Ferramenta para Análise de Características de *Spams* e Mecanismos Anti-*Spam*

Danilo Michalczuk Taveira[1], **Diogo Menezes Ferrazani Mattos**[1] **e Otto Carlos Muniz Bandeira Duarte**[1]*†

[1]GTA/COPPE/UFRJ
http://www.gta.ufrj.br

Abstract. *Sending spam is a profitable activity for spammers that is constantly rising. This paper presents the implementation of a tool called ADES which analyze the process of sending the messages and also commonly used anti-spam mechanisms. The results show the low efficiency of current anti-spam mechanisms and important characteristics of the harvest and email sending processes.*

Resumo. *O envio dos spams é uma atividade em constante crescimento e que se torna cada vez mais lucrativa para os spammers. Neste artigo, é apresentada a implementação da ferramenta ADES (***A***nálise* **DE S***pam), que permite avaliar tanto o processo de envio de spams, quanto os mecanismos anti-spam. Os resultados das análises revelam não somente a baixa eficiência dos mecanismos analisados de combate aos spams, como também importantes características do processo de envio de spams.*

1. Introdução

O número de mensagens não solicitadas, denominadas *spams*, já ultrapassou o número de mensagens legítimas, representando 68,7% de todas as mensagens que circulam na Internet. Os prejuízos financeiros causados por esta atividade alcançam cifras extraordinárias devido ao consumo de recursos como banda passante, memória e processamento e aos gastos na aquisição de programas anti-*spams*. Além disso, os *spams* afetam diretamente os destinatários, que se sentem cada vez mais insatisfeitos por perderem tempo na recepção e na leitura das mensagens, como também pela possibilidade de disseminação de vírus e de outros programas que causam a perda de dados e comprometem a segurança. A maior motivação para o envio de *spams* está relacionada ao retorno financeiro, já que esta atividade é eficaz na divulgação de produtos e serviços, podendo até mesmo ser utilizada para o enriquecimento ilícito através de golpes ou tentativas de estelionato [Taveira et al., 2006].

A avaliação do desempenho dos mecanismos anti-*spams* é imprescindível no combate aos *spams*, para avaliar a efetividade das soluções anti-*spam* adotadas. Os *spammers* constantemente mudam as técnicas utilizadas para enviar os *spams*, com o objetivo de burlar os sistemas anti-*spam*. Dessa forma, é importante tanto avaliar o desempenho dos mecanismos anti-*spam* como avaliar o processo utilizado pelos *spammers* para enviar as mensagens. Nesse trabalho é apresentada uma ferramenta que analisa o desempenho de diveros mecanismos anti-*spam* atuais, medindo os falso-positivos e os falso-negativos. Além disso, a ferramenta desenvolvida também analisa características do processo utilizado pelos *spammers* para enviar as mensagens. A arquitetura da ferramenta ADES é

*Este trabalho foi realizado com recursos do CNPq, FINEP, RNP, FAPERJ e CAPES
†A ferramenta pode ser avaliada em http://www.gta.ufrj.br/ades

modular e flexível, permitindo a análise de diferentes características das mensagens, do processo de envio de *spams* e dos mecanismos anti-*spam*.

Este trabalho está organizado da seguinte forma. Na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 é apresentada a ferramenta de avaliação dos mecanismos anti-*spam* e do comportamento dos *spammers* denominada ADES. Detalhes referentes aos testes e à análise dos resultados são abordados na Seção 4. Por fim, na Seção 5 são apresentadas as conclusões deste trabalho.

2. Trabalhos Relacionados

Gomes *et al.* analisam características do tráfego gerado pelos *spams*, como a distribuição temporal das mensagens *spams* e legítimas, o tamanho das mensagens, o número de destinatários e remetentes, entre outras [Gomes et al., 2004]. Entretanto, resultados como a distribuição do tamanho das mensagens se modificaram devido à rápida evolução das estratégias dos *spammers* em burlar os sistemas anti-*spams*.

Na proposta de Andreolini *et al.* endereços eletrônicos gerados aleatoriamente são divulgados com o objetivo de contaminar as listas dos *spammers*, fazendo com que eles enviem mensagens para endereços que não pertencem a usuários legítimos [Andreolini et al., 2005]. Entretanto, não são feitas análises em relação ao tempo entre a coleta e o envio das mensagens, o tempo de permanência dos endereços divulgados nas listas dos *spammers* e se os processos de coleta e envio são realizados por uma mesma máquina ou por máquinas diferentes.

Ramachandran e Feamster mostram resultados a respeito das características dos *spams* e análises sobre as técnicas utilizadas pelos *spammers* para enviar as mensagens [Ramachandran e Feamster, 2006]. Atualmente, a técnica mais comum para o envio de *spams* utiliza máquinas alheias controladas por *spammers*, também chamadas de máquinas zumbis.

3. Sistema de Análise de *Spam* - ADES

A grande parte dos sistemas anti-*spam* existente avalia as mensagens, as classificam e as descartam quando são consideradas *spams*. Dessa forma, por causa do descarte da mensagem, não é possível avaliar a eficiência de cada um dos sistemas individualmente para cada mensagem recebida. A ferramenta ADES (**A**nálise **DE S**pam) permite realizar a análise da eficiência dos mecanismos anti-*spam* de forma individualizada e também fornece características do processo de coleta de endereços e envio de *spams*. No ADES, cada mensagem é analisada por cada um dos mecanismos anti-*spam* que são avaliados de forma totalmente isolada. A partir da ação que seria tomada por cada mecanismo-*spam* estatísticas e relatórios são gerados. O sistema ADES também possui um Pote de Mel que permite analisar características dos *spammers*. O Pote de Mel divulga endereços 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. 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.

O módulo de verificação dos mecanismos tem como objetivo realizar os testes de desempenho dos mecanismos anti-*spam* analisados. O módulo foi implementado na linguagem Perl e funciona como um servidor de políticas [Postfix, 2006] 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 (*Domain Name System*) reverso e SPF (*Sender Policy Framework*). Para o mecanismo de listas negras são consultadas as cinco listas negras[1] mais utilizadas [Ramachandran e Feamster, 2006]. 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 empregado 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* Razor [Prakash, 2007]. 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 avaliação 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 falso-positivos e de falso-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, as mensagens recebidas devem ser manualmente classificadas pelos usuários, para comparar a classificação do usuário com a classificação efetuada pelo mecanismo. O usuário precisa apenas mover as mensagens que são *spams* 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 do usuário e compara o resultado da classificação do usuário e a classificação dos mecanismos anti-*spam* para determinar a taxa de falso-positivos e falso-negativos de cada mecanismo. Mensagens na pasta de *spams* e na lixeira com mais de trinta dias são apagadas automaticamente, para não utili-

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

zarem espaço em disco desnecessariamente. Entretanto, antes de apagar essas mensagens, a linha do cabeçalho da mensagem com o resultado dos mecanismos anti-*spam* é guardada em um arquivo especial. O módulo de análise de resultados além de analisar as mensagens em todas as pastas, também considera o arquivo especial com as linhas de resultado das mensagens apagadas. Dessa forma, mesmo sem armazenar as mensagens antigas, elas ainda são consideradas.

4. Resultados

Os resultados mostrados nesta seção correspondem ao conjunto de todos os usuários que utilizam a ferramenta. No entanto, a ferramenta apresenta também os resultados separadamente para cada usuário. Nos resultados apresentados, dezoito usuários diferentes utilizaram a ferramenta. As mensagens enviadas para endereços divulgadas no Pote de Mel foram automaticamente consideradas como *spams*. O número total de mensagens utilizadas na análise foram 63.325 mensagens legítimas e 3.392.931 *spams*, recebidas pelos dezoito endereços de usuários legítimos e pelos endereços do Pote de Mel no período de um ano e seis meses.

4.1. Análise de desempenho dos Mecanismos

A Figura 1 mostra os percentuais de falso-positivos e falso-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 falso-positivos para os mecanismos de SPF, listas negras e DNS reverso foi consideravelmente alta, chegando a 17,1% das mensagens para o mecanismo SPF. Isso ocorre já que muito *spammers* publicam registros SPF válidos para tentarem burlar os mecanismos anti-*spam*. A alta taxa de falso-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 falso-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 falso-positivos, causando uma insatisfação dos usuários. Os mecanismos de pesos e regras e filtros bayesianos obtiveram as menores taxas de falso-positivos, embora ainda sejam altas ao considerar-se o impacto negativo que podem causar aos usuários. Em todos os mecanismos a taxa de falso-negativos é alta, chegando a 67,4% para o mecanismo de DNS reverso, mostrando a ineficiência desse mecanismo. A alta taxa de falso-negativos para o DNS reverso é causada por *spammers* que invadem máquinas de terceiros com o DNS reverso configurado corretamente para enviar *spams*.

4.2. Características dos *spams*

Os *spams* possuem características que diferem das mensagens legítimas. Os *spammers* possuem uma relação diferente dos usuários legítimos, principalmente em seus comportamentos. Os *spammers* tentam enviar *spams* para um grande número de destinatários. Por outro lado, os usuários legítimos trocam mais mensagens durante o período de trabalho e durante discussões por correio eletrônico.

A Figura 2(a) apresenta o percentual de mensagens recebidas em função das horas do dia. Pode-se notar que no período entre duas e seis horas da manhã a quantidade

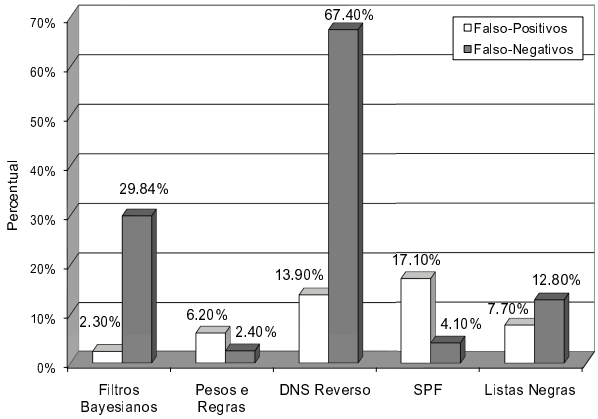


Figura 1. Percentual de falso-positivos e falso-negativos.

de mensagens legítimas trocadas é muito baixa. Em contrapartida, o envio de *spams* é relativamente constante, independente do horário, pois as mensagens são enviadas por mecanismos automatizados que podem ser executados continuamente, diferente das pessoas, que geralmente trabalham durante o dia e dormem à noite.

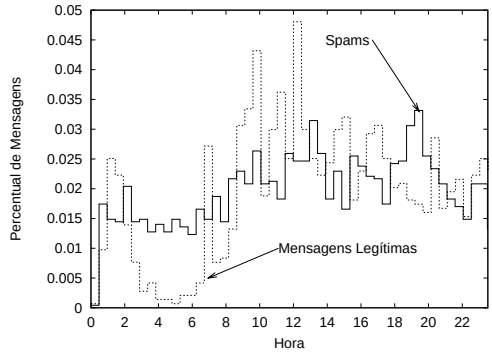
A Figura 2(b) mostra o gráfico do percentual de mensagens em função do tamanho das mensagens. A maior parte dos *spams* possui um tamanho menor que as mensagens legítimas. Uma das preocupações iniciais dos *spammers* era a criação de mensagens com um tamanho pequeno, para enviá-las rapidamente. Atualmente, muitos *spammers* estão utilizando imagens para tentar enganar os mecanismos que realizam análise por conteúdo, o que torna as mensagens maiores do que as mensagens que possuem somente texto. Os resultados obtidos se diferenciam de outros resultados já apresentados na literatura, como em [Gomes et al., 2004]. Isso porque os testes em [Gomes et al., 2004] foram realizados em 2004 e as mensagens *spams* com imagens ainda eram incomuns.

Além das características citadas anteriormente, outra característica que difere as mensagens legítimas dos *spams* é a distribuição dos remetentes por faixas de endereço IP. A Figura 2(c) mostra o percentual de mensagens em função da faixa de endereços IP. Na faixa *146.164.0.0/16* existe uma grande concentração de mensagens legítimas, pois essa faixa pertence à Universidade Federal do Rio de Janeiro e são trocadas muitas mensagens entre os usuários da universidade.

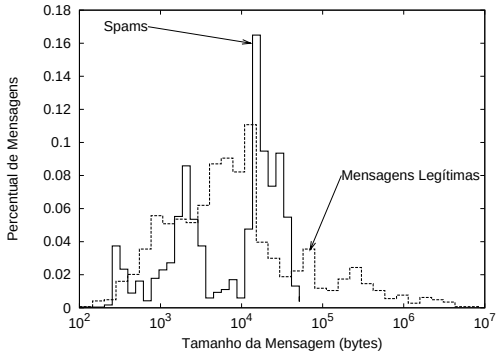
Para comparar a distribuição dos endereços IP que enviam *spams* com a distribuição de endereços IP de máquinas ativas, foi realizado um teste onde foram enviados 500.000 pacotes de *ping* para endereços IP escolhidos aleatoriamente. A distribuição dos endereços IP das máquinas alcançáveis é mostrada na Figura 2(d) e possui um formato similar aos endereços IP das máquinas que enviam *spam*. Esse resultado mostra que o envio dos *spams* é feito de máquinas dispersas por todo o mundo, o que é causado pelo uso das *botnets*.

4.3. Análise do Pote de Mel

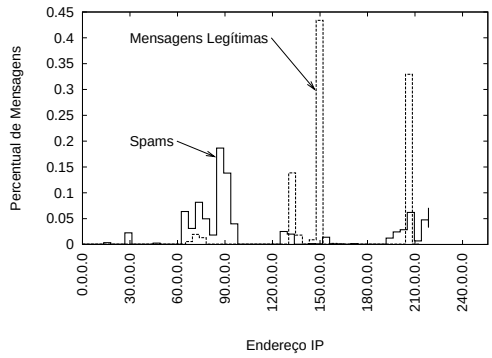
O Pote de Mel implementado permite a análise de características relativas à coleta de endereços e envio de *spams*. O número de endereços divulgados no Pote de Mel que receberam pelo menos uma mensagem foi 5.960. A Figura 3(a) mostra a função de distribuição cumulativa do percentual de mensagens recebidas em função do tempo entre



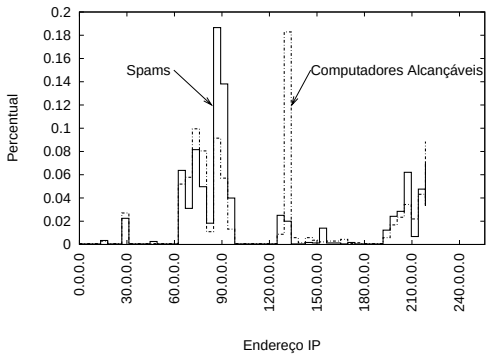
(a) Percentual de mensagens em função da hora do dia.



(b) Percentual de mensagens em função do tamanho das mensagens.



(c) Percentual de mensagens em função das faixas de endereços IP.



(d) Percentual de computadores alcançáveis e computadores que enviam *spam* em função das faixas de endereços IP.

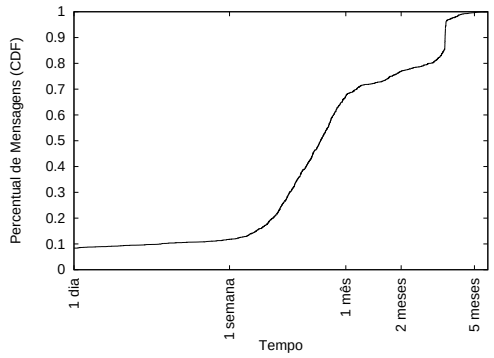
Figura 2. Características dos *spams*.

a divulgação do endereço e o recebimento do primeiro *spam*. Observa-se que, em quase 90% dos casos, os *spams* só começaram a ser recebidos uma semana após terem sido coletados na página principal do GTA. Portanto, mesmo que o endereço eletrônico deixe de ser divulgado, ele ainda receberá *spams* durante um longo período, pois o tempo entre o processo de coleta e o envio dos *spams* tende a ser longo.

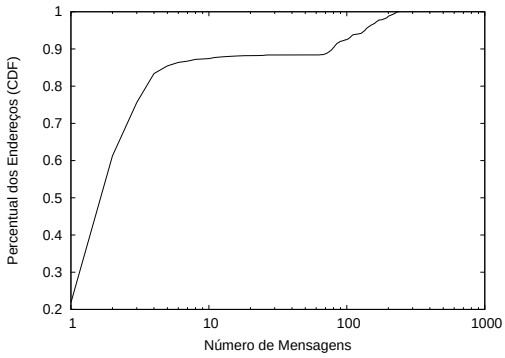
A Figura 3(b) apresenta a função de distribuição cumulativa do percentual dos endereços em função do número de mensagens recebidas. Esse gráfico mostra que aproximadamente 13% dos endereços divulgados receberam mais do que 10 mensagens, mostrando que o número de mensagens que cada *spammer* envia para os endereços em sua lista é reduzido. Por outro lado, para um pequeno percentual de endereços, o número de mensagens enviadas foi maior do que cem.

Para avaliar o tempo de permanência dos endereços nas listas dos *spammers* foi calculada a diferença entre o tempo em que a primeira e a última mensagem foi enviada para cada um dos endereços do Pote de Mel. No entanto, somente essa diferença não permite uma avaliação do tempo de permanência dos endereços nas listas dos *spammers*, já que a diferença entre o tempo da primeira e última mensagem é pequena para os endereços que receberam a primeira mensagem próximo do período final de observação. Para contornar esse problema, a diferença entre a primeira e última mensagem foi normalizada pela diferença entre o tempo final da observação e o tempo no qual a primeira

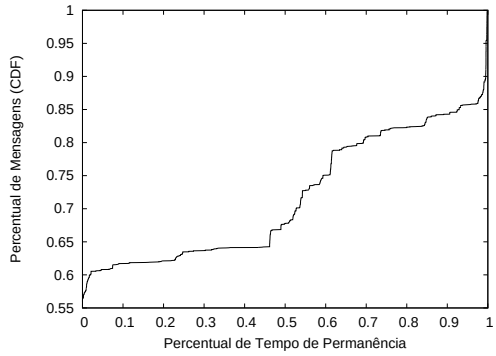
mensagem foi recebida. A Figura 3(c) mostra a função de distribuição cumulativa do percentual de mensagens em função da diferença de tempo normalizada entre a primeira e a última mensagem que cada endereço recebeu. Observa-se que em aproximadamente 55% dos casos o tempo de permanência na lista foi quase zero, indicando que o endereço só recebeu mensagens por um curto intervalo de tempo. Já para 15% dos endereços, foram recebidas mensagens durante todo o período entre a chegada da primeira mensagem e o final do tempo de observação.



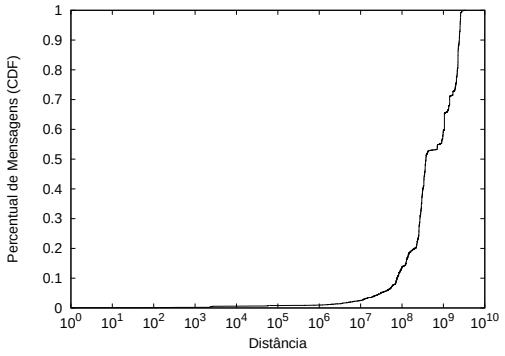
(a) Percentual acumulado de mensagens em função do tempo entre coleta e envio.



(b) Percentual acumulado dos endereços em função do número de mensagens recebidas.



(c) Percentual acumulado de mensagens em função do percentual de tempo de permanência na lista.



(d) Percentual acumulado de mensagens em função da distância entre o endereço IP utilizado para a coleta e o envio.

Figura 3. Análises do pote de mel.

Procurou-se verificar se a coleta de dados e o envio de spams eram feitos pela mesma máquina. Para isso foi avaliada a distância[2] entre o endereço IP que realizou o processo de coleta e o endereço IP que realizou o processo de envio do *spam*. A Figura 3(d) apresenta a função de distribuição cumulativa do percentual de mensagens em função da distância. Em 90% dos casos, a distância é maior do que 10^8. Se for computada a diferença entre os dois endereços da forma $x.y.0.0$ e $x.y.255.255$, que representam os extremos de uma rede classe C, a distância entre os dois endereços é 65.535. Esse fato sugere que as máquinas utilizadas na coleta e no envio não se encontram na mesma rede, o que pode ser causado pela execução paralela dos mecanismos de coleta e envio em máquinas separadas ou então pela utilização de máquinas zumbis que ficam responsáveis por coletar ou enviar *spams*.

[2]A distância entre dois endereços IP foi definida como o módulo da diferença entre a representação decimal dos endereços IP. Para converter o endereço IP que é escrito no formato $x.y.z.w$, simplesmente realizou-se a operação: $x * 256 * 256 * 256 + y * 256 * 256 + z * 256 + w$.

5. Conclusão

Neste trabalho foi apresentada a ferramenta ADES (**A**nálise **DE S**pam), desenvolvida para avaliar a eficiência dos mecanismos anti-*spam* e, também, para analisar características e técnicas utilizadas por *spammers* para enviar *spams*. Os resultados mostraram uma taxa de falso-negativos alta, entre 2,3% e 67,4%. Também foi observada uma taxa de falso-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 falso-positivos.

O módulo de Pote de Mel do sistema ADES mostrou que o tempo entre a coleta dos endereços e o envio da primeira mensagem é longo, da ordem de algumas semanas. Além disso, verificou-se que o tempo de permanência dos endereços nas listas dos *spammers* é geralmente longo. Dessa forma, retirar um endereço já divulgado em uma página pode ajudar a reduzir o número de *spams* recebidos, uma vez que ele não será incluído em novas listas, mas ele continuará recebendo por um período longo *spams* das listas em que ele já se encontra. As análises do Pote de Mel também mostraram que os processos de coleta e envio de mensagens geralmente são realizados de máquinas diferentes, sendo que em mais de 90% dos casos as máquinas não pertencem à mesma rede. Essa característica mostra a ineficiência de propostas de mecanismos que identifiquem as máquinas que estão realizando o processo de coleta para incluí-las em uma lista negra.

Não existe hoje nenhum indício que permita inferir que a atividade de enviar *spams* diminuirá nos próximos anos. Ao contrário, os *spammers* vêm se especializando e usando técnicas cada vez mais elaboradas para burlar os sistemas anti-*spam*. Os *spammers* estão constantemente evoluindo, tentando adaptar-se aos novos mecanismos anti-*spam*, fazendo com que os mecanismos anti-*spam* também evoluam. Assim, uma ferramenta que permite avaliar o desempenho dos mecanismos anti-*spam* e do comportamento dos *spammers* pode ajudar na criação de mecanismos anti-*spam* mais eficientes.

Referências

Andreolini, M., Bulgarelli, A., Colajanni, M. e Mazzoni, F. (2005). Honeyspam: Honeypots fighting spam at the source. Em *SRUTI05: Steps to Reducing Unwanted Traffic on the Internet Workshop*, páginas 77–83.

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)*, páginas 356–369. ACM Press.

Postfix (2006). Postfix SMTP access policy delegation. `http://www.postfix.org/SMTPD_POLICY_README.html`.

Prakash, V. V. (2007). Vipul's razor. http://razor.sourceforge.net/.

Ramachandran, A. e Feamster, N. (2006). Understanding the network-level behavior of spammers. Em *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, páginas 291–302. ACM Press.

Taveira, D. M., Moraes, I. M., Rubinstein, M. G. e Duarte, O. C. M. B. (2006). Técnicas de defesa contra spam. Em *Livro Texto dos Mini-cursos do VI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, páginas 202–250. Sociedade Brasileira de Computação.