Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Fabio Rafael de Almeida Beato Filho

Avaliação comparativa de técnicas de filtragem de mensagens indesejadas

São Paulo
2013

Fabio Rafael de Almeida Beato Filho

Avaliação comparativa de técnicas de filtragem de mensagens indesejadas

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da computação.

Data da aprovação ____/_____/_____

_____
Prof. Dr. Wagner Luiz Zucchi (Orientador)
IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Membros da Banca Examinadora:

Prof. Dr. Wagner Luiz Zucchi (Orientador)
IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Prof. Dr. Volnys Borges Bernal (Membro)
IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Profa. Dr. José Eduardo Zindel Deboni (Membro)
IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Fabio Rafael de Almeida Beato Filho

Avaliação comparativa de técnicas de filtragem de mensagens indesejadas

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Computação.

Área de Concentração: Redes de Computadores.

Orientador: Prof. Dr. Wagner Luiz Zucchi

São Paulo
Agosto/2013

Ficha Catalográfica
Elaborada pelo Departamento de Acervo e Informação Tecnológica – DAIT
do Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT

B369a **Beato Filho, Fabio Rafael de Almeida**
Avaliação comparativa de técnicas de filtragem de mensagens indesejadas. / Fabio Rafael de Almeida Beato Filho. São Paulo, 2013.
70p.

Dissertação (Mestrado em Engenharia de Computação) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Área de concentração: Redes de Computadores.

Orientador: Prof. Dr. Wagner Luiz Zucchi

1. Eficácia 2. Técnica antispam 3. Spam 4. Internet (redes de computadores) 5. Correio eletrônico 6. Segurança de computadores 7. Tese I. Zucchi, Wagner Luiz, orient. II. IPT. Coordenadoria de Ensino Tecnológico III. Título

13-64 CDU 004.056.53(043)

DEDICATÓRIA

Dedico este trabalho à minha família, à minha namorada e aos amigos que me apoiaram nessa longa caminhada árdua e desafiadora. Um agradecimento especial a minha mãe que me orientou nos momentos mais complicados.

AGRADECIMENTOS

A empresa Mira OTM Transportes que forneceu os equipamentos que foram utilizados para o desenvolvimento prático e ao meu amigo Havary Camara que me ajudou na configuração do laboratório em ambiente *Linux*.

Agradeço ao professor Zucchi pela orientação, paciência e anos de convivência que foram fundamentais para o andamento deste trabalho.

RESUMO

Para evitar mensagens indesejadas, administradores de sistemas utilizam variadas técnicas *antispam*. Algumas dessas técnicas têm sido incorporadas em produtos, serviços e softwares para diminuir a carga de mensagens indesejadas nos sistemas e correios eletrônicos. Não há uma técnica ou uma solução completa que resolva o problema da atividade *spam*. Existem duas características principais que apoiam a construção das ferramentas *antispam*: as taxas de falsos positivos e falsos negativos. Falsos positivos são todas as mensagens lícitas que são classificadas como *spam* e os falsos negativos são as mensagens ilícitas que as ferramentas *antispam* não conseguem identificar. O objetivo dessa pesquisa é realizar uma análise de técnicas *antispam* e comparar a eficácia de cada uma, considerando a retenção dos falsos positivos e falsos negativos. O trabalho utiliza software livre para construção do ambiente de testes, realização da recepção dos *e-mails*, configuração das técnicas *antispam* e para realizar o processamento e armazenamento das informações resultantes dos testes que são feitos em cada técnica. A principal contribuição dessa pesquisa é possibilitar uma orientação na implementação de novas ferramentas *antispam* através da escolha de técnicas que redundem em maior eficácia na retenção de falsos positivos ou falsos negativos.

Palavras Chaves: *spam*; segurança; técnicas *antispam*; *e-mail*, internet, correio eletrônico.

ABSTRACT

Comparative evaluation of unwanted messages filtering techniques

To avoid unwanted messages, systems administrators use several *antispam* technics. Some of these techniques have been incorporated in products, services and software's to reduce the unwanted messages load in electronic mailing systems. There is no technique or solution which solves the *spam* activity problem completely. Two characteristics corroborate with *antispam*'s tools construction: the rates of false positives and false negatives. False positives are all licit messages which are classified as *spam* and the false negatives are illicit messages which *antispam* tools can't identify. The research's objective is to perform a technical analysis of *antispam* techniques and compare the effectiveness of each one, considering the retention of false positives and false negatives. This work is free software based in order to build a test environment, perform electronic mailing reception, *antispam*'s technical configuration and to perform the processing and storage of the test resulting information done with each technique. The main contribution of this research is to enable an orientation in the implementation of new *antispam* tools through the choice of techniques that result in efficacious retention of false positives or false negatives.

Key Words: *spam*, security, *antispam* techniques, *e-mail*, internet, electronic mail.

Lista de ilustrações

Figura 1	Funcionamento do *Greylisting*	25
Figura 2	Exemplo de raiz DNS e DNS reverso	30
Figura 3	Visualização dos *bad neighborhoods*	32
Figura 4	Arquitetura do *SpamTracker*	34
Figura 5	Arquitetura do sistema *GNU* ADES	36
Figura 6	Exemplo de linha adicionada ao cabeçalho da mensagem	37
Figura 7	Resultados dos testes das ferramentas *antispam*	40
Figura 8	Eficácia na retenção de spam x geração de falsos positivos	41
Figura 9	Funcionamento lógico do ambiente de teste	46
Figura 10	Exemplo do cabeçalho de mensagem alterado	47
Figura 11	Gráfico com os resultados por taxa de falsos positivos	48
Figura 12	Gráfico com os resultados por taxa de falsos negativos	49
Figura 13	Gráfico com os resultados da técnica DNSBL	50
Figura 14	Gráfico com os resultados da técnica DNS reverso	51
Figura 15	Gráfico com os resultados da técnica SPF	52
Figura 16	Gráfico com os resultados da técnica filtro *Bayesiano*	53
Figura 17	Gráfico com os resultados da técnica *Greylisting*	54
Figura 18	Gráfico com os resultados da técnica filtro de conteúdo	55
Figura 19	Gráfico com os resultados da técnica filtro com base em assinaturas	56
Figura 20	Exemplo de teste para análise conjunta de técnicas	57

Lista de tabelas

Tabela 1 - Resultados para retenção de falsos positivos 58

Tabela 2 - Resultados para retenção de falsos negativos 59

Lista de abreviaturas e siglas

ADES	Análise de *Spam*
DCC	*Distributed Checksum Clearinghouses*
DKIM	*DomainKeys Identified Mail*
DNS	*Domain Name System*
DNSBL	*Domain Name System Black List*
GNU	*General Public License*
HD	*Hard Disk*
IP	*Internet Protocol*
IPV6	*Internet Protocol version 6*
MTA	*Mail Transfer Agent*
MX	*Mail Exchange*
PTR	*Domain Name Pointer*
RBL	*Real Time Black List*
SATA	*Serial AT Attachment*
SMTP	*Simple Mail Transfer Protocol*
SPF	*Sender Policy Framework*
SQL	*Structured Query Language*
SRS	*Sender Rewriting Scheme*
TXT	*Text Strings*
URL	*Uniform Resource Locator*

Lista de Símbolos

@ : “Arroba” - endereço ou localização

Sumário

1 INTRODUÇÃO 16
1.1 Motivação 17
1.2 Objetivo 19
1.3 Método de Trabalho 19
1.4 Contribuições 21
1.5 Organização do Trabalho 21
2 ESTADO DA ARTE 23
2.1 Técnicas *Antispam* 23
2.1.1 DNSBL 23
2.1.2 *Greylisting* 24
2.1.3 Filtro com base em assinaturas 25
2.1.4 Filtro *bayesiano* 26
2.1.5 Filtro de conteúdo 27
2.1.6 SPF 28
2.1.7 DNS Reverso 29
2.2 Pesquisas relacionadas à construção de ferramentas *antispam* 30
2.3 Pesquisas relacionadas à comparação de técnicas *antispam* 35
3 MONTAGEM DO AMBIENTE E TESTE 43
3.1 Ambiente de teste 43
3.2 Base de teste 44
3.3 Instalação e configuração das ferramentas 45
3.4 Teste das técnicas *antispam* 45
4 ANÁLISE INDIVIDUAL DAS TÉCNICAS E DISCUSSÕES 48
4.1 Resultados gerais 48
4.2 DNSBL 50
4.3 DNS reverso 51
4.4 SPF 52
4.5 Filtro *bayesiano* 53
4.6 *Greylisting* 54
4.7 Filtro de conteúdo 55
4.8 Filtro com base em assinaturas 56
5 ANÁLISE CONJUNTA DAS TÉCNICAS E DISCUSSÕES 57

5.1 Configuração do ambiente e teste 57
5.2 Testes para falsos positivos 58
5.3 Testes para falsos negativos 58
5.4 Validade e confiabilidade do teste 59
6 CONCLUSÕES 61
6.1 Considerações finais 62
6.2 Trabalhos futuros 62
REFERÊNCIAS 63
ANEXO A 66
A.1 Instalação do *spamassassin* 66
A.2 Instalação do módulo de análise do *GNU* ADES 66
A.3 Script do banco de dados 69

1 INTRODUÇÃO

Spam, no ambiente da Internet, é considerado um abuso e se refere ao envio de um grande volume de mensagens não solicitadas, ou seja, o envio de mensagens indiscriminadamente à vários usuários, sem que estes tenham requisitado tal informação. O conteúdo do *spam* pode ser propaganda de produtos e serviços, pedido de doações para obras assistenciais, correntes da sorte, propostas de ganho de dinheiro fácil, boatos desacreditando o serviço prestado por determinada empresa, entre outros. (TEIXEIRA, 2001)

É uma atividade economicamente viável, pois o remetente tem custos operacionais baixos com o gerenciamento de seus *e-mails* e os controles de segurança adotados para o bloqueio desse tipo de mensagem, nem sempre tem eficácia.

Os sistemas *antispam* são as principais ferramentas no combate da atividade *spam*. O objetivo desses sistemas é reduzir ao máximo o número de mensagens indesejadas aos usuários. São compostos por diversas técnicas que atuam de forma preventiva e inibitiva em todas as etapas do processo de envio e recepção das mensagens.

Existem duas características principais que ajudam a avaliação da eficácia das ferramentas *antispam*: as taxas de falsos positivos e falsos negativos. Falsos positivos são todas as mensagens lícitas que são classificadas como *spam* e não chegam ao seu destino. Eles possuem um impacto grande na produtividade do usuário, pois a mensagem acaba sendo perdida ocasionando atrasos no processo de comunicação e eventualmente perda de informações que podem ser estratégicas gerando consequências graves ao usuário. Já os falsos negativos possuem um impacto menor, pois o usuário que recebe a mensagem ilícita terá apenas o trabalho de apagá-la ou simplesmente ignorá-la. O custo com falsos negativos é a perda de produtividade na realização das atividades e a possibilidade de contaminação do sistema operacional com vírus e outros malefícios.

1.1 Motivação

O envio em massa de mensagens não desejadas e irrelevantes, conhecido como '*spam*', é um problema sério que, se não for combatido, poderá em breve ser considerado um ataque por negativa de serviço (*denial of service attack*) contra a própria infraestrutura de *e-mail* na Internet.

O termo "*spam*" provém de um esquete do grupo inglês *Monty Python*, passado em uma cafeteria, onde os personagens tentam se fazer ouvir enquanto um grupo de vikings canta "*SPAM*", referindo-se a um produto fabricado pela empresa *Hormel* e conhecido no Brasil como "presuntado". Embora tenha havido mudanças, em parte devido a questões comerciais envolvendo o uso de expressões como "*e-mail* comercial não solicitado" (*unsolicited commercial e-mail*), a expressão coloquial "*spam*" é utilizada neste trabalho por ser ao mesmo tempo mais familiar e mais apropriada.

A característica que define o *spam* é o seu envio indiscriminado, apesar do conhecimento de que a mensagem não será bem-vinda pela grande maioria dos destinatários. Há *e-mails* não solicitados que muitas vezes são bem-vindos e, de fato, o que torna tão difícil reduzir ou eliminar o *spam* é a necessidade de distinguir as mensagens não solicitadas, mas desejadas, das não solicitadas e não desejadas.

Certos tipos de *e-mails* comerciais também são bem-vindos, em especial as comunicações relativas a faturas e extratos de contas, *newsletters* e, em muitos casos, certos tipos de anúncios de alguma forma relevantes para o destinatário. Convocações para apresentação de trabalhos ou para participação em conferências acadêmicas circularam e ainda circulam em redes de computadores, sem provocar qualquer reclamação.

Um sistema ideal de controle de *spam* deve ter as seguintes propriedades (LEVINE, 2005):

• Eliminar todos os *e-mails* não desejados;
• Não eliminar *e-mails* desejados;
• Não exigir que os remetentes e destinatários forneçam informações de identificação;

• Ser compatível com todos os usos de *e-mail;*
• Ser compatível com todas as configurações de infraestrutura de *e-mail;*
• Ser escalável, ou seja, manter a eficiência mesmo que 90% dos usuários de Internet o adotem;
• Resistir às tentativas de ludibriar o sistema.

Nenhuma solução perfeita de controle de *spam* foi encontrada até hoje. As soluções de filtragem são compatíveis com uma grande variedade de usos e infraestruturas de *e-mail*, mas nenhum filtro identifica perfeitamente os *e-mails* não desejados sem eliminar pelo menos alguns que seriam bem-vindos. Além disso, quanto mais abrangente é o uso de um filtro, maior é o incentivo para que aqueles que costumam enviar *spam* o testem, com o objetivo de garantir que seus *e-mails* vão “furar” o bloqueio.

Os métodos de autenticação "leves", baseados em circuitos de retorno (*callback loops*) são da mesma forma compatível com a maioria das infraestruturas de *e-mail* e não exigem qualquer intervenção por parte do destinatário. Infelizmente, a intervenção exigida pelo remetente é significativa e em breve se tornaria inaceitável, caso a solução fosse adotada em grande escala.

As técnicas de autenticação "forte", baseadas em criptografia, oferecem um método para garantir que nenhum *e-mail* desejado com origem autenticada seja eliminado, mas não fornecem qualquer outra orientação.

A *Osterman Research* (2011) publicou um artigo cujo objetivo foi analisar os impactos que as empresas sofrem devido à alta taxa de falsos positivos nos sistemas *antispam* disponíveis no mercado.

O trabalho aqui proposto se aprofunda nessa pesquisa analisando técnicas *antispam*, individualmente, considerando além da taxa de falsos positivos, os falsos negativos. Ainda faltam trabalhos, tanto no campo de desenvolvimento das ferramentas quanto na comparação de técnicas existentes, que façam uma análise individual de técnicas *antispam* e meçam a eficácia utilizando as taxas de falsos positivos e principalmente falsos negativos para considerar se a ferramenta ou sistema é eficaz ou não.

1.2 Objetivo

O objetivo dessa pesquisa é analisar e comparar a eficácia, que é a taxa de falsos positivos e falsos negativos, das principais técnicas utilizadas nas ferramentas de mercado e nos trabalhos que foram pesquisados para compor essa pesquisa. As técnicas *antispam* são: *domain name system blacklist* (DNSBL), consulta de DNS reverso, *sender policy framework* (SPF), filtro de conteúdo, filtro *bayesiano*, filtro com base em assinaturas e *greylisting*.

O conceito das principais técnicas decorre da utilização e de práticas do mercado que serão detalhados na seção 2.

1.3 Método de Trabalho

O trabalho é composto pelas seguintes atividades:

a) Estado da arte:

Para desenvolvimento deste trabalho é realizada uma pesquisa com o levantamento bibliográfico das informações relevantes ao tema *spam* e técnicas *antispam*. A revisão bibliográfica é separada por trabalhos de comparações entre técnicas e ferramentas, desenvolvimento de ferramentas *antispam* e artigos sobre eficácia na retenção de falsos positivos e falsos negativos.

b) Desenvolvimento do ambiente de testes:

Para a montagem do ambiente de testes, é utilizado um computador com configurações de *hardware* padrões de um servidor. A ferramenta *antispam spamassassin* é instalada e configurada com as configurações padrão de instalação e a ferramenta *GNU* ADES é configurada para poder processar e armazenar os *e-mails* que serão testados nas técnicas *antispam*.

c) Implementação:

Para realizar os testes de medição dos falsos negativos é utilizada uma base de *spam*s do site *Spam Arquive* (BRUCE, 2011). Esse site fornece *spam*s

recebidos desde o ano 2000 para ajudar as pesquisas na área. Para realizar os testes de medição dos falsos positivos é utilizada uma base de *e-mails* lícitos própria, que está sendo armazenada desde o começo do desenvolvimento do trabalho.

d) Testes:

Todas as técnicas *antispam* são configuradas no *spamassassin* com a configuração padrão, para que nenhuma configuração customizada possa melhorar ou piorar o desempenho de alguma técnica involuntariamente. Para cada *e-mail* processado, a ferramenta *GNU ADES* testa todas as técnicas configuradas e irá gravar num banco de dados as informações necessárias para realizar o trabalho de comparação.

e) Critério e Método de Comparação:

O critério para realizar as comparações das técnicas é a eficácia na retenção de falsos positivos e falsos negativos. As informações gravadas no banco de dados são separadas, para que seja possível realizar um cálculo, baseado nos resultados obtidos de cada técnica sobre o total de *e-mails* processados nos dois testes. A seguir, todas as técnicas terão duas taxas de erro nos testes realizados, uma para falsos positivos e outra para falsos negativos.

f) Análise de resultados:

São analisadas quais técnicas tiveram uma melhor eficácia na retenção dos falsos positivos e dos falsos negativos e como esse trabalho pode contribuir para que ferramentas *antispam* sejam construídas com maior eficácia para retenção de determinados tipos de *spam*. Também é realizado outro teste onde as três melhores técnicas nos testes individuais sejam configuradas conjuntamente com o objetivo de melhorar os resultados obtidos no teste anterior.

1.4 Contribuições

As seguintes contribuições são esperadas com essa pesquisa:

a) Possibilitar que ferramentas *antispam* sejam implementadas com precisão da taxa de *e-mails* falsos positivos ou falsos negativos a partir da escolha de técnicas que possuam maior eficácia nos testes que forem realizados nessa pesquisa;
b) Atualizar estudos e pesquisas na área de comparação de técnicas *antispam* utilizando diferentes técnicas e métrica na realização dos testes para medir a eficácia na retenção de falsos positivos e falsos negativos;
c) Expandir e testar a aplicabilidade da ferramenta *GNU* ADES, que é utilizada para realizar os testes com as diferentes técnicas *antispam*.

1.5 Organização do Trabalho

Na seção 2, Estado da Arte, são descritos os principais estudos encontrados e relacionados ao tema de comparação entre as técnicas *antispam*, análise de artigos que comparam eficiências entre técnicas considerando falsos positivos e falsos negativos e a explicação sobre o funcionamento técnico de todas as técnicas que são testadas.

Na seção 3, Montagem do Ambiente e Teste, descrevem-se os equipamentos selecionados, a construção do ambiente de teste, a configuração das ferramentas e técnicas, os monitores utilizados e a descrição dos testes realizados.

Na seção 4, Análise individual das técnicas e discussões, são apresentados os resultados individuais das técnicas testadas considerando a retenção de falsos positivos e falsos negativos, a comparação dos desempenhos entre todas as técnicas e as limitações encontradas e não previstas inicialmente.

Na seção 5, Análise conjunta das técnicas e discussões, é apresentado o laboratório configurado com as três melhores técnicas nos desempenhos para os falsos positivos e falsos negativos e os resultados do teste refeito.

Na seção 6, Conclusões, são apresentadas as conclusões, considerações finais e sugestões para trabalhos futuros.

Também são listadas as referências bibliográficas utilizadas neste trabalho e os anexos são disponibilizados com os scripts para instalação e configuração dos softwares que são utilizados para montar o laboratório de teste.

2 ESTADO DA ARTE

O objetivo desta seção é analisar artigos e pesquisas científicas de ferramentas para filtragem de *spam* e comparações de técnicas *antispam*. Também é explicado o funcionamento da ferramenta *GNU* ADES e das técnicas *antispam* que são analisadas e testadas neste trabalho.

2.1 Técnicas *Antispam*

Nesta subseção é descrito o funcionamento das técnicas *antispam* testadas neste trabalho. São elas: DNSBL, consulta de DNS reverso, SPF, filtro de conteúdo, filtro *bayesiano*, filtro com base em assinaturas e *greylisting*. Essas técnicas foram escolhidas por serem muito utilizadas nas ferramentas *antispam* de mercado (*Osterman Research*, 2011) e (*Snyder*, 2012).

2.1.1 DNSBL

A técnica *Domain Name System Black List* (TIPTON; KRAUSE, 2004) faz a publicação de listas para que, se o emissor do *e-mail* estiver cadastrado em uma DNSBL, qualquer servidor de *e-mail* possa encontrá-lo por meio de consultas possibilitando classificar uma mensagem como um *spam*.

Existem inúmeros servidores DNSBL disponíveis na Internet. Cada um deles trata de categorias específicas de *spam*s, ou seja, o usuário de um serviço de DNSBL deve escolher qual servidor se enquadra na categoria de *spam* que deseja pesquisar.

A técnica requer três itens para classificar um *spam*: domínio, servidor de e-mails desse domínio e lista de endereços para publicar *spammers*. Para fazer uma consulta a um servidor DNSBL, é necessário configurar quais servidores DNSBL serão consultados. Dessa forma, no momento da chegada de cada *e-mail*, a técnica fará uma consulta a todos os servidores DNSBL listados e analisará algumas características do *e-mail*. Quando o software *antispam* recebe uma resposta positiva, ou seja, a indicação de que o *e-mail*

em questão pertence a uma lista de *spammers*, ele decide se o *e-mail* deve ser classificado ou não como *spam*.

2.1.2 *Greylisting*

O funcionamento dessa técnica baseia-se na postergação da entrega da mensagem, utilizando um sistema sincronizado de comunicação para trafegar as mensagens. (LEVINE, 2005)

Quando um *e-mail* é recebido em um servidor, a mensagem é temporariamente rejeitada, retornando como “tente novamente mais tarde”. É armazenado em um banco de dados que contém um conjunto de informações suficientes para identificar, unicamente, cada mensagem. Em um curto intervalo de tempo o servidor fará uma nova tentativa de envio. Ao receber novamente a mensagem para tentar mais tarde, o servidor remetente irá enviar novamente. Ao receber a mensagem reenviada, o servidor que utiliza a técnica *greylisting* pesquisará na base de dados o histórico da mensagem rejeitada. Obtendo resposta positiva, ela será liberada e chegará ao destinatário.

O protocolo de *e-mail* SMTP é considerado não confiável por não checar a identificação e a autenticidade do usuário que o está utilizando; portanto, a possibilidade de falhas temporárias está embutida em seu núcleo. Todo servidor de *e-mail* corretamente implementado promove tentativas de entrega de uma mensagem que tenha obtido um código de falha temporária. Isso geralmente ocorre, por exemplo, quando a fila de um servidor-destino está muito longa para ser processada, ou o servidor tem uma carga muito alta (de operações de entrada/saída).

É nesse aspecto que a técnica *greylisting* é bem sucedida. Na Figura 1, é demonstrado o funcionamento dessa técnica: Para ser viável, um *spammer* utiliza um servidor de *e-mail* desenvolvido para desconsiderar mensagens de erro no destino. Por exemplo, servidores de *e-mail* recebem uma carga de mensagens que devem ser entregues para usuários gerados a partir de uma lista de nomes. Para o *spammer*, processar cada mensagem com falha geraria um custo alto, pois consumiria largura de banda e recursos computacionais.

Ainda, é possível retransmitir a mensagem de rejeição por até 35 dias, normalmente essa funcionalidade é utilizada para identificar *spammers*.

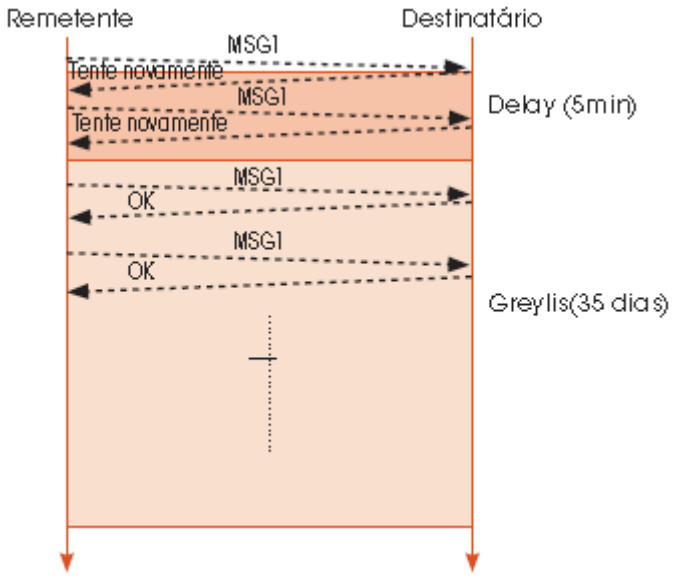

Figura 1. Funcionamento do *Greylisting*
Fonte: Levine, 2005 (traduzido).

Também deve ser considerada a eficácia de novos softwares utilizados por *spammers* que, muitas vezes, podem detectar a técnica *greylisting* por meio da análise de tempo de resposta, ou até mesmo possuindo filas de *e-mail* que simulam um servidor de *e-mail* padrão, que pode gerenciar mensagens não entregues e reenviar o *spam* ao usuário final.

2.1.3 Filtro com base em assinaturas

Um filtro que se baseie na análise de assinaturas (FABRE, 2005), possui a vantagem de classificar uma mensagem enviada, analisando se ela é idêntica à que foi recebida pelo destinatário. Além disso, quando o texto de uma mensagem contiver o nome ou o *e-mail* com assinaturas, esse filtro poderá identificar se a mensagem é um *spam* ou não, pois não só um administrador de sistemas poderá fazer esse reconhecimento, como também o próprio usuário, por meio da simples comparação entre as assinaturas. A técnica com base em assinaturas constantemente atualizadas emprega uma rede distribuída que

contém um catálogo de cada *spam* detectado por essa rede, sendo possível realizar consultas simples que poderão retornar um *spam* conhecido seguido de uma assinatura.

Uma mensagem é validada por meio de sua reputação (classificação de uma mensagem como legítima, possivelmente um *spam*, ou algo parecido) e, dessa forma, ela é classificada na rede distribuída.

2.1.4 Filtro *bayesiano*

O filtro *bayesiano* foi proposto, inicialmente, por Mehran Sahami (SAHAMI, 1996). Ele analisou um documento por um sistema de classificação, popularizado e proposto por Paulo Graham, que criou a chamada “classificação bayesiana” para detectar e julgar se uma mensagem é *spam* ou não. Uma vez configurados, os filtros *bayesiano*s não requerem manutenção. Por outro lado, usuários devem marcar mensagens como *spam* ou não *spam*, e o software de filtro apreenderá de acordo com essas marcações, criando uma base de dados com conhecimento de todos os *e-mails* processados até aquele momento. Dessa forma, o filtro *bayesiano* não reflete as técnicas de programação utilizadas pelo programador ou administrador do software.

A vantagem de um analisador *bayesiano* é a rápida aprendizagem quanto à constante mudança de palavras, de forma automática, sem intervenções administrativas.

Os *spammers* tentam, constantemente, enganar filtros *bayesiano*s inserindo caracteres estranhos junto às palavras, fazendo com que um analisador *bayesiano* classifique o *spam* como uma mensagem autêntica.

Utilizando algoritmos de filtro *bayesiano* é possível realizar a categorização automática de *e-mails*, permitindo reconhecer uma mensagem como “familiar” ou “de trabalho”. Para exemplificar como funciona o algoritmo *bayesiano*, a seguir, é explicado através de um exemplo como seria calculado a probabilidade de uma mensagem conter vírus com base no tamanho da mensagem.

A probabilidade direta de uma hipótese H condicionada a um corpo de dados E, P (H|E) está relacionada ao inverso da probabilidade dos mesmos dados e sujeita à hipótese H, P (E|H).

Matematicamente: $P(H|E) = P(E|H) \cdot P(H)$

Definem-se as variáveis: probabilidade (P); hipótese (H); corpo de dados (E).
Exemplo: Classificação de *spam*
H = *spam* contendo vírus.
E = mensagem com tamanho conhecido (indica a presença de um vírus).
Os dados utilizados pelo software *antispam* foram:

$P(H|E) = 0,5$
$P(H) = 1/1000$
$P(E) = 1/50$
$P(H|E) = P(E|H) \cdot P(H)$
$P(E) = 0,5 \cdot (1/1000) = 0,0005 = 1/50$

A probabilidade de um *spam* conter vírus, levando-se em consideração uma mensagem de tamanho conhecido, é de 0,05.

2.1.5 Filtro de conteúdo

Até recentemente, técnicas de filtro (FABRE, 2005), com base em conteúdo utilizavam palavras cadastradas pelos operadores ou administradores das ferramentas *antispam*. Dessa forma, se um servidor de *e-mail* recebesse um *spam* contendo a palavra “Viagra”, o administrador da ferramenta deveria adicioná-la à configuração, para que o servidor rejeitasse qualquer mensagem que a contivesse. Porém, modernos filtros de conteúdo podem também executar testes adicionais, como analisar o cabeçalho de um *e-mail*.

Spammers podem ser inseridos em endereços falsos no cabeçalho de uma mensagem com a intenção de esconder as identidades. Além disso, há

muitas formas de manipulação de um cabeçalho, que servem de base para a realização da análise.

São grandes as desvantagens de um filtro de conteúdo: o consumo elevado de recurso de processamento e alto número de falsos-positivos. Um administrador de sistemas que rejeite um *spam* por meio desse filtro pode acabar recusando algum outro *e-mail* válido.

Finalmente, os *spammers* podem modificar as frases ou a forma de escrever uma determinada palavra, utilizando técnicas de inserção de hífens ou espaços entre uma sílaba e outra. Como exemplo, se o *spammer* emprega a palavra “Viagra”, como “V1agra” ou “Via_gra”, torna-se difícil para o analisador de conteúdo identificar essa mensagem como um *spam*.

2.1.6 SPF

Sender Policy Framework (GÖRLING, 2007) é uma técnica para combater a falsificação de endereços de retorno dos *e-mails* (*return-path*). O mecanismo permite ao administrador de um domínio definir e publicar uma política SPF, onde são designados os endereços das máquinas autorizadas a enviar mensagens em nome deste domínio e ao administrador de um serviço de *e-mail* estabelecer critérios de aceitação de mensagens em função da checagem das políticas SPF publicadas para cada domínio.

O processo de publicação de uma política SPF é independente da implantação de checagem de SPF por parte do MTA, estes podem ou não serem feitos em conjunto. Ao publicar uma política de SPF, o administrador de um domínio está autorizando determinados MTAs a enviar *e-mails* em nome deste domínio. O objetivo é evitar que terceiros enviem mensagens indevidamente em nome de seu domínio, e que mensagens de erro causadas por *spam,* com envelope falso, sejam enviadas para o seu servidor.

Estas políticas são publicadas através de registros TXT do DNS, em formato ASCII. Um exemplo desse registro é:

Exemplo: examplo.com. IN TXT "v=spf1 a mx ip4:192.0.2.32/27 -all"

Neste caso, a política estabelece que pode enviar mensagens em nome do domínio example.com uma máquina que satisfaça um dos seguintes critérios: seu endereço IP deve ser um RR tipo A do domínio examplo.com (a); seja designada como MX do domínio examplo.com (mx); ou pertença ao bloco de endereços IP 192.0.2.32/27 (ip4).

A cláusula "-all" diz que devem ser recusados ("-", prefixo Fail) *e-mails* partindo de qualquer outro endereço IP (all).

Todas as opções de prefixos são:

"+" Pass
"-" Fail
"~" SoftFail
"?" Neutral

O prefixo é opcional, e se omitido o valor utilizado é o "+" (Pass).

A cláusula "all" deve ser sempre a cláusula mais à direita. Ela define qual resposta será retornada em uma consulta SPF, caso nenhuma das outras cláusulas se aplique. O administrador de um MTA que consulte a política SPF do domínio do remetente de um *e-mail*, como definido no envelope, poderá rejeitar ou marcar como suspeita, uma mensagem que não satisfaça à política SPF daquele domínio.

Mensagens legítimas, mas que tenham passado por um *relay* ou tenham sido redirecionadas, podem ser recusadas por MTAs que checam SPF. Para evitar que estas mensagens sejam rejeitadas, devem ser adotadas algumas estratégias, como SRS (*Sender Rewriting Scheme*) e autorizações especiais.

2.1.7 DNS Reverso

O *Domain Name System* reverso é o mecanismo que permite obter o nome de um determinado recurso, através do seu endereço IP, ou seja, é o inverso da tradução de nomes em endereços (CHESWICK; BELLOVIN, 2005). A função do DNS reverso permite a constatação da autenticidade de endereços, fazendo uma consulta no servidor DNS e avaliando se um

determinado IP realmente é o correspondente ao domínio. Se o reverso do DNS não for configurado, todas as mensagens podem ser consideradas spam, é uma opção de configuração. É um mecanismo bastante utilizado para dificultar os spams, pois um *spammer* pode estar forjando um domínio que não lhe pertença para o envio dessas mensagens indesejadas. A figura 2 mostra um exemplo de raiz de DNS e DNS reverso.

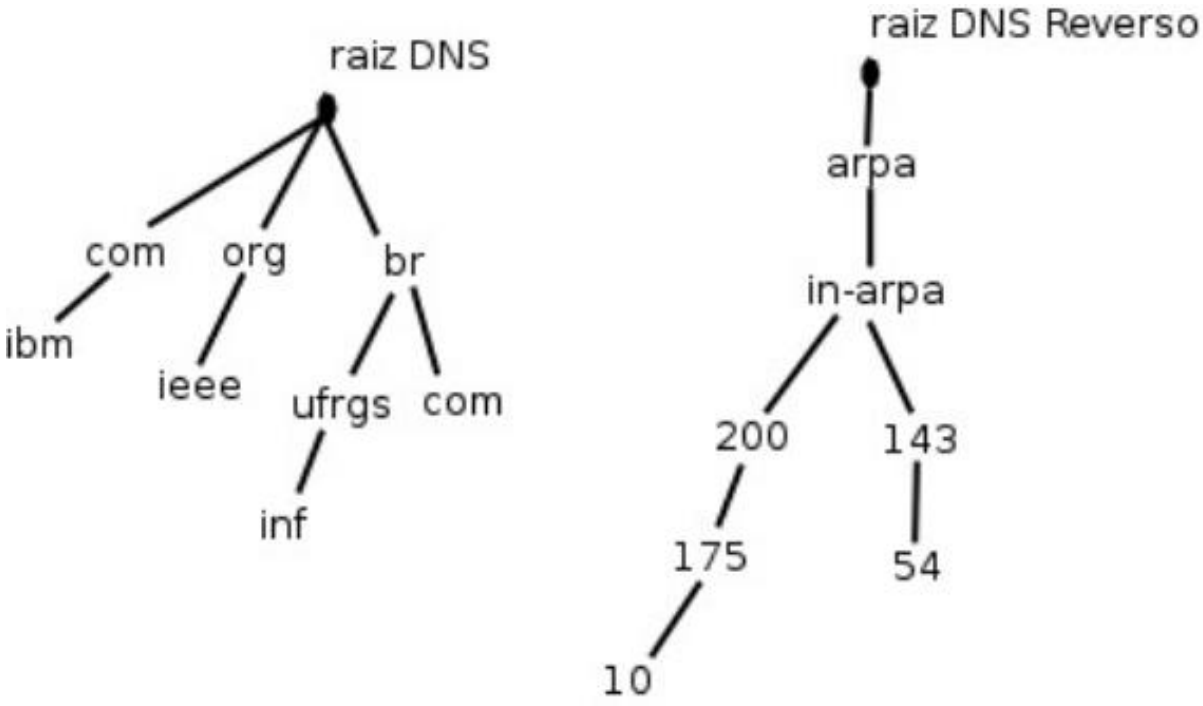

Figura 2. Exemplo de raiz DNS e DNS reverso
Fonte: Cheswick; Heswick, 2005 (adaptado)

O espaço de nomes de DNS é estruturado em árvores. Para facilidade de operação, sub-árvores podem ser delegadas a outros servidores. São utilizadas duas árvores logicamente distintas.

A primeira mapeia nomes de host, como www.uol.com.br para endereços como 200.221.2.45. A segunda árvore é para consultas inversas e contém registros PTR. Nesse caso, ela mapearia 45.2.221.200.in-addr.arpa para www.uol.com.br. Não há nenhum relacionamento imposto entre as duas árvores, embora alguns sites tenham tentado impor esse link para alguns serviços. A árvore inversa raramente é tão bem mantida e atualizada como a árvore de mapeamento direto comumente utilizada.

2.2 Pesquisas relacionadas à construção de ferramentas *antispam*

Na área de pesquisa de técnicas *antispam*, o trabalho de Wanrooij, Pras, (2010) propõe a criação de um mecanismo baseado no mapeamento de sub-redes que são denominadas de bairros. Esses bairros são vulneráveis a origem de ataques *spam* e a probabilidade de que também outros sistemas dentro do mesmo bairro sejam comprometidos é alta. Além disso, é assumido que os sistemas, uma vez infectados, vão primeiro tentar comprometer os sistemas nas proximidades. Isto leva ao conceito, dado pelos autores, de vizinhos maus (*bad neigborhoods*), a partir do qual a probabilidade de um sistema do mesmo bairro receber *spam* é maior do que de outros bairros.

Para identificar os *spam*s, foram utilizadas DNSBLs (*Domain Name System Blacklists*) que registraram quantos *spammers* foram encontrados em cada bairro. Quanto maior o número, maior as chances de que esse é um bairro ruim. Os resultados das listas DNSBLs foram comparados e combinados para aumentar a chance de acerto.

A Figura 3 mostra os bairros que mais geraram *spam* na simulação realizada. Cada sub-rede é um pixel e a intensidade das alterações de cor de cada bairro, do claro para o escuro, indicam o número de mensagens de *spam* originária dessa sub-rede.

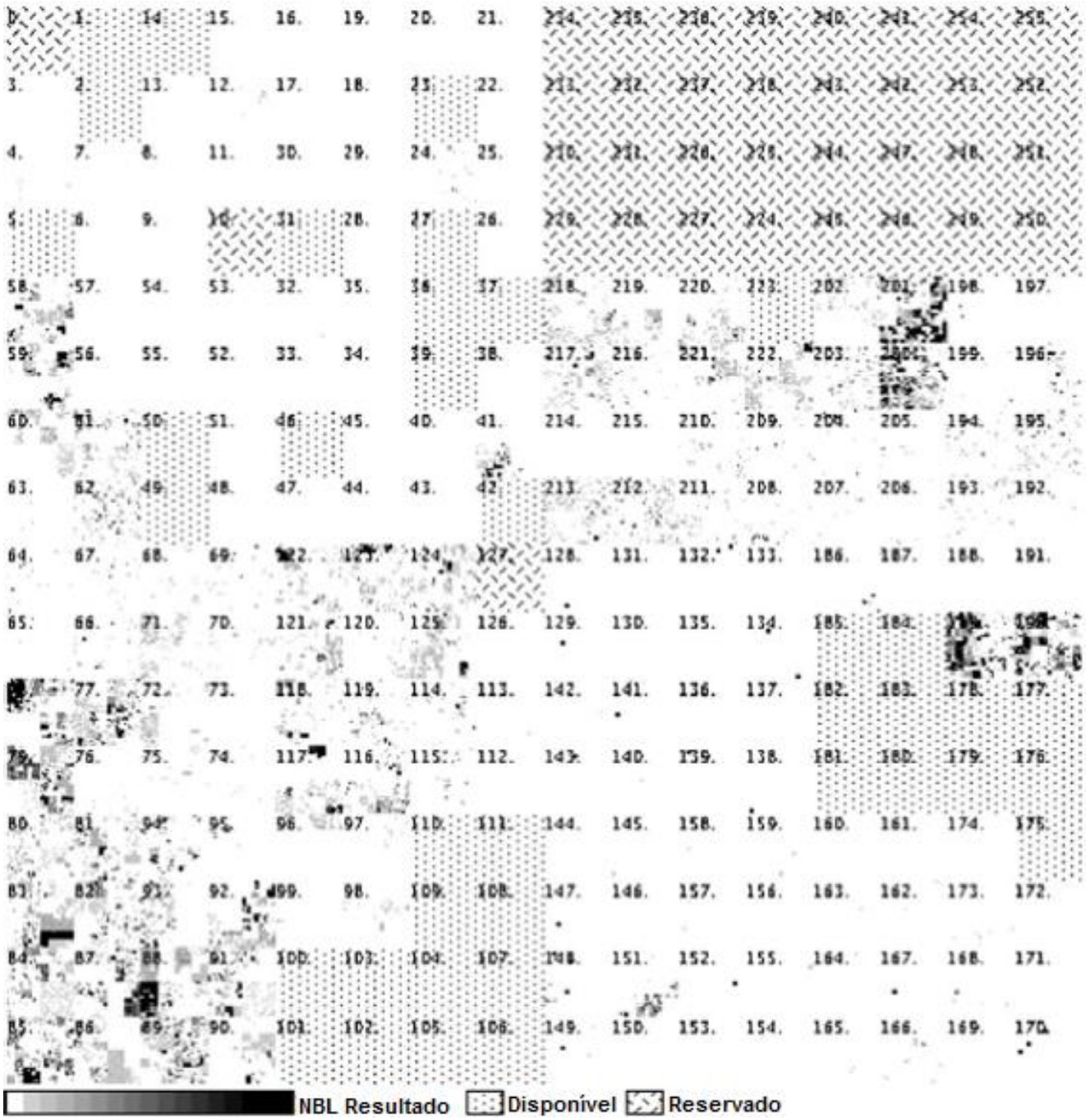

Figura 3. Visualização dos *bad neighborhoods*

Fonte: WANROOIJ, PRAS, 2010. (traduzido)

Os resultados dos testes efetuados mostraram uma taxa de retenção de *spam* maior que a ferramenta *spamassassin,* que foi utilizada como comparativo, porém não fica claro quais foram as técnicas *antispam* configuradas dentro da ferramenta para realizar os testes. Entretanto, os autores ressaltam a importância de se realizar mais testes comparando outras

técnicas *antispam* e utilizar diferentes tipos de *e-mail* em diferentes períodos de tempo.

Ramachandran, Feamster. Vempala (2007) propõem um mecanismo chamado *SpamTracker*, cujo objetivo é classificar os *spammers* utilizando uma técnica denominada *behavioral blacklisting.* Essa técnica é baseada na premissa de que os padrões de envio de *spam* podem ser mapeados e servem para padronizar o comportamento dos *spammers* no envio de *spam*. Os autores denominam esse padrão de “assinatura digital”.

Ramachandran, Feamster. Vempala (2007) afirmam que os filtros muitas vezes utilizam a reputação de um endereço IP (*Internet Protocol*) para classificar um *spam*, porém, essa técnica funciona bem apenas para os endereços IP fixos e hoje muitos endereços IP dinâmicos são utilizados. Como consequência, as DNSBLs precisam estar em constante atualização para manter um índice considerável de retenção de *spam*.

A figura 4 mostra a arquitetura da ferramenta *SpamTracker.*

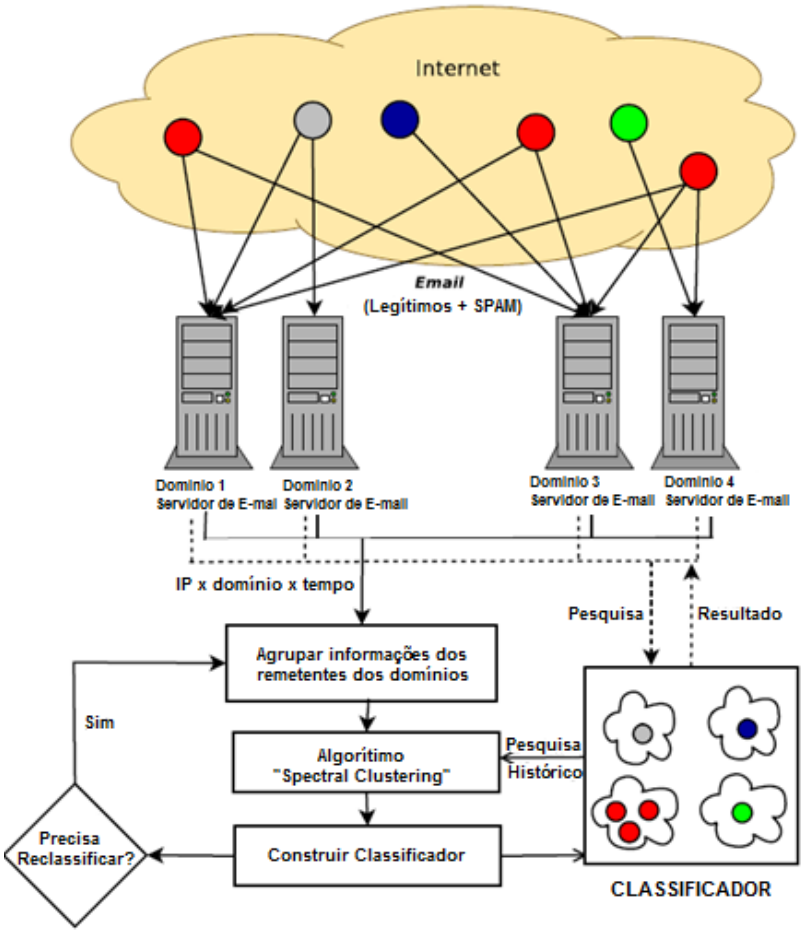

Figura 4. Arquitetura do *SpamTracker*

Fonte: Ramachandran, Feamster. Vempala (2007) (traduzido)

A ferramenta utiliza *clusters* para armazenar as informações das mensagens recebidas dos domínios e utiliza esses dados para construir *blacklist clusters*, onde é utilizado um vetor de medida para calcular as assinaturas digitais de cada *cluster*. Quando um *e-mail* é recebido, o sistema calcula a similaridade de seu padrão (assinatura digital) e compara com uma base de pontuação de *spam* para concluir se a mensagem é ou não um *spam*.

Na avaliação, Ramachandran, Feamster. Vempala (2007) mostrou que conseguiu complementar as *blacklists,* distinguindo o *spam* de um *e-mail* legítimo além de detectar muitos *spammers* antes de serem consultados em qualquer DNSBL, pela avaliação do comportamento baseado nos dados das *blacklist clusters*.

As técnicas propostas por Wanrooij, Pras (2010) e Ramachandran, Feamster. Vempala (2007) utilizaram como métrica principal para medir a eficácia das ferramentas *antispam* a taxa total de retenção de *spam*. Não é possível afirmar se as ferramentas são eficazes na retenção de falsos positivos. O trabalho aqui proposto utiliza a implementação de DNSBL como uma das técnicas a ser comparada e considerada como métrica eficaz na retenção de falsos positivos e falsos negativos para que seja possível avaliar individualmente cada mensagem recebida, *spam* ou não.

2.3 Pesquisas relacionadas à comparação de técnicas *antispam*

Taveira (2008) avalia a eficácia de técnicas *antispam* e, também, analisa características e técnicas utilizadas por *spammers* para enviar *spam*s. Para realizar a análise, foi desenvolvida a ferramenta *GNU* ADES (Análise de *Spam*) que realiza a análise da eficácia dos mecanismos *antispam* de forma individualizada e também fornece características do processo de coleta de endereços e envio de *spam*s.

A ferramenta *GNU* ADES permite realizar a análise da eficácia das técnicas *antispam* de forma individualizada. Cada mensagem é analisada por cada uma das técnicas *antispam* de forma isolada. A partir da classificação realizada por cada mecanismo *antispam* estatísticas e relatórios são gerados.

A Figura 5 apresenta a arquitetura do sistema ADES.

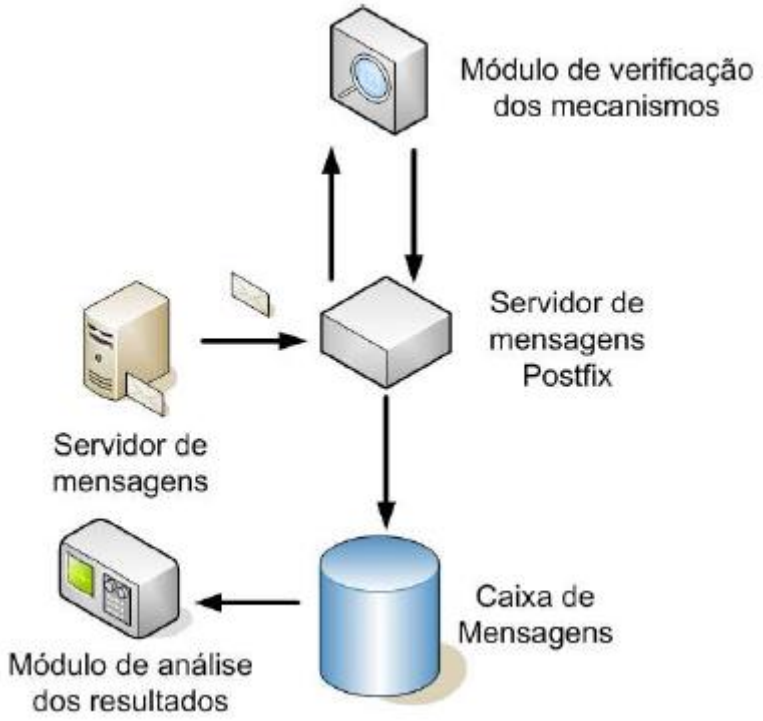

Figura 5. Arquitetura do sistema *GNU* ADES
Fonte: Taveira (2008) (adaptado)

O módulo de verificação de mecanismos realiza os testes das técnicas *antispam* analisadas em todas as mensagens recebidas pelo servidor de mensagens. O módulo é implementado na linguagem *Perl* e funciona como um servidor de políticas do servidor de mensagens *Postfix*. A arquitetura de servidores de políticas do *Postfix* permite a criação de módulos que recebam informações das mensagens que chegam ao servidor de correio eletrônico e definam uma ação a ser tomada com a mensagem. O servidor de políticas pode realizar vários testes e/ou procedimentos e, por fim, enviar um comando ao servidor de mensagens. Os principais comandos são: descartar a mensagem, aceitar a mensagem ou adicionar informações ao cabeçalho da mensagem.

O servidor de correio eletrônico é configurado para utilizar o mecanismo de pesos e regras *spamassassin* com a configuração padrão e a utilização da base de dados para cada técnica *antispam* que seja testada. Dessa forma, as informações deste mecanismo são adicionadas automaticamente ao cabeçalho da mensagem pelo servidor de correio eletrônico.

Após a realização dos testes de cada uma das técnicas, os resultados são armazenados adicionando-se uma linha ao cabeçalho da mensagem. A linha adicionada é composta por um nome identificador e um valor. O nome identificador utilizado é X-ADES e em seguida existem vários campos separados pelo símbolo “@”. O primeiro campo é obrigatoriamente o endereço IP da máquina que enviou a mensagem e em seguida outros campos que apresentam os resultados dos mecanismos aplicados.

Para cada resultado de mecanismo, o primeiro campo é o código do nome do mecanismo e o segundo campo contém os resultados do mecanismo. A figura 6 ilustra um exemplo da linha que é adicionada ao cabeçalho da mensagem:

X-ADES-ANALYSIS: @192.168.0.1@SPF@OK@DNSREVERSO@FALHOU

Figura 6. Exemplo de linha adicionada ao cabeçalho da mensagem
Fonte: O autor

No exemplo acima a mensagem foi enviada através do endereço IP 192.168.0.1. Além disso, observa-se que duas técnicas foram aplicadas, a primeira com nome código SPF que teve como resultado OK e a segunda com nome-código DNSREVERSO que teve como resultado FALHOU.

O módulo de análise de resultados tem como objetivo realizar as análises dos resultados das técnicas *antispam*. Antes de calcular os índices de falsos positivos e de falsos negativos de cada uma das técnicas, é necessário saber se a mensagem é legítima ou se é *spam*, para comparar com o resultado do mecanismo. Para isso, as mensagens recebidas são manualmente classificadas, para comparar a classificação do usuário com a classificação efetuada pelo mecanismo. As mensagens que forem *spam*s são movidas para uma pasta chamada *spam* e as mensagens legítimas podem ser colocadas em qualquer outra pasta.

O módulo de análise dos resultados analisa todas as mensagens de todas as pastas e compara o resultado da classificação do usuário com a classificação das técnicas *antispam* para determinar a taxa de falsos positivos e falsos negativos de cada mecanismo.

Nos testes que foram realizados no trabalho de Taveira (2008), foram utilizadas cinco técnicas: filtro *bayesiano*, pesos e regras, DNS reverso, SPF e *Black Lists*. O número total de mensagens utilizadas na análise foram 63.325 mensagens legítimas e 3.392.931 *spam*s, recebidas por dezoito endereços de usuários legítimos num período de um ano e seis meses.

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.

Assim como Taveira (2008), também é utilizado nesta pesquisa como métrica a retenção para falsos positivos e falsos negativos e o software *GNU* ADES para analisar individualmente cada técnica *antispam*, entretanto, uma maior quantidade e técnicas distintas são utilizadas para realizar as análises e testes. Taveira (2008) utiliza contas pessoais para recepcionar as mensagens *spam* que foram testadas, diferentemente do que é proposto para este trabalho , que utiliza uma base pública de *e-mails spam*s disponível na Internet para estudos e desenvolvimentos neste campo de pesquisa.

Wiehe, Hjelm, Wolthusen (2006) fizeram um levantamento de técnicas *antispam* e realizaram diversas comparações práticas com o objetivo de verificar as taxas de retenção de *spam* e o número de falsos positivos gerados.

Os resultados apresentados foram obtidos da coleta e análise de 300.000 mensagens entre os anos de 2005 e 2006 provenientes de contas de *e-mails* da faculdade *Gjovik University College*, na Noruega.

O estudo foi realizado em duas etapas: na primeira todas as mensagens foram classificadas como *spam* ou não *spam* de forma manual, por cada colaborador ou voluntário, e foram armazenadas em uma base de dados para, numa segunda etapa, serem comparadas individualmente com as técnicas *antispam* avaliadas.

Para montar o ambiente de testes, foi utilizado um servidor com sistema operacional *GNU / Linux,* o servidor MTA *(Message Transfer Agent) Sendmail* para receber os *e-mails* avaliados e a ferramenta de banco de dados *PostgreSQL* onde foram armazenados todos os dados coletados. As técnicas e

ferramentas *antispam* avaliadas foram: *Greylisting*, SPF (*Sender Policy Framework*), RBL (*Real Time Black Lists*), DKIM (*DomainKeys Identified Mail)*, *Razor*, DCC *(Distributed Checksum Clearinghouses*) e *spamassassin*.

Wiehe, Hjelm, Wolthusen (2006) avaliam que os resultados, de uma forma geral, foram satisfatórios. Os mecanismos analisados tiveram uma taxa de retenção de *spam* entre 37% e 75% quando analisados individualmente. Entretanto, quando as técnicas são combinadas para aumentar a efetividade na retenção de *spam*, a taxa de falsos positivos aumenta consideravelmente devido a maior probabilidade de pelo menos uma técnica considerar erroneamente um *e-mail* lícito como *spam*.

O trabalho de Wiehe, Hjelm, Wolthusen (2006) é um pouco confuso por realizar testes e comparações entre técnicas e ferramentas *antispam.* Considerando que uma ferramenta pode ter um conjunto de técnicas configuradas, que não fica claro no trabalho, de todas as técnicas testadas quais foram configuradas conjuntamente e individualmente, podendo haver distorções nos resultados se o desempenho das técnicas for analisado individualmente. O trabalho aqui proposto utiliza uma única ferramenta (*spamassassin*) e todas as técnicas que são testadas estão configuradas nessa única ferramenta.

Snyder (2012) publicou um artigo com testes realizados com as ferramentas *antispam* mais utilizadas no mercado de acordo com o “quadrante mágico para *gateways* de segurança de *e-mail*” do *Gartner* (SNYDER, 2012). O objetivo dos testes é verificar qual é a ferramenta que possui uma melhor retenção de *spam*, ou seja, menor taxa de falsos negativos. Os fabricantes que tiveram as ferramentas testadas foram: *Barracuda Networks, Cisco, Google, McAfee, Microsoft, Proofpoint, Sophos, Symantec* e *Trend Micro.*

A figura 7 apresenta os resultados dos testes.

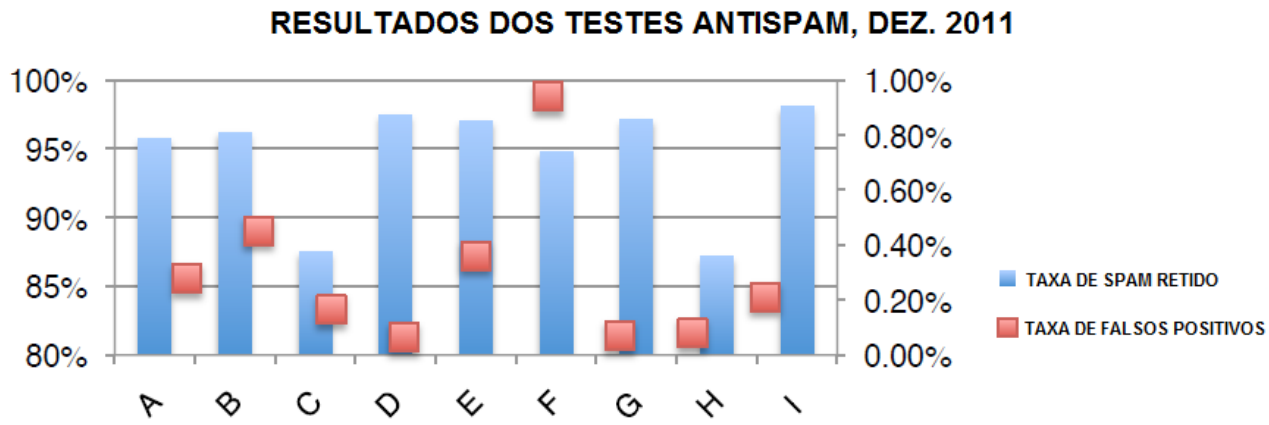

Figura 7. Resultados dos testes das ferramentas *antispam*

Fonte: *Snyder* (2012) (traduzido)

A ferramenta “I” teve o melhor desempenho nos testes com uma taxa de retenção de spam de 98.06%. Snyder (2012) afirma que comparar resultados de diferentes produtos não é uma tarefa simples, porque todos os produtos possuem grande variedade de opções de configuração e os resultados foram obtidos com o uso das opções de configuração padrão para cada produto. O trabalho aqui proposto usa a taxa de 98,06% como comparação para a retenção dos *spams* nos testes conjuntos.

A *Osterman Research* (2011) publicou um artigo cujo objetivo foi analisar os impactos, financeiros e operacionais, que as empresas sofrem devido à alta taxa de falsos positivos nos sistemas *antispam* disponíveis no mercado.

A taxa de falsos positivos é a principal métrica para avaliar as soluções de filtragem de *spam*, e em 2010, 63% de usuários de *e-mail* precisaram, regularmente, acessar uma área de quarentena, para verificar se as mensagens eram ou não *spam*, ocasionando perda de produtividade.

A figura 8 apresenta a comparação na retenção de *spam* com a taxa de falsos positivos gerados.

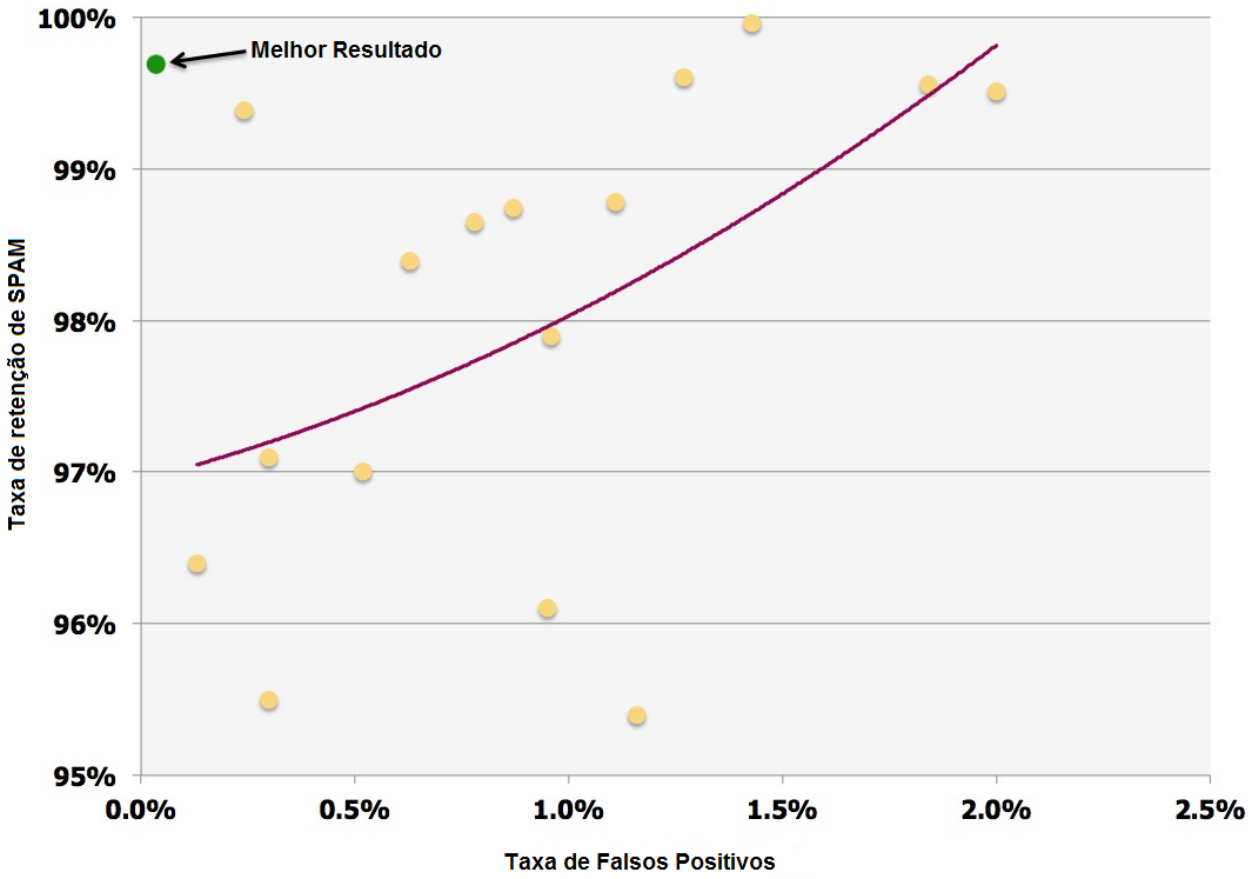

Figura 8. Eficácia na retenção de *spam* x geração de falsos positivos
Fonte: *Osterman Research* (2011) (traduzido)

A linha vermelha monstra que quanto maior for a eficiência na retenção de *spam* maior será a taxa de falsos positivos e os pontos amarelos representam as ferramentas de mercado que foram testadas e comparadas. É importante ressaltar que a medição é baseada no número total de *e-mails* legítimos e não no número total de *e-mails* recebidos.

O melhor resultado foi 0,15% de falsos positivos gerados, 1,5 para cada 1000 *e-mails*. Segundo *Osterman Research* (2011), a situação ideal seria se os sistemas *antispam* tivessem uma taxa de falso positivo nulo, entretanto, o artigo considera a possibilidade dessas ferramentas adotarem um padrão de qualidade que permitiria apenas 0,00025% de falsos positivos, ou seja, um falso positivo para cada 400 mil *e-mails*.

Quanto mais configurações são feitas para fortalecer os filtros, menor quantidade de *spam* chegará ao usuário, porém, a taxa de falsos positivos é

maior. Por outro lado, se essas ferramentas forem configuradas de forma menos eficaz, maior é o números de *spam*s que chegará ao usuário, porém a taxa de falsos positivos é menor.

Osterman Research (2011) conclui que atualmente, a maioria dos sistemas *antispam* não podem chegar nesse nível de desempenho, e que todas as decisões com relação à retenção de *spam* são baseadas no impacto no ‘negócio’ e não na métrica da taxa de falsos positivos.

O artigo da *Osterman Research* (2011) aponta uma necessidade das ferramentas serem eficazes na retenção de *spam* e também na retenção de falsos positivos. O trabalho aqui proposto possibilitará um melhor entendimento para que ferramentas sejam implementadas de acordo com técnicas que possuam melhor eficácia na retenção de falsos positivos ou falsos negativos.

3 MONTAGEM DO AMBIENTE E TESTE

Nesta seção é descrita a construção do ambiente e a realização dos testes para validação em laboratório do método proposto. Para realizar a análise de todas as técnicas citadas, foi desenvolvido um ambiente com a configuração padrão de instalação das ferramentas utilizadas para não haver nenhuma influência em possíveis configurações de *tunning* que podem distorcer os resultados finais dos testes em favorecimento a uma determinada técnica *antispam*.

Considerando que as mensagens são pré-selecionadas e inseridas diretamente na ferramenta *antispam spamassassin*, não é necessário uma infraestrutura conectada à Internet, somente é utilizado um *hardware* com os requisitos computacionais que são capazes de testar todos os *e-mails* e realizar a classificação no sistema *GNU* ADES.

Com o ambiente montado e configurado, é possível realizar os testes individuais de cada técnica *antispam* e realizar análises para medir a eficácia dessas técnicas considerando a retenção de falsos positivos e falsos negativos. Essa análise também permite observar se alguma técnica pode oferecer alguma vantagem em ser implementada em combinação com outra, para que, em cenários específicos, possa ocorrer ganho de eficácia. Muitas vezes combinar algumas técnicas podem não trazer nenhum benefício a mais na filtragem de *spam*s, apenas ocupar mais banda de *link* ou tempo de processamento da mensagem.

3.1 Ambiente de Teste

Para montar e configurar o ambiente de testes é utilizada a distribuição *Debian Linux* (DEBIAN, 2013), configurada e instalada em um computador com a configuração de *hardware*: processador Intel *core* i3-2100 com 2 núcleos, 4 *gigabyte* de memória e um HD (*hard disk*) SATA (*Serial AT Attachment*). Foi escolhido o ambiente *Linux* pela facilidade na instalação e configuração das ferramentas e, especificamente, a distribuição *Debian* devido a sua característica de aproveitamento em obter a máxima performance do *hardware*,

além de sua facilidade de instalação, personalização e atualização de pacotes pela Internet. A ferramenta *antispam spamassassin* é instalada e configurada com as configurações padrões de instalação e a ferramenta *GNU* ADES é configurada para poder processar e armazenar os *e-mails* que são testados nas técnicas *antispam*.

O servidor de *e-mail Postfix* é instalado e configurado como servidor de correio eletrônico para receber os *e-mails* e executar os comandos para as ferramentas de testes *spamassassin* e *GNU* ADES. Esses comandos visam descartar a mensagem, aceitar a mensagem ou adicionar informações ao cabeçalho da mensagem para poder viabilizar a classificação.

Para a base de dados que armazena os resultados é configurado o banco de dados *Mysql* (MYSQL, 2013), em sua versão 5.6.10. A escolha por armazenar os resultados em banco de dados deve-se ao fato de que a posterior consulta com diversas combinações fica facilitada e principalmente porque as realizações de todas as verificações em todas as mensagens demandam muito tempo.

3.2 Base de Teste

Para realizar a análise de todas as técnicas, é necessária uma base grande e diversificada de *e-mails* pré-classificados como *spam* e não *spam*. Os *e-mails* classificados como *spam* foram selecionados em um repositório público na Internet chamado "*Spam Arquive*" (BRUCE, 2011). Esse site fornece *spam*s recebidos desde o ano 2000 para contribuir com pesquisas e desenvolvimentos na área.

Para realizar os testes de medição com *e-mails* lícitos é utilizada uma base de *e-mails* com mensagens recebidas em contas de correio eletrônico do próprio autor que foram armazenadas desde o começo do desenvolvimento deste trabalho. São utilizados 500.000 *e-mails spam*s e 5.000 não *spam*s, todos dos anos entre 2011 e 2013. A discrepância no número de mensagens *spam* e não *spam* deve-se ao fato de que os testes focam nas mensagens que podem causar maiores prejuízos técnicos e financeiros, os falsos positivos, conforme o artigo da *Osterman Research* (2011).

3.3 Instalação e configuração das ferramentas

A ferramenta *spamassassin* e as técnicas *antispam* foram instaladas, através do gerenciador de pacotes do *Gentoo Linux* (GENTOO, 2013), com as configurações padrão.

Para o *spamassassin*, é utilizado um conjunto de regras para detectar também *spam*s na língua portuguesa disponível em (LAFRAIA, 2013). O restante das configurações das técnicas foi mantido em seu modo padrão (*default*). Essa medida foi tomada para que a obtenção dos resultados refletissem a operação normal dessas ferramentas, e também não favorecer nenhuma delas.

A técnica *antispam* filtro *bayesiano* necessita de uma base de "treinamento" antes da realização dos testes para criar o dicionário utilizado na classificação de *e-mails*. Portanto, são separados aleatoriamente da base de testes 3000 mensagens, sendo 500 consideradas não *spam* e 2500 consideradas como *spam*s.

Na ferramenta *GNU* ADES são implementadas as técnicas DNSBL, consulta de DNS reverso, SPF, filtro de conteúdo, filtro *bayesiano*, filtro com base em assinaturas e *greylisting*. Para cada uma dessas técnicas analisadas é inserido no cabeçalho da mensagem a informação com o resultado individual de cada técnica *antispam* testada.

O software *spamassassin* também insere novas linhas no cabeçalho da mensagem somente para as técnicas de filtro de conteúdo e filtro com base em assinatura, pois essas duas técnicas fazem parte da ferramenta *spamassassin*, que posteriormente também são analisados pela ferramenta *GNU* ADES.

3.4 Teste das técnicas *antispam*

As técnicas *antispam* implementadas na ferramenta *GNU* ADES são acionadas todas as vezes que o servidor de *e-mail* recebe uma mensagem. Embora todas as técnicas sejam de grande importância na classificação de

uma mensagem como *spam*, o filtro *antispam spamassassin* também precisa ser modificado para ser acionado a cada verificação realizada.

A figura 9 apresenta o funcionamento lógico do ambiente de testes:

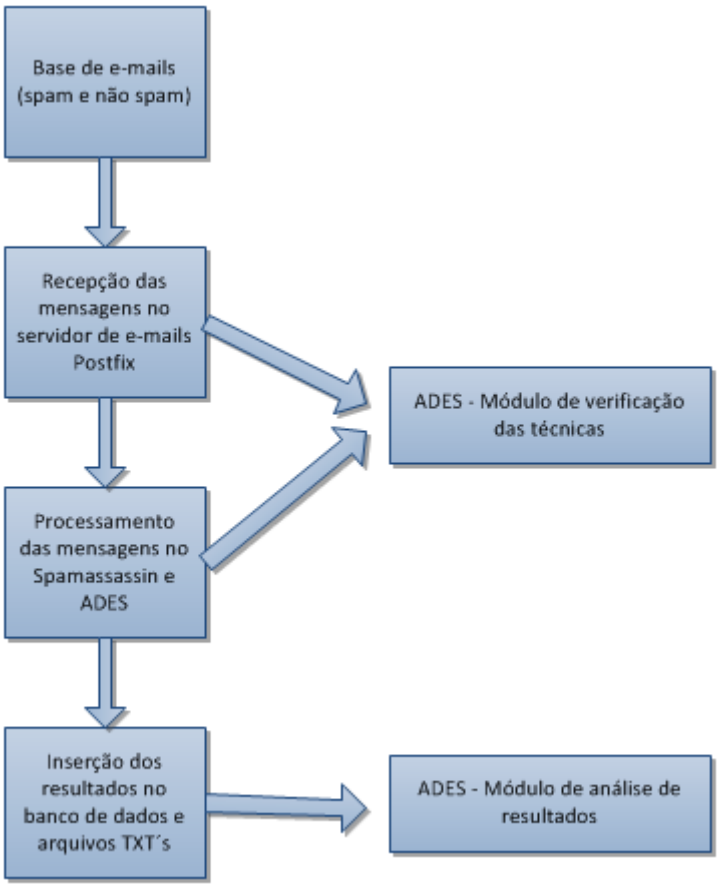

Figura 9. Funcionamento lógico do ambiente de teste
Fonte: O autor

O software *spamassassin* possui uma série de regras que permitem classificar as mensagens. Durante esse processo, são analisados campos do corpo e cabeçalho da mensagem. No corpo da mensagem são procuradas expressões regulares que são comuns em *spam*s, a cada expressão encontrada é adicionado um valor à soma total de pontos dessa mensagem, o valor adicionado depende da regra e do peso. Para uma mensagem ser considerada *spam* a soma total de pontos dessa mensagem deve ser igual ou superior a 5, valor padrão do *spamassassin*. Além do uso de expressões regulares, também são usadas todas as técnicas *antispam* que são testadas nesse trabalho. Ao fim da análise, o *spamassassin* modifica o cabeçalho da

mensagem criando novos campos, nos quais são inseridas as informações geradas pelo filtro *antispam*.

Após a realização dos testes de cada uma das técnicas, os resultados são armazenados adicionando uma linha ao cabeçalho da mensagem. A linha adicionada é composta por um nome identificador e um valor. O nome identificador utilizado é X-ADES e os campos separados pelo símbolo @. O primeiro campo é obrigatoriamente o endereço IP da máquina que enviou a mensagem e em seguida outros campos que definem o resultado das técnicas *antispam*.

Para cada técnica, o primeiro campo é o código do nome dessa técnica e o segundo campo contém os resultados. A figura 10 ilustra um exemplo da linha que é adicionada ao cabeçalho das mensagens:

```
X-ADES-ANALYSIS:
@200.212.106.85@DNSBL@OK@DNSREVERSO@FALHOU@SPF@OK@C
ONTEUDO@OK@BAYSIANO@FALHOU@ASSINATURAS@OK@GREYLISTI
NG@FALHOU
```

Figura 10. Exemplo do cabeçalho de mensagem alterado
Fonte: O autor

O módulo de análise das mensagens gera arquivos texto, com as informações de cada mensagem, além de armazenar todas as linhas no banco de dados. No módulo de análise de dados, as informações contidas nos arquivos texto são classificadas e agrupadas, permitindo visualizar as características mais comuns em cada tipo de mensagem, legítimas e *spam*s.

Após realizar todos os testes, o banco de dados *MySQL* possuirá todos os resultados armazenados, podendo ser facilmente verificado através de consultas em SQL (structured query language).

4 ANÁLISE INDIVIDUAL DAS TÉCNICAS E DISCUSSÕES

Nesta seção são apresentadas os resultados do experimento obtido no teste descrito na seção anterior. Os resultados são analisados e discutidos, separados por técnica *antispam*, considerando a retenção dos falsos positivos e falsos negativos.

4.1 Resultados gerais

Os resultados apresentados nessa seção correspondem ao conjunto de mensagens que foram testadas simultaneamente em todas as técnicas *antispam*. Todas as técnicas *antispam* apresentam retenções para mensagens *spam* e não *spam*. A figura 11 mostra a retenção dos falsos positivos.

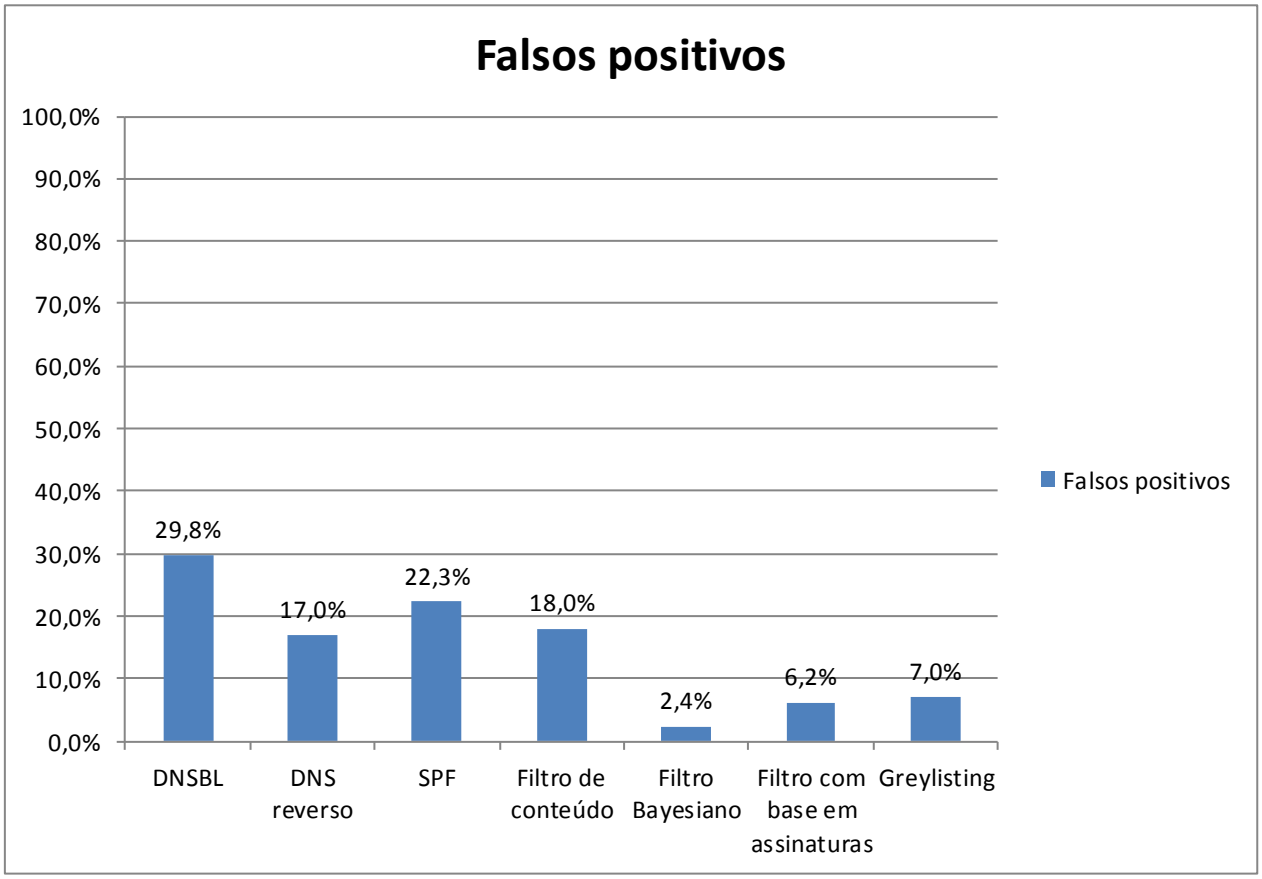

Figura 11. Gráfico com os resultados por taxa de falsos positivos
Fonte: O autor

De um modo geral, as taxas de falsos positivos foram consideravelmente altas se for comparado com as recomendações de *Osterman Research* (2011). A taxa de falsos positivos para os mecanismos de DNSBL, SPF, filtro de conteúdo e DNS reverso foram as maiores, chegando a 29,1% das mensagens para o mecanismo DNSBL.

A figura 12 apresenta os resultados dos falsos negativos.

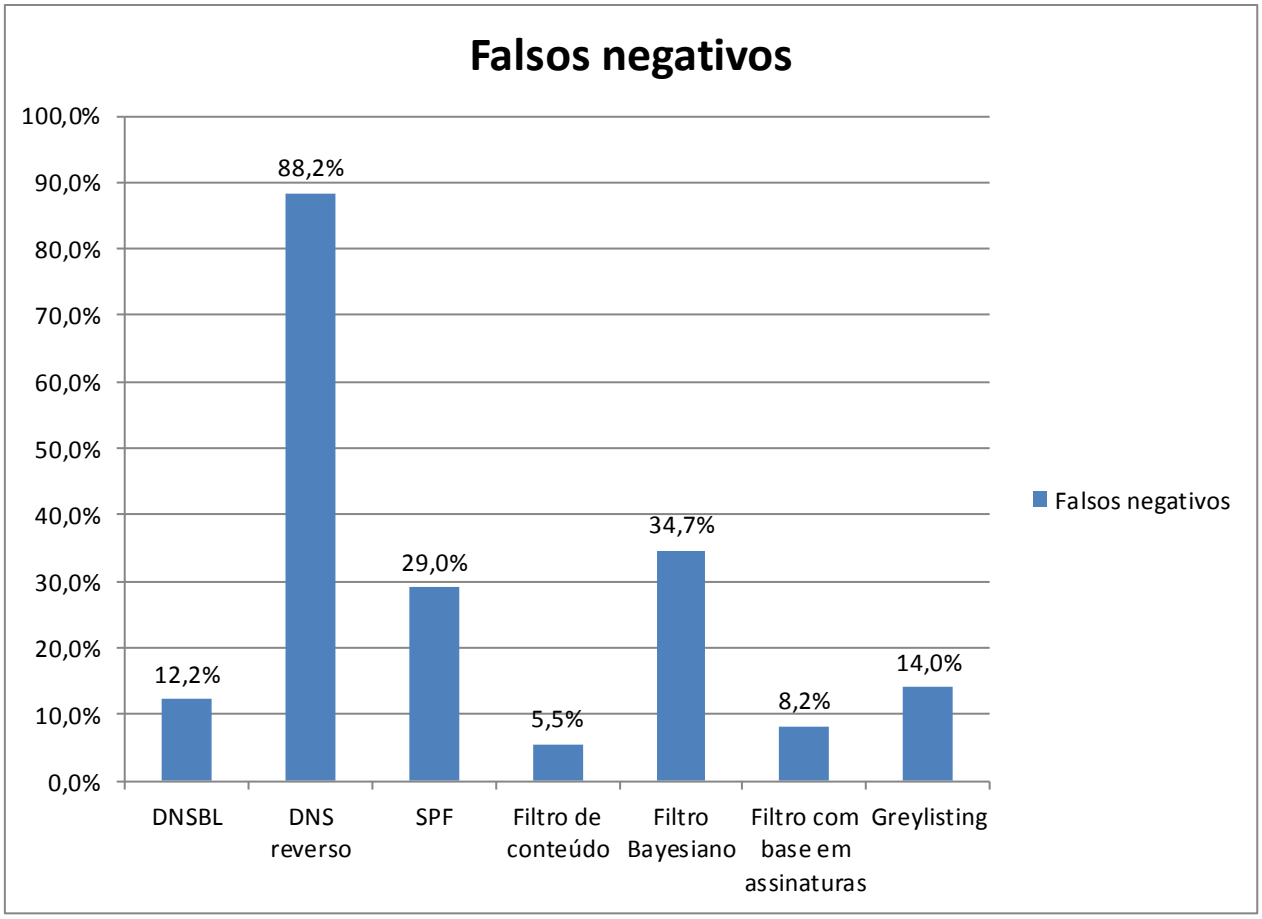

Figura 12. Gráfico com os resultados por taxa de falsos negativos
Fonte: O autor

Apesar dos falsos negativos possuírem um impacto menos prejudicial ao usuário, fica evidente a ineficiência das técnicas de filtro *Bayesiano*, SPF e DNS reverso, que resultou em 88,2% de falsos negativos. As técnicas de filtro de conteúdo e filtro com base em assinaturas tiveram os melhores desempenhos no teste.

4.2 DNSBL

Existem diferentes critérios para adicionar endereços IP em diferentes listas negras de DNS (DNSBL). Alguns *spammers* selecionam servidores com falhas de segurança e adicionam URLs (*Uniform Resource Locator)* conhecidas para enviar *spam*. Para os usuários que enviarem *e-mails*, mesmo que verdadeiros, através de um servidor vulnerável, frequentemente terão suas mensagens rejeitadas, consideradas como *spam* ou excluídas.

A Figura 13 mostra os resultados na retenção dos falsos positivos e falsos negativos.

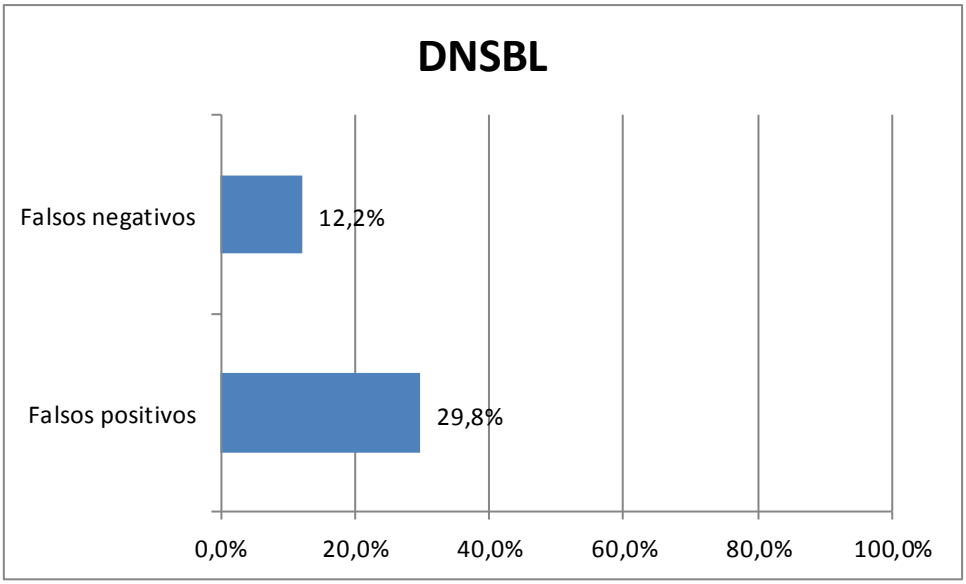

Figura 13. Gráfico com os resultados da técnica DNSBL
Fonte: O autor

Uma configuração mais detalhada e específica das listas negras poderia diminuir essas taxas, entretanto, a constante mudança que os *spammers* encontram para “escravizar” servidores vulneráveis e usá-los para enviar *spam*, faz com que as listas precisem ser atualizadas o tempo todo, além de poderem enviar códigos para que outros computadores sejam infectados tornando-se também um *relay* de *spam*. Na versão IPV6 (Internet Protocol version 6), que hoje ainda é pouco utilizada na Internet, o espaço de endereço será muito maior, possibilitando os *spammers* serem capazes de alterar os endereços de *relay* mais frequentemente.

4.3 DNS Reverso

A taxa de falsos positivos para a técnica *antispam* de consulta ao DNS reverso mostra que muitas mensagens legítimas provêm de servidores legítimos com o DNS reverso mal configurado. A figura 14 mostra os resultados para essa técnica.

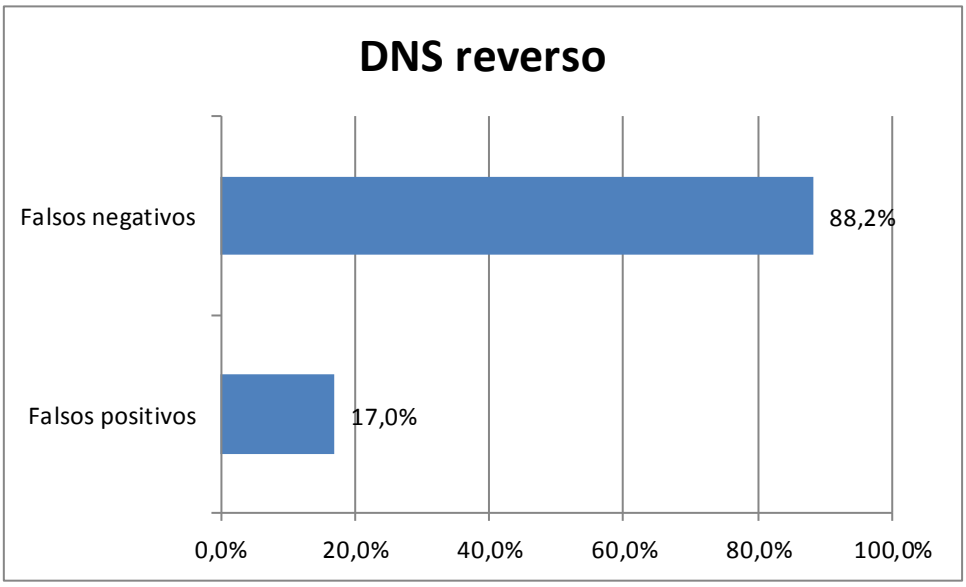

Figura 14. Gráfico com os resultados da técnica DNS reverso
Fonte: O autor

Já a taxa de falsos negativos é alta, chegando a 88,2%, mostrando a ineficiência dessa técnica. A alta taxa de falsos negativos para o DNS reverso é causada por *spammers* que invadem máquinas vulneráveis com o DNS reverso configurado corretamente para enviar *spam*s.

A maioria dos agentes SMTP usa a verificação de DNS reverso para confirmar a veracidade do domínio destino. Os endereços que são provenientes de conexões dial-up ou endereços atribuídos dinamicamente por provedores de acesso são mais suscetíveis a enviar *spam* devido ao alto índice de endereços com o DNS reverso mal configurado ou sem configuração definida.

A verificação do DNS reverso, uma vez configurado corretamente, pode criar uma autenticação válida entre o provedor que fornece a conexão para

Internet e o usuário da rede que recebe o IP para o acesso, e essa autenticação pode ser segura o suficiente para que o domínio de acesso seja inserido em *whitelists* conhecidas. Isso dificulta a ação de *spammers* já que, geralmente, não se pode ignorar esta verificação quando é usada na tentativa de forjar os domínios.

4.4 SPF

As taxas de falsos positivos e falsos negativos para a técnica SPF foram bem semelhantes, diferentemente da discrepância de outras técnicas testadas. A figura 15 apresenta os resultados.

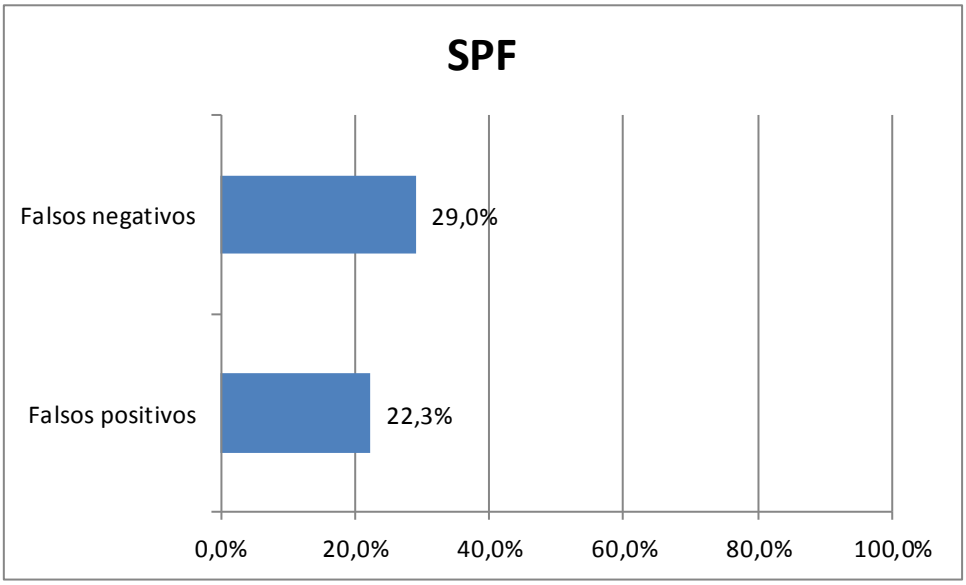

Figura 15. Gráfico com os resultados da técnica SPF
Fonte: O autor

Dentre as razões para justificar os resultados, a principal delas é que essa técnica ainda não é um padrão de comunicação na Internet. Há também a possibilidade de que os filtros de *e-mails* considerados válidos são marcados como *spam* devido à má configuração da aplicação, que nesse experimento utilizou as configurações padrões de instalação.

As técnicas que utilizam padrões de autenticação do remetente, como o SPF, provavelmente vão atenuar consideravelmente a disseminação de *spam* e ajudarão a impedir ataques como fraude e phishing. Entretanto, os padrões

precisam ser melhores definidos e difundidos, pois, hoje os *spammers* ainda conseguem contorná-los.

4.5 Filtro *Bayesiano*

A técnica de filtro *bayesiano* teve a menor taxa de falsos positivos entre todas as técnicas testadas, em contrapartida, a taxa de falsos negativos foi alta. A figura 16 mostra os resultados.

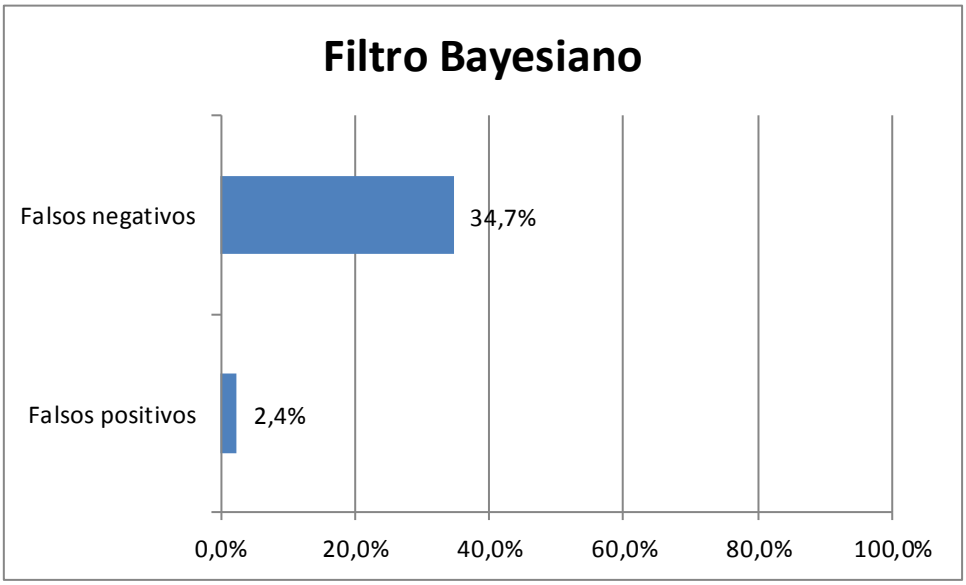

Figura 16. Gráfico com os resultados da técnica filtro *Bayesiano*
Fonte: O autor

Devido à lógica que a técnica utiliza para classificar uma mensagem como *spam* e o treinamento efetuado na ferramenta com mensagens pré-selecionadas, a técnica obteve um número muito baixo de falsos positivos. Para que o filtro pudesse funcionar corretamente foi necessário "treinar" a técnica para que a lógica de probabilidade ficasse atualizada utilizando uma amostragem de mensagens lícitas e *spam*s selecionados previamente.

Diferentemente das outras técnicas, o filtro *bayesiano* depende de aprendizagem, que pode não ser exata, para classificar uma mensagem, e como esse processo é estatístico, terá sempre o risco, mesmo que

gradativamente menor, de classificação incorreta ocasionando mais falsos negativos, que ficou evidente no resultado do teste.

Spammers tentam driblar a classificação de conteúdo, introduzindo no meio das palavras do texto, caracteres especiais ou letras a mais, visando dificultar a determinação de padrões automáticos nas estatísticas, sem prejudicar muito a legibilidade do texto.

4.6 *Greylisting*

Com o conceito diferente das outras técnicas testadas, que consiste em recusar temporariamente uma mensagem e esperar por sua retransmissão, a técnica *greylisting* obteve os resultados apresentados na figura 17.

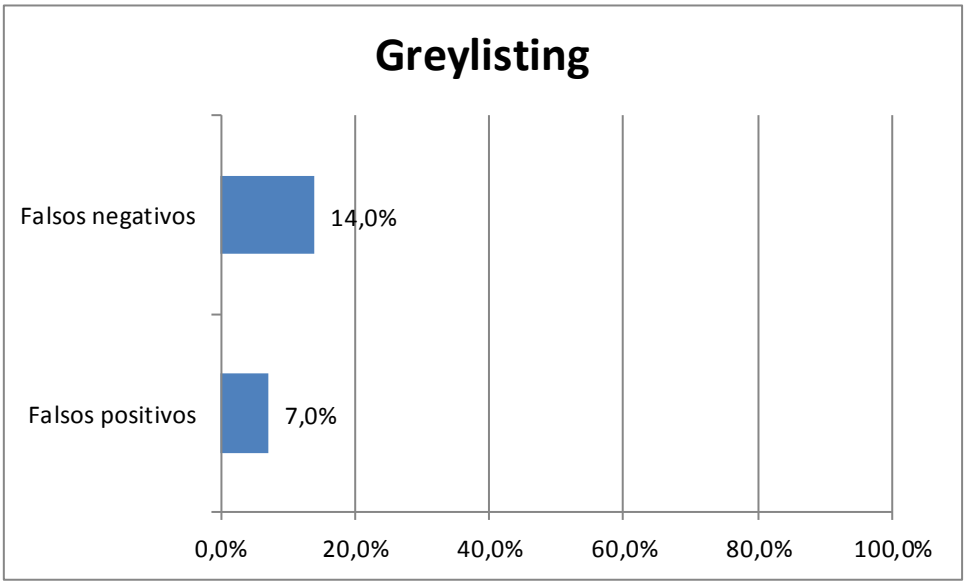

Figura 17. Gráfico com os resultados da técnica *Greylisting*
Fonte: O autor

A técnica considera que *e-mails* válidos são enviados a partir de MTAs legítimos, que mantém filas e possuem políticas de retransmissão em caso de erros temporários. Porém, *spammers* usam códigos maliciosos que, nem sempre, usam MTAs legítimos para envio. Através dos testes, ficou evidente a utilização de MTAs legítimos, certamente, com o objetivo de contornar esta técnica. Ainda assim, a técnica *greylisting* mostrou-se mais eficiente na retenção dos falsos positivos. Uma alternativa para diminuir os falsos negativos

seria manter em uma *whitelist*, endereços IP que tem passagem livre pelo *greylisting* porque são máquinas confiáveis ou porque seus MTAs não conseguem tratar corretamente erros temporários.

A eficácia da técnica depende do quanto ela será usada na Internet. Se grande parte dos servidores SMTP estiver com a técnica *greylisting* configurada, provavelmente o custo dos *spammers* para contorná-la será maior. Mas se o uso for comedido, eles provavelmente irão usar os mesmos métodos para o envio de *spam* como é feito hoje.

4.7 Filtro de conteúdo

A técnica de filtro de conteúdo teve um bom desempenho, principalmente na retenção dos falsos negativos, como mostra a figura 18.

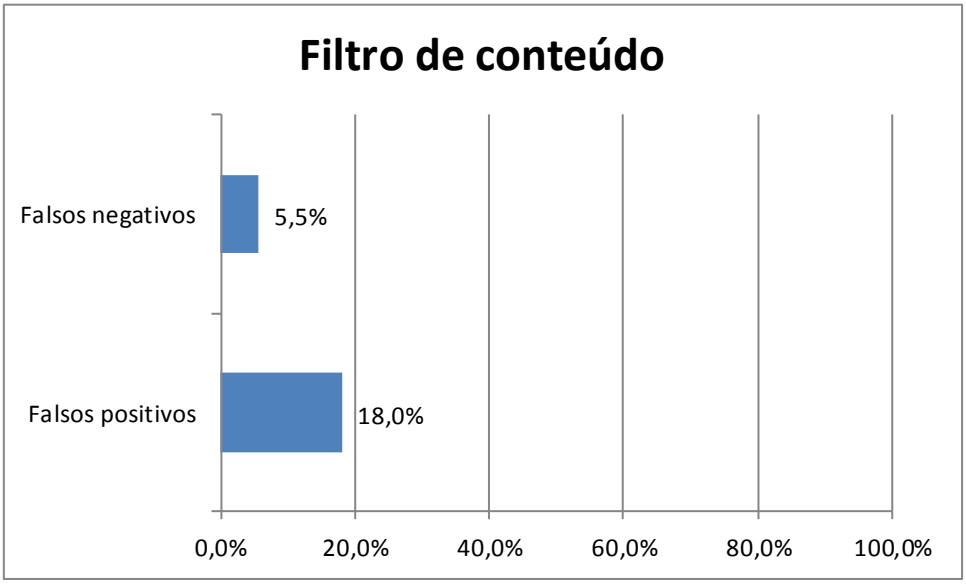

Figura 18. Gráfico com os resultados da técnica filtro de conteúdo
Fonte: O autor

Os resultados foram considerados surpreendentes, pois o arquivo da lista de conteúdo, que foi utilizado no teste, foi o padrão na instalação do *spamassassin*, sem nenhuma atualização de sites que publicam listas de conteúdos para atualização. Ainda assim, técnicas que utilizam regras estáticas para bloqueios de *spam* geralmente são fáceis de serem burladas por *spammers.*

A manutenção e atualização das regras estáticas do filtro com base em conteúdo é uma tarefa que, se for realizada de forma rotineira, ainda pode ser eficiente para determinados cenários de retenção de *spam*, tanto falsos negativos quanto falsos positivos.

4.8 Filtro com base em assinaturas

Os resultados na retenção dos falsos positivos e falsos negativos no teste para a técnica de filtro com base em assinaturas, apresentados na figura 19, foi similar aos resultados da técnica de filtro de conteúdo.

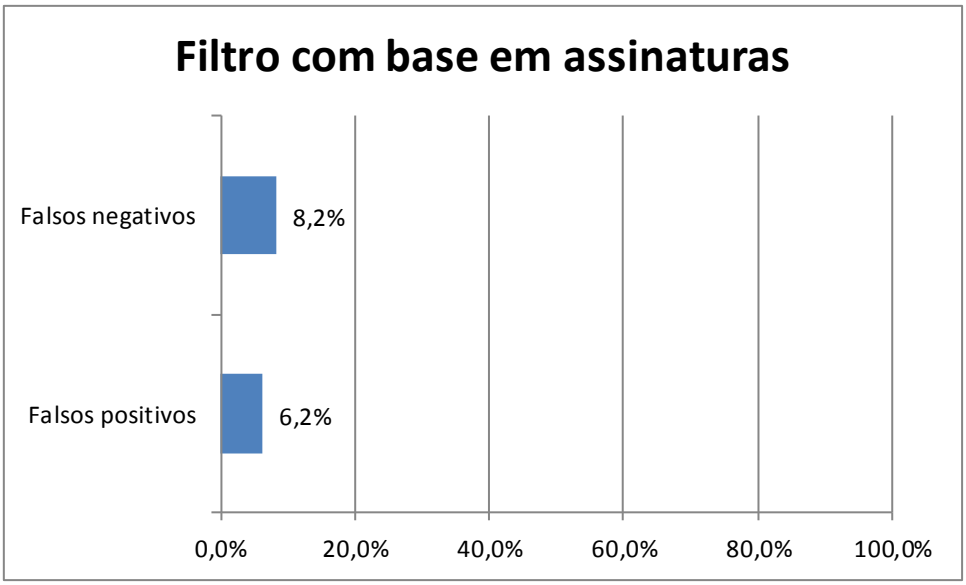

Figura 19. Gráfico com os resultados da técnica filtro com base em assinaturas
Fonte: O autor

Essa técnica também foi configurada com as configurações padrões do *spamassassin*, portanto, era esperado um desempenho pior em comparação as outras técnicas testadas, já que para manter uma eficiência aceitável essa técnica precisa constantemente manter a base de catálogo das assinaturas atualizadas, mas os resultados mostraram que mesmo com o catálogo padrão foram melhores que a maioria das outras técnicas.

5 ANÁLISE CONJUNTA DAS TÉCNICAS E DISCUSSÕES

Esta seção apresenta e discute dois cenários, uma seleção conjunta para retenção de falsos positivos e outra para falsos negativos, com testes realizados utilizando as três técnicas que tiveram as melhores eficácias individualmente para cada tipo de retenção analisadas na seção anterior.

5.1 Configuração do ambiente e teste

Com base na análise de resultados da seção 4, são selecionadas as três técnicas que tiveram o melhor desempenho para retenção de falsos positivos: filtro *bayesiano*, filtro com base em assinaturas e *greylisting*. Para os falsos negativos as três técnicas mais eficientes foram: DNSBL, filtro com base em assinaturas e filtro de conteúdo.

Essas técnicas são configuradas na ferramenta *GNU* ADES e o teste é refeito usando sempre a mesma condição de configuração padrão e base de *e-mails spam* e não *spam* utilizado na realização dos testes individuais. Nestes dois cenários, apenas um retorno “OK” de um das técnicas é necessário para acusar um falso positivo ou um falso negativo. A figura 20 apresenta um exemplo para essa condição.

X-ADES-ANALYSIS:@200.221.2.45
@BAYSIANO@OK@ASSINATURAS@FALHO@*GREYLISTING*@FALHOU

Figura 20. Exemplo de teste para análise conjunta de técnicas
Fonte: O autor

Nesse exemplo, supondo que a mensagem não é um spam, será considerado um falso positivo devido à técnica de filtro bayesiano acertar no teste, mesmo que as outras duas falharem ela não será entregue na caixa postal. Se essa mesma mensagem for um spam, será barrada e não irá gerar um falso negativo.

5.2 Testes para falsos positivos

No cenário para testar as técnicas com maior desempenho na retenção dos falsos positivos, o laboratório é configurado apenas com as três técnicas e os testes são realizados.

A tabela 1 apresenta os resultados combinados, considerando: **A** para filtro *bayesiano*, **B** para filtro com base em assinaturas e **C** para *greylisting*.

Tabela 1 – Resultados para retenção de falsos positivos

Técnicas *antispam*	Resultados
A+B	**1,8%**
A+C	**2,2%**
B+C	**4,3%**

Fonte: O autor

Os resultados mostram que para todos os testes ocorreram reduções na taxa de falsos positivos, mas apenas a combinação entre filtro *bayesiano* e filtro com base em assinaturas conseguiu um desempenho melhor que uma técnica, o próprio filtro *bayesiano* com 2,2%, nos testes individuais.

Mesmo assim, se considerarmos que a melhor taxa de retenção dos *softwares* testadas no artigo da *Osterman Research* (2011) foi 0,15% de falsos positivos, ou 1,5 para cada 1000 *e-mails,* o resultado está distante do esperado.

5.3 Testes para falsos negativos

Para os testes dos falsos negativos o laboratório também é modificado para validar as três melhores técnicas e os testes são realizados. A tabela 2 mostra os resultados combinados, considerando: **A** para DNSBL, **B** para filtro com base em assinaturas e **C** para filtro de conteúdo.

Tabela 2 – Resultados para retenção de falsos negativos

Técnicas *antispam*	Resultados
A+B	**6,5%**
A+C	**5,2%**
B+C	**2,9%**

Fonte: O autor

Assim como os testes de falsos positivos, os testes de falsos negativos também tiveram resultados combinados com reduções em comparação às técnicas individuais, entretanto, apenas o cenário B+C obteve um desempenho mais eficiente do que o melhor resultado de uma técnica individual, que foi o filtro de conteúdo com 5,5%.

A combinação das técnicas filtro de conteúdo e filtro com base em assinaturas teve o melhor desempenho com 2,9% de falsos negativos gerados. Entretanto, em comparação com ferramentas de mercado (SNYDER, 2012), onde o melhor software teve um desempenho de 1,94%, o resultado também não é satisfatório.

5.4 Validade e confiabilidade do teste

Validade é definida como "na medida em que qualquer instrumento de medição mede o que é destinado a medir" (CARMINES, 1979). A validade desta experiência em relação à redução de *spam*, na caixa de correio do usuário final, é reduzida, pois alguns *spammers* utilizam endereços IP aleatórios para o envio de *spam* e o servidor do experimento não tem acesso aos endereços válidos, já que as mensagens que são consideradas *spam*, utilizadas no teste, foram retiradas de um repositório próprio para esse tipo de experimento. Bugs no software também podem causar redução na validade interna. Modificações no software relacionado ao experimento também são susceptíveis de ter mais bugs do que o software original.

Confiabilidade é definida como "à medida que um experimento, teste ou qualquer tipo de procedimento de medição produz os mesmos resultados nos testes repetidos" (CARMINES, 1979). O mesmo experimento pode ser

realizado por alguém com conhecimento e acesso ao software e suas configurações. Com o acesso a um ambiente similar pode-se realizar o mesmo experimento e, se for dentro de um prazo razoável, resultados semelhantes podem ser obtidos. Com o passar do tempo, os *spammers* e as técnicas *antispam* irão aprimorar e desenvolver métodos mais avançados para retenção de *spam* e o conteúdo de uma experiência com base nessa metodologia deve refletir isso.

6 CONCLUSÕES

Neste trabalho diferentes técnicas *antispam* foram estudadas com o objetivo de analisar a eficácia na retenção de mensagens que são classificadas como falsos positivos e falsos negativos. Foi projetado e implementado um laboratório de teste que permite a execução e medição individual de cada técnica *antispam* a partir de dados reais e fazer uma comparação para identificar quais técnicas possuem uma melhor eficácia.

Não existe uma combinação perfeita de técnicas que gere uma taxa de 0% de *spam*. Quanto mais perto de 0% de *spam* retido, maior será o número de falsos positivos gerados.

Os resultados individuais mostraram uma taxa alta de falsos negativos entre 5,5% e 88,2%. Também foi comprovada uma taxa de falso positivos acima de 2,4% para todos os mecanismos, o que é alto considerando o impacto negativo que um falso positivo pode causar para os usuários. O mecanismo de filtro *bayesiano* obteve o melhor resultado para os falsos positivos, com 2,4% e a técnica de filtro de conteúdo, com 5,5%, foi a melhor na retenção dos falsos negativos.

O laboratório foi modificado para realizar testes de combinações das técnicas *antispam* que tiveram os melhores desempenhos individuais. Para retenção de falsos positivos, o melhor resultado foi de 1,8% na combinação entre as técnicas de filtro com base em assinaturas e filtro *bayesiano* e 2,9% para os falsos negativos com a combinação entre as técnicas de filtro com base em assinaturas e de filtro de conteúdo. Mesmo com desempenho melhor, os testes não conseguiram superar as taxas de 0,15% (*Osterman Research*, 2011) para os falsos positivos e 1,94% (*Snyder*, 2012) para os falsos negativos, mostrando que somente configurando uma ferramenta com técnicas que possuem uma boa eficácia na retenção de falsos positivos ou falsos negativos ela não foi suficiente para chegar aos resultados esperados.

6.1 Considerações finais

Hoje, não há nenhum indício que permita inferir que a atividade de enviar *spam*s diminuirá nos próximos anos. Ao contrário, os *spammers* vêm se especializando e usando técnicas cada vez mais elaboradas para burlar sistemas *antispam*. Os *spammers* estão constantemente evoluindo, tentando adaptar-se aos novos mecanismos *antispam*, fazendo com que as técnicas também evoluam. Assim, um ambiente que permita avaliar o desempenho das técnicas *antispam* individualmente pode orientar na criação de mecanismos *antispam* mais eficientes focando na retenção dos falsos positivos ou dos falsos negativos.

Este trabalho possui um forte componente de obsolescência, pois, na área de Segurança da Informação (SI) as técnicas *antispam* e os *spammers* estão em constante evolução, portanto, recomenda-se refazer os testes periodicamente.

6.2 Trabalhos futuros

A continuidade desse trabalho pode envolver um estudo mais detalhado de cada técnica abordada e de seus algoritmos, utilizando bases de dados temporais para avaliar a evolução do desempenho das técnicas *antispam* ao longo do tempo.

Poderia ser feito também um estudo do uso de recursos de *hardware* de cada técnica, para estimar o limite máximo de sua implantação.

O uso de uma alternativa ao protocolo SMTP que garantisse um maior controle da autenticidade da mensagem seria interessante e poderia dar origem a uma nova proposta de funcionamento do protocolo SMTP.

REFERÊNCIAS

BRUCE, G. ***Spam* Archive.** Disponível em: <http://untroubled.org/*spam*>. Acesso em: 19 set. 2011.

CARMINES, Edward. **Quantitative Applications in the Social Sciences.** Londres: Sage Publications, 1979. 17 v.

CHESWICK, W. R.; BELLOVIN, S. M.; RUBIN, A. D. **Firewalls e Segurança na Internet**: Repelindo o hacker ardiloso. 2.ed. Porto Alegre: Bookman, 2005.

DAVID, E. S. ***Spam* Laws.** Disponível em: <http://www.*spam*laws.com/*spam*-stats.html>. Acesso em: 24 set. 2011.

DEBIAN *Linux* Disponível em: <http://www.debian.org/>. Acesso em: 04 fev. 2013.

FABRE, R. C. **Métodos avançados para controle de *spam*.** 2005. Dissertação (Mestrado) - Laboratório de Administração e Segurança de Sistemas, Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2005.

GENTOO *Linux* Disponível em: <http://www.gentoo.org/>. Acesso em: 04 fev. 2013.

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

GÖRLING S., **An overview of the Sender Policy Framework (SPF) as an anti-phishing mechanism**, Vol. 17 Iss: 2, pp.169 – 179, 2007

J. R. LEVINE, **Experiences with *greylisting***, CEAS 2005: Second Conference on *E-mail* and *Antispam*, 2005

LAFRAIA, D. Disponível em: <http://www.lafraia.com.br/spambr/>. Acesso em: 04 fev. 2013.

MYSQL Disponível em: <http://www.mysql.com/>. Acesso em: 04 fev. 2013.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (Org.). ***Antispam*.br.** Disponível em: <http://www.*antispam*.br/>. Acesso em: 18 set. 2011.

OSTERMAN RESEARCH (Usa) (Org.). **Why You Need to Eliminate False Positives in Your *E-mail* System.** Washington, 2011. 14 p. Disponível em: <http://www.trustsphere.com/index.php?option=com_content&view=article&id=184:view-webinar-by-osterman-why-you-need-to-eliminate-false-positives-in-your-*e-mail*-system&catid=49:press-liner-newsflash-blue-coin>. Acesso em: 11 out. 2011.

PFLEEGER, S. L. E BLOOM, G. (2005). **Canning *spam*: Proposed solutions to unwanted *e-mail*.** IEEE Security & Privacy Magazine, 3(2):40–47.

RAAD, M., YEASSEN, A., Gazi, M. **Impact of *spam* advertisement through *e-mail*: A study to assess the influence of the *antispam* on the *e-mail* marketing**. **African Journal of Business Management**: Academic Journals, v. 11, n.4, p.2362-2367, 04 set. 2010.

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.

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.

SAHAMI M., **Learning Limited Dependence Bayesian Classifiers**, In KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pág. 335-338, Menlo Park, CA: AAAI Press, 1996.

SNYDER, Joel. **Comparative Performance of *Antispam* Gateways.** Disponível em: <http://www.opus1.com/www/whitepapers/*antispam*dec2011.pdf>. Acesso em: 04 abr. 2013.

STEDING, K. **Uso de *Honeypot*s para o Estudo de *Spam* e Phishing.** 2008. 204 f. Tese (Doutorado) - Inpe, São José Dos Campos, 2008.

TAVEIRA, Danilo Michalczuk. **Mecanismo *antispam* baseado em autenticação e reputação**. 2008. 97 f. Dissertação (Mestrado) - Universidade Federal do Rio, Rio de Janeiro, 2008.

TIPTON, H. F.; KRAUSE, M. **Information security management handbook**.5. ed. Danvers: Auerbach, 2004

WANROOIJ, W. V., PRAS, A. **Filtering *spam* from *bad neighborhoods*.** International Journal Of Network Management, EUA, n., p.433-434, 15 out. 2010.

WITTEL, G. L., E WU, S. F. **On attacking statistical *spam* filters.** Em First Conference on *E-mail* and *Antispam* (CEAS2004) (julho de 2004)

WIEHE, Anders; HJELMAS, Erik; WOLTHUSEN, Stephen D.. **Quantitative Analysis of Ecient *Antispam* Techniques.** West Point: IEEE Press, 2006. 7 p.

ANEXO A

Neste anexo é mostrado o modo como cada uma das ferramentas foram instaladas. São mostrados os scripts usados na instalação e configuração das ferramentas *spamassassin*, *GNU* ADES e o script para instalar a base no banco de dados.

A.1 Instalação do *spamassassin*

A filtragem *antispam* é feita pelo software *spamassassin*, que pode ser instalado através dos seguintes comandos, com as configurações padrões:

```
$ su
$ apt-get update
$ apt-get install spamassassin
```

A.2 Instalação do módulo de análise do *GNU* ADES

Alguns pacotes do *Debian* devem ser instalados antes do sistema *GNU* ADES. Para instalá-los, use os seguintes comandos:

```
$ su
$ apt-get update
$ apt-get install apache2.0 php5 mysql-server python2.5 pythonmysqldb
```

Faça o download do sistema na seguinte *URL*: http://www.gta.ufrj.br/ades. Após o download, para instalar a ferramenta *GNU* ADES, deve-se digitar os seguintes comandos no terminal.

```
$ tar xzvf ades.tar.gz
$ cp ades/ /var/www
$ cp ades/files/policy.pl /etc/postix
```

Na instalação do *honeypot* deve-se substituir <PASTA> pela localização do diretório no qual se encontra o sítio que será usado para a divulgação dos *e-mails* do *honeypot*.

$ cp ades/*honeypot_e-mail*.php <PASTA>

Após executar os comandos, o sistema já está instalado, porém ainda não está configurado. A primeira parte da configuração é alterar dois arquivos do *Postfix*:

$ nano /etc/postfix/master.cf

Insira a seguinte linha ao fim do arquivo:

policy unix - n n - - spawnuser=no body argv=/usr/bin/perl /etc/postfix/policy.pl

A seguir, edite o outro arquivo de configuração do *Postfix*:

$ nano /etc/postfix/main.cf

Na seção *smtpd_recipient_restrictions* insira a seguinte linha:

check_policy_service unix:private/policy

Para que o *honeypot* divulgue os endereços de *e-mail*, adicione as seguintes linhas a um página PHP, que seja acessada frequentemente, como a página principal:

```
<?php
require("honey_pot_e-mail.php");
honey_pot_e-mail();
honey_pot_e-mail();
honey_pot_e-mail();
```

honey_pot_*e-mail*();
honey_pot_*e-mail*();
?>

Altere o arquivo honey_pot_*e-mail*.php para que as variáveis $usuario_bd, $senha_bd e $servidor_bd apontem respectivamente para o usuário do banco de dados, senha e qual o servidor do banco de dados. No diretório files também são encontrados os arquivos de backup do banco de dados que devem ser importados durante a instalação do sistema:

$ cd /var/www/ades/files
$ mysql -u <usuario> -p < *honeypot*_db.sql
$ mysql -u <usuario> -p < language_db.sql

O <usuario> é o nome de usuário do banco de dados M*ysql*. O próximo procedimento é alterar os arquivos do *GNU* ADES:

$ cd /var/www/ades
$ nano system/const.pl

Nesse arquivo devem-se alterar as variáveis $user_base_dir e $*honeypot*_base_dir, que devem apontar respectivamente para diretório base dos usuários e o usuário base do *honeypot*.

$ nano usuarios-participando

Esse arquivo contém todos os usuários que estão participando do sistema, ele deve ser preenchido com o complemento do endereço da *home* do usuário em relação ao caminho estabelecido no const.pl.

O ultimo arquivo a ser alterado é o ades.py.

$ nano /system/ades.py

As variáveis servidor_bd, usuário e senha, que devem ser preenchidas com o endereço do servidor de banco de dados, o usuário do banco de dados e a sua senha, respectivamente.

Para executar a verificação de todos os usuários participantes basta executar o seguinte comando como *root* (super usuário):

```
$ /var/www/ades/system/run
```

Para manter o sistema sempre com os dados mais atuais dos usuários, esse script de execução pode ser inserido no cron:

```
$ nano /etc/crontab
```

Inserir a seguinte linha ao fim do arquivo:

```
00 3 *** root cd /var/www/ades/system && ./run
```

Essa linha executa o Sistema ADES todos os dias às 3:00 hs.

A.3 Script do banco de dados

Script para criar a base de dados:

```
-- Database: `honeypot`
--
-- --------------------------------------------------------
--
-- Table structure for table `Info`
--
CREATE TABLE `Info` (
  `Cd_Data` int(11) NOT NULL auto_increment,
  `Cd_User` int(11) NOT NULL default '0',
  `Nm_Agente` varchar(200) NOT NULL default '',
```

```
  `Nm_Refer` varchar(100) NOT NULL default '',
  `Nu_IP` varchar(100) NOT NULL default '',
  `Nm_Pagina` varchar(255) NOT NULL default '',
  `Dt_Inclusao` timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
  PRIMARY KEY  (`Cd_Data`)
) ENGINE=MyISAM AUTO_INCREMENT=391767 DEFAULT CHARSET=latin1 AUTO_INCREMENT=391767 ;
-- --------------------------------------------------------
--
-- Table structure for table `User`
--
CREATE TABLE `User` (
  `Cd_User` int(11) NOT NULL auto_increment,
  `Nm_User` varchar(100) NOT NULL default '',
  `Nm_E-mail` varchar(100) NOT NULL default '',
  PRIMARY KEY  (`Cd_User`),
  KEY `Nm_E-mail` (`Nm_E-mail`)
) ENGINE=MyISAM AUTO_INCREMENT=391768 DEFAULT CHARSET=latin1 AUTO_INCREMENT=391768 ;
```