press, pág. 291–302.

[13] REKHTER, Y., E LI, T. A Border Gateway Protocol 4 (BGP-4), março de 1995. RFC 1771.

[14] RAMACHANDRAN, A., FEAMSTER, N., E VEMPALA, S. Filtering spam with behavioral blacklisting. Em *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security* (New York, NY, USA, 2007), ACM, pág. 342–351.

[15] JUNG, J., E SIT, E. An empirical study of spam traffic and the use of DNS black lists. Em *ACM SIGCOMM conference on Internet measurement (IMC' 04)* (2004), ACM Press, pág. 370–375.

[16] THE APACHE SPAMASSASSIN PROJECT. SpamAssassin: The powerful #1 open-source spam filter. http://spamassassin.apache.org/. Acessado em 4 de Fevereiro de 2008.

[17] THE APACHE SPAMASSASSIN PROJECT. SpamAssassin tests performed: v3.1.x. http://spamassassin.apache.org/tests_3_1_x.html. Acessado em 4 de Fevereiro de 2008.

[18] PEARL, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[19] WONG, M., E SCHLITT, W. *Sender Policy Framework (SPF) for Authorizing Use of Domains in E-MAIL, Version 1*. RFC 4408, abril de 2006.

[20] TAVEIRA, D. M., MATTOS, D. M. F., E DUARTE, O. C. M. B. Ferramenta para análise de características de spams e mecanismos anti-spam. Em *Salão de Ferramentas do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'08)* (maio de 2008).

[21] SCHRYEN, G. An e-mail honeypot addressing spammers' behavior in collecting and applying addresses. Em *Systems, Man and Cybernetics (SMC) Information Assurance Workshop* (2005), pág. 37–41.

[22] ANDREOLINI, M., BULGARELLI, A., COLAJANNI, M., E MAZZONI, F. Honeyspam: Honeypots fighting spam at the source. Em *SRUTI05: Steps to Reducing Unwanted Traffic on the Internet Workshop* (2005), pág. 77–83.

[23] POSTFIX. Postfix SMTP access policy delegation. `http://www.postfix.org/SMTPD_POLICY_README.html`, 2006. Acessado em 4 de Fevereiro de 2008.

[24] PRAKASH, V. V. Vipul's razor. http://razor.sourceforge.net/, 2007. Acessado em 4 de Fevereiro de 2008.

[25] TAVEIRA, D. M., E DUARTE, O. C. M. B. Mecanismo anti-spam baseado em autenticação e reputação. Em *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'08)* (maio de 2008).

[26] TAVEIRA, D. M., E DUARTE, O. C. M. B. A monitor tool for anti-spam mechanisms and spammers behavior. Em *6th IEEE Workshop on End-to-End Monitoring Techniques and Services (E2EMon 2008)* (abril de 2008).

[27] GOLBECK, J., E HENDLER, J. Reputation network analysis for email filtering. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (2004).

[28] BOYKIN, P. O., E ROYCHOWDHURY, V. P. Leveraging social networks to fight spam. *IEEE Computer Magazine 38*, 4 (abril de 2005), 61–68.

[29] GOMES, L. H., BETTENCOURT, L. M. A., ALMEIDA, V. A. F., ALMEIDA, J. M., E CASTRO, F. D. O. Quantifying social vs. antisocial behavior in email networks. *ArXiv Physics e-prints* (janeiro de 2006).

[30] GOMES, L. H., ALMEIDA, R. B., BETTENCOURT, L. M. A., ALMEIDA, V. A. F., E ALMEIDA, J. M. Comparative graph theoretical characterization of networks of spam and legitimate email. Em *Second Conference on Email and Anti-Spam (CEAS2005)* (julho de 2005).

[31] GOMES, L. H., CASTRO, F. D. O., ALMEIDA, R. B., BETTENCOURT, L. M. A., ALMEIDA, V. A. F., E ALMEIDA, J. M. Improving spam detection based on structural similarity. Em *Steps to Reducing Unwanted Traffic on the Internet (SRUTI2005)* (julho de 2005).

[32] MACINTOSH, R., E VINOKUROV, D. Detection and mitigation of spam in IP telephony networks using signaling protocol analysis. *IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication* (abril de 2005), 49–52.

[33] BALASUBRAMANIYAN, V. A., AHAMAD, M., E PARK, H. CallRank: Combating SPIT using call duration, social networks and global reputation. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (2007).

[34] SEIGNEUR, J.-M., DIMMOCK, N., BRYCE, C., E JENSEN, C. D. Combating spam with TEA. Em *Conference on Privacy, Security and Trust* (2004).

[35] MCGIBNEY, J., E BOTVICH, D. A trust overlay architecture and protocol for enhanced protection against spam. Em *ARES '07: Proceedings of the The Second International Conference on Availability, Reliability and Security* (Washington, DC, USA, 2007), IEEE Computer Society, pág. 749–756.

[36] DE OLIVEIRA, L. B. Trustmail: Um modelo de confiança entre servidores de e-mail. Tese de Mestrado, Pontifícia Universidade Católica do Paraná Programa de Pós-Graduação em Informática Aplicada, março de 2005.

[37] TAYLOR, B. Sender reputation in a large webmail service. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (julho de 2006).

[38] KAMVAR, S. D., SCHLOSSER, M. T., E GARCIA-MOLINA, H. The eigentrust algorithm for reputation management in P2P networks. Em *WWW '03: Proceedings of the 12th international conference on World Wide Web* (2003), pág. 640–651.

[39] GRAY, A., E HAAHR, M. Personalised, collaborative spam filtering. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (2004).

[40] DAMIANI, E., DI VIMERCATI, S. D. C., PARABOSCHI, S., E SAMARATI, P. P2p-based collaborative spam detection and filtering. Em *P2P '04: Proceedings of the Fourth International Conference on Peer-to-Peer Computing (P2P'04)* (Washington, DC, USA, 2004), IEEE Computer Society, pág. 176–183.

[41] CHIRITA, P.-A., DIEDERICH, J., E NEJDL, W. Mailrank: using ranking for spam detection. Em *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management* (New York, NY, USA, 2005), ACM, pág. 373–380.

[42] BOYKIN, P. O., E ROYCHOWDHURY, V. P. Personal email networks: An effective anti-spam tool. *IEEE Computer 38*, 4 (abril de 2005), 61–68.

[43] ALLMAN, E., CALLAS, J., DELANY, M., LIBBEY, M., FENTON, J., E THOMAS, M. *DomainKeys Identified Mail (DKIM) Signatures*. RFC 4871, maio de 2007.

[44] KLENSIN, J. *Simple Mail Transfer Protocol*. RFC 2821, abril de 2001.

[45] GARFINKEL, S. Email-based identification and authentication: an alternative to PKI? *Security & Privacy Magazine, IEEE 1*, 6 (Nov.-Dec. 2003), 20–26.

[46] KRAUT, R. E., SUNDER, S., TELANG, R., E MORRIS, J. H. Pricing Electronic Mail to Solve the Problem of Spam. Relatório Técnico 05-24, Yale ICF, 2005.

[47] ABADI, M., BURROWS, M., MANASSE, M., E WOBBER, T. Moderately hard, memory-bound functions. *ACM Trans. Inter. Tech. 5*, 2 (2005), 299–327.

[48] BACK, A. Hash cash - a denial of service counter-measure. http://www.hashcash.org/. Acessado em 4 de Fevereiro de 2008.

[49] GOMES, L. H., CAZITA, C., ALMEIDA, J. M., ALMEIDA, V., E WAGNER MEIRA, J. Characterizing a spam traffic. Em *ACM SIGCOMM conference on Internet measurement (IMC'04)* (2004), ACM Press, pág. 356–369.

[50] GOMES, L. H., CAZITA, C., ALMEIDA, J. M., ALMEIDA, V., E WAGNER MEIRA, J. Workload models of spam and legitimate e-mails. *Performance Evaluation 64*, 7-8 (2007), 690–714.

[51] DOUCEUR, J. R. The sybil attack. Em *IPTPS '01: Revised Papers from the First International Workshop on Peer-to-Peer Systems* (London, UK, 2002), Springer-Verlag, pág. 251–260.

[52] BORISOV, N. Computational puzzles as sybil defenses. Em *P2P '06: Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing* (Washington, DC, USA, 2006), IEEE Computer Society, pág. 171–176.

[53] FALL, K., E VARADHAN, K. *The ns Manual*. UC Berkeley, LBL, USC/ISI e Xerox, 2006.

[54] GOODMAN, J. T., E ROUNTHWAITE, R. Stopping outgoing spam. Em *ACM conference on Electronic commerce (EC'04)* (2004), ACM Press, pág. 30–39.

[55] CLAYTON, R. Stopping spam by extrusion detection. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)* (2004).

[56] CLAYTON, R. Stopping outgoing spam by examining incoming server logs. Em *Proceedings of the Second Conference on Email and Anti-Spam (CEAS)* (2005).

[57] MORI, G., E MALIK, J. Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. Em *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2003), pág. 134–141.