Detecção de Spammers na Rede de Origem

Pedro Henrique B. Las Casas[1], Dorgival Guedes[2], Jussara M. Almeida[2], Artur Ziviani[3], Humberto T. Marques-Neto[1]

[1] Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
30.535-901 - Belo Horizonte - Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
31.270-010 - Belo Horizonte - Brasil

[3]Coordenação de Ciência da Computação
Laboratório Nacional de Computação Científica (LNCC)
25.651-075 - Petrópolis - Brasil

pedro.casas@sga.pucminas.br, {dorgival,jussara}@dcc.ufmg.br

ziviani@lncc.br, humberto@pucminas.br

Abstract. *The volume of unsolicited messages (spam) sent over the Internet represents more than 85% of all e-mails. Even with the evolution of the filtering techniques such as the analysis of the message content and the blocking of IPs, network resources are wasted given that such a filtering is usually performed at the e-mail destination server. This paper proposes a method for detecting spammers in the origin network using some metrics which do not require inspection of message contents. This method uses a supervised classification technique that has been applied to the two real-world datasets from a Brazilian broadband Internet service provider. Results show that the adopted method is efficient, being able to correctly identify most spammers still in the origin network. In this way, network resources are saved because of the reduced number of spams in transit that likely would be discarded at their destination.*

Resumo. *A quantidade de mensagens não-solicitadas (spams) enviadas na Internet representa mais de 85% de todos os e-mails. Mesmo com a evolução de técnicas de filtragem como a análise do conteúdo de mensagens e o bloqueio de IPs, recursos da rede são desperdiçados, uma vez que essa filtragem é realizada normalmente no servidor de destino dos e-mails. Este trabalho propõe um método para detecção de spammers na rede de origem utilizando métricas que não requerem a inspeção do conteúdo das mensagens enviadas. Esse método utiliza uma técnica de classificação supervisionada, a qual foi aplicada em dois conjuntos de dados reais de um provedor de internet de banda larga brasileiro. Os resultados mostram que o método utilizado é eficaz, sendo capaz de identificar a maioria dos spammers ainda em sua rede de origem. Desta forma, os recursos da rede são preservados a partir da diminuição do número de spams em circulação que provavelmente seriam descartados em seu destino.*

1. Introdução

O crescimento e a importância do serviço de correio eletrônico proporciona um contexto favorável para geração de um grande volume de mensagens não-solicitadas (*spams*). Há algum tempo, os *spams* representam mais de 85% do total de *e-mails* que trafegam pela Internet [IronPort 2009, MessageLabs 2010]. Além do seu caráter indesejado, mensagens de *spam* estão diretamente relacionadas à propagação de *malwares*, como cavalos de tróia, vírus e *worms* [Newman et al. 2002], o que as torna ainda mais nocivas para a rede e seus usuários.

A adoção de filtros *anti-spam* é a principal medida adotada por provedores do serviço de correio eletrônico para diminuir a quantidade de *spam* na caixa de entrada de seus usuários. Normalmente, esses filtros classificam as mensagens como legítimas ou não e, se for o caso, efetuam seu descarte ou utilizam um mecanismo de quarentena para alertar o usuário. Entretanto, mesmo sendo razoavelmente eficiente, essa filtragem ocorre apenas depois que as mensagens são entregues ao servidor de correio eletrônico de destino (ou um intermediário adequado). Nesse ponto, o *spam* já consumiu recursos da rede, como banda passante, e a própria aplicação do filtro consome recursos como memória e capacidade de processamento do servidor de destino.

Uma possível forma de evitar esse desperdício de recursos causado pelos *spams* seria complementar a filtragem no servidor receptor com o uso de técnicas de filtragem prévia, capazes de evitar o envio do *spam* e, com isso, o desperdício de recursos associado. Tais técnicas podem ser aplicadas, por exemplo, em provedores de acesso à Internet de banda larga que, através da análise do tráfego SMTP (*Simple Mail Transfer Protocol*), poderiam detectar a ação de possíveis *spammers*, bloqueando o tráfego na sua origem. Além deste tipo de bloqueio, outras medidas podem ser tomadas para não prejudicar usuários possivelmente classificados como falsos positivos, como por exemplo, o envio de mensagens de alerta, uso (periódico) de desafios para testar a legitimidade de usuários suspeitos, em conjunto com a introdução de atrasos em mensagens destes usuários.

Este artigo propõe um novo método para detecção de *spammers* na rede de origem chamado SpaDeS - *Spammer detection at the Source* - que é baseado em um algoritmo de classificação supervisionada e explora apenas métricas que não requerem a inspeção do conteúdo das mensagens. O método proposto foi aplicado e validado utilizando dois conjuntos de dados reais contendo informações agregadas e anonimizadas de transações SMTP de usuários de um provedor brasileiro de Internet de banda larga residencial coletadas em 2009 e 2010. Os resultados apresentados mostram que a utilização do SpaDeS é capaz de diferenciar *spammers* de usuários legítimos ainda na rede de origem, sem inspecionar o conteúdo de suas mensagens. Com o uso da técnica de classificação supervisionada, validada através da comparação com uma base de dados real de denúncias de *spam* e da inspeção de uma amostra dos usuários classificados, estima-se que cerca de 98% dos usuários legítimos e 94% dos *spammers* foram classificados corretamente. Ou seja, as taxas de falsos positivos e de falsos negativos foram, respectivamente, de 2% e 6%. O estudo mostra também que classes de usuários legítimos representaram cerca de 83% dos usuários e realizaram cerca de 1,6% do total de transações SMTP observadas nos dados de 2010. Enquanto isso, os usuários classificados como *spammers*, (cerca de 17%) originaram mais de 98% de todas as transações SMTP no período observado.

O restante deste artigo está organizado da seguinte forma: os trabalhos relacionados são discutidos na Seção 2; a Seção 3 apresenta o método de detecção de *spammers* proposto e a Seção 4 discute os resultados mais relevantes deste trabalho. Finalmente, a Seção 5 apresenta algumas conclusões e sugere trabalhos futuros.

2. Trabalhos Relacionados

Entender as características dos *spams* é uma tarefa importante para o desenvolvimento de métodos para detecção desse tipo de mensagem assim como de seus remetentes, os *spammers*. No que tange à análise de *spams*, Gomes *et al.* analisaram uma carga de trabalho de mensagens de usuários de uma universidade brasileira e destacaram uma série de características capazes de diferenciar *spams* de mensagens legítimas [Gomes et al. 2007]. Para isso, os autores utilizaram as mensagens recebidas na rede da universidade. Em uma extensão daquele trabalho, os mesmos autores indicam que o tráfego legítimo apresenta menor entropia que o tráfego gerado pelos *spammers*, os quais, geralmente, enviam e-mails indistintamente para os seus alvos [Gomes et al. 2009]. De forma similar, Kim *et al.* caracterizam o tráfego de *spam* de uma universidade sul-coreana também com dados da camada de aplicação no destino da mensagem, mostrando que o intervalo entre chegadas de *spams* é bem inferior ao intervalo entre e-mails legítimos (menor que 5 segundos em 95% dos casos) [Kim e Choi 2008].

No que tange à análise de *spammers*, diversos trabalhos propõem soluções para sua identificação em pontos intermediários da rede. Ramachandran *et al.* investigam características de tráfego, coletadas da camada de rede, tais como a persistência de endereços IP e de rotas e características específicas de *botnets*, que sejam comuns a *spammers* [Ramachandran e Feamster 2006], enquanto Guerra *et al.* analisam os padrões de comunicação presentes em uma campanha de *spam* [Guerra et al. 2009]. Já Hao *et al.* aplicam técnicas de aprendizado de máquina em dados coletados da camada de rede para classificar *usuários* em legítimos e *spammers* em um servidor posicionado entre as redes de origem e destino [Hao et al. 2009]. Schatzmann *et al.* propõem detectar *spammers* no nível de sistemas autônomos (AS), coletando e combinando as visões locais de múltiplos servidores de e-mail destinatários [Schatzmann et al. 2009]. Ao contrário desses trabalhos, este artigo propõe a detecção dos *spammers* ainda na rede de origem, para minimizar o desperdício de recursos devidos ao processo de recepção das mensagens.

Através da análise de características de fluxos de pacotes SMTP, Sperotto *et al.* propõem um algoritmo para detecção de *spams* utilizando apenas informações da camada de rede (p.ex: tempo de inatividade e quantidade de picos no fluxos de pacotes) [Sperotto et al. 2009]. Os autores utilizam dados da rede de uma universidade holandesa em conjunto com informações de lista de bloqueio de DNS (*blacklists*) para validar o algoritmo proposto. Taveira *et al.* propõem um mecanismo *anti-spam* baseado em autenticação e reputação dos usuários objetivando minimizar falsos positivos ao classificar *spams* [Taveira e Duarte 2008]. Por outro lado, outros métodos de detecção de *spams* utilizam técnicas de classificação supervisionada, porém exploram características do conteúdo das mensagens [Kolcz e Alspector 2001]. O trabalho aqui proposto explora técnicas semelhantes, mas considera apenas métricas relacionadas aos protocolos envolvidos, sem inspecionar o conteúdo das mensagens, para garantir a privacidade dos usuários legítimos, e tem como alvo a detecção de *spammers*.

Em um trabalho anterior, os autores caracterizaram o tráfego SMTP de usuários de um provedor de Internet residencial de banda larga [Castilho et al. 2010], identificando várias características da camada de rede e do protocolo SMTP, coletadas na rede de origem, que permitem classificar os usuários em usuários legítimos e usuários com comportamento abusivo (potencialmente *spammers*). Esses resultados motivaram o desenvolvimento do método aqui proposto, conforme será detalhado na próxima seção.

3. SpaDeS: Detector de Spammers na Rede de Origem

O método proposto para detecção de *spammers* na rede de origem, denominado SpaDeS (*Spammer Detection at the Source*), tem como principal componente um algoritmo de classificação supervisionada, que "aprende" um modelo de classificação de usuários a partir de um conjunto de exemplos (usuários) previamente classificados (conjunto de treino). O classificador recebe como entrada o número de classes distintas C e exemplos de usuários de cada uma. Após a fase de aprendizado, o modelo derivado pode então ser aplicado para classificar novos usuários (conjunto de teste) nas classes pré-definidas. A Seção 3.1 apresenta as classes de usuários consideradas assim como o modelo de representação dos mesmos. O algoritmo de classificação utilizado é apresentado na Seção 3.2, enquanto a Seção 3.3 discute como obter o conjunto de treino.

3.1. Modelo de Representação de Usuários e suas Classes

Cada usuário é representado por um vetor de N atributos que conjuntamente descrevem seu comportamento quanto ao uso do protocolo SMTP. Para detectar *spammers* na rede de origem com eficiência, foram utilizados N=6 atributos, que são métricas que não envolvem processamento do corpo da mensagem. As métricas são: número de transações SMTP realizadas, número de remetentes distintos[1], número de servidores SMTP distintos acessados, tamanho médio das transações SMTP, distância geodésica[2] média entre origem e destino e tempo médio entre transações consecutivas (aqui referenciado como IATs, *inter-arrival times*). Os atributos foram mantidos sem normalização, dado que experimentos preliminares com diferentes estratégias de normalização não levaram a melhorias significativas nos resultados da classificação.

A escolha das métricas foi inspirada nos resultados de um trabalho anterior[3] [Castilho et al. 2010]. Utilizando o algoritmo de agrupamento *X-means* [Pelleg e Moore 2000], demonstrou-se que essas características podem ser utilizadas para distinguir 4 classes de comportamento, sendo que duas refletem padrões de usuários legítimos (classes 1 e 2) enquanto as outras refletem padrões abusivos (classes 3 e 4), potencialmente de *spammers*. Por exemplo, o número de transações por usuário é útil para distinguir usuários que fazem pouco uso de SMTP daqueles que o utilizam com grande intensidade. Mais ainda, enquanto o uso de poucos servidores de SMTP é o esperado para usuários legítimos, o acesso a um número muito grande pode indicar a operação de *open proxies* ou de *open mail relays* sendo explorados para o envio de spam por usuários maliciosos ou *bots* [Guerra et al. 2009]. O uso da distância geodésica como

[1]Essa informação pode ser obtida durante a negociação do protocolo SMTP, sem necessidade de inspecionar as mensagens propriamente ditas.

[2]Menor distância entre dois pontos ao longo da superfície da Terra.

[3]O número de remetentes distintos não foi considerado naquele trabalho, entretanto, em experimentos preliminares, observamos uma melhoria da classificação com a sua inclusão.

métrica se baseia na hipótese de que conexões SMTP de *spammers* tendem a ocorrer entre endereços IPs mais distantes que as conexões de usuários legítimos, já que *spammers* tendem a ocultar suas presença usando máquinas em outros países [Guerra et al. 2008]. Como naquele trabalho, foram consideradas C=4 classes distintas, discutidas em mais detalhes na Seção 4.

3.2. Algoritmo de Classificação Supervisionada

O algoritmo de classificação utilizado neste trabalho é o *Lazy Associative Classifier* (LAC) [Veloso et al. 2006], que tem ótima escalabilidade, com complexidade de tempo polinomial. O LAC fornece uma estimativa da confiança na predição feita em cada caso. Essa confiança, que pode ser interpretada como uma probabilidade de acerto da classificação, será explorada na geração do conjunto de treino (Seção 3.3). O LAC explora o fato de que, frequentemente, há fortes associações entre os valores dos atributos e as classes. Tais associações estão geralmente implícitas no conjunto de treino e, quando descobertas, revelam aspectos que podem ser utilizados para prever as classes dos usuários. O LAC "aprende" o modelo de classificação em duas etapas. Inicialmente, ele extrai, do conjunto de treino, regras de associação do tipo $\mathcal{X} \rightarrow c_i$, que indicam a associação entre um conjunto de valores de atributos $\mathcal{X}$ e uma classe c_i, atribuindo uma confiança a cada regra. O LAC então prevê a classe de um usuário u no conjunto de teste combinando as confianças de todas as regras $\mathcal{X} \rightarrow c_i$ tal que $\mathcal{X}$ contém valores de atributos que coincidem com os de u. A classe de u será aquela que tiver a maior confiança agregada. Dois parâmetros principais do LAC são o tamanho máximo das regras (número de atributos em $\mathcal{X}$) e a confiança mínima permitida. Considerou-se o tamanho máximo como 5 e uma confiança mínima de 0,01, valores de referência comumente usados.

Outras técnicas de classificação também foram consideradas, como por exemplo, SVM e Naive Bayes. O LAC foi escolhido por produzir estimativas de confiança que, conforme observações experimentais, tendem a ter uma confiabilidade maior do que as estimativas oferecidas por outras técnicas.

3.3. Coleção de Treino

O funcionamento de qualquer método de classificação supervisionada depende primariamente de um conjunto de treino contendo usuários pré-classificados. A obtenção desse conjunto para a classificação de usuários em *spammers* e legítimos é um grande desafio, uma vez que tais dados tipicamente não estão disponíveis publicamente. Um fator complicador é que almeja-se detectar *spammers* ainda na rede de origem. Logo, faz-se necessário um conjunto de treino coletado naquele ponto do sistema. Caso contrário, os padrões levantados poderiam não generalizar para o conjunto de teste, resultando em um desempenho pobre do classificador. Uma estratégia seria realizar inspeção manual em mensagens de um subconjunto dos usuários. Entretanto, o custo para esse esforço manual seria muito alto, uma vez que um número considerável de usuários (e as mensagens enviadas por eles) teriam que ser inspecionados a fim de se obter um número suficiente de representantes dos vários padrões observados. Além disto, ele exigiria acesso ao conteúdo das mensagens enviadas, o que pode não ser viável. Assim, são propostas duas estratégias para geração da coleção de treino, uma baseada em informação externa ao algoritmo de aprendizado, enquanto a segunda utiliza a informação de confiança provida pelo LAC.

A primeira estratégia parte das 4 classes de usuários identificadas no trabalho anterior dos autores [Castilho et al. 2010] e acrescidas de informações sobre usuários apontados como *spammers* por mecanismos de relato de abusos do sistema de correio. Para cada uma das duas classes representando usuários legítimos identificadas [Castilho et al. 2010], foram selecionados os M usuários mais próximos do centróide de cada classe, de forma a obter bons representantes de cada uma. As duas classes de usuários abusivos (possíveis *spammers*) identificadas naquele trabalho apresentam uma variabilidade maior de comportamentos. Por esse motivo, optou-se por não usar o mesmo mecanismo de seleção, mas baseou-se a escolha em uma informação confiável externa. Para isso, utilizou-se a identificação de usuários cujas máquinas foram apontadas como origem de *spam* por relatos oriundos de outros provedores. Tais relatos, enviados para o endereço `abuse` do provedor que forneceu os dados utilizados nesse trabalho, são gerados por provedores tanto a partir de reclamações de seus usuáriso (como o recurso "*Report spam*" do Gmail) ou por mecanismos automáticos, como listas de bloqueio ou outros mecanismos de detecção automática de *spam*. Como será visto, um número razoável de usuários denunciados estavam presentes naquelas duas classes, sendo portanto bons representantes das mesmas.

A segunda estratégia parte do pressuposto que o SpaDeS deve ser aplicado continuamente, em diferentes conjuntos de teste (p.ex: dados referentes a diferentes semanas, meses ou anos). Logo, propôs-se retreinar o LAC a partir do resultado da sua execução anterior, explorando as confianças reportadas naquela classificação. Ou seja, considerando sucessivos conjuntos de teste $t_1, t_2 \cdots t_n$, seleciona-se como treino do LAC para a classificação do teste t_i, os usuários do conjunto t_{i-1} que foram classificados com uma confiança superior a um certo limiar. Para a classificação de t_1, um conjunto de treino inicial é necessário, podendo ser obtido pela primeira estratégia. O algoritmo 1 apresenta a estratégia utilizada. Ele garante que pelo menos $\alpha\%$ dos usuários de cada classe sejam selecionados, mantendo uma confiança mínima uniforme entre todas as classes.

Algoritmo 1: Dados os usuários classificados pelo LAC na iteração anterior, faça:
1. Ordene os usuários de cada classe em ordem decrescente de confiança;
2. Selecione $\alpha\%$ dos usuários de cada classe, ordenados anteriormente;
3. Seja c_i^{min} a menor confiança dos usuários selecionados da classe i ($i = 1..4$);
4. Seja $c = min(c_1^{min}, c_2^{min}, c_3^{min}, c_4^{min})$;
5. Selecione para o conjuto de treino todos os usuários que possuem confiança $\geq c$, mantendo, para cada um, a classe definida pelo LAC na iteração anterior.

Neste trabalho utilizou-se a primeira estratégia, baseada no algoritmo de agrupamento, apenas na iteração inicial e continuou-se o treinamento a partir do resultado da classificação anterior para as iterações seguintes. Esse enfoque se baseia no fato de que, considerando que um número suficiente de bons exemplos de treinos sejam fornecidos, técnicas de classificação supervisionadas (p.ex., LAC) tendem a ser superiores às técnicas não supervisionadas (p.ex., algoritmos de agrupamento) [Veloso et al. 2006]. Note que a coleção de treino pode ser estendida e/ou refinada para incluir exemplos pré-classificados por outros meios, potencialmente mais confiáveis, se tais exemplos estiverem disponíveis. Por exemplo, assim como feito na iteração inicial, havendo conhecimento sobre usuários locais denunciados como *spammers*, os mesmos poderiam ser incluídos no treino das ite-

rações seguintes. Como ficará claro na Seção 4, optou-se por não incluir tais usuários no conjunto de treino nas iterações seguintes para que os mesmos pudessem ser utilizados para validação do método proposto no treinamento.

A estratégia proposta é completamente automatizada e não exige esforço manual de classificação. Note que a natureza iterativa do processo, que utiliza resultados da iteração anterior como treino da próxima iteração, pode afetar a classificação ao longo do tempo. Entretanto, nos experimentos realizados, observou-se que os padrões de cada classe de usuários se mantêm estáveis em duas bases de dados incluindo tráfego em 2009 e 2010. Mais ainda, a classificação dos usuários da base de 2010, seguindo a abordagem iterativa descrita, apresentou excelente efetividade (Seção 4). De qualquer maneira, considera-se que, para melhor refletir os padrões de comportamento dos usuários, que podem evoluir com o tempo, e também para interromper uma possível propagação de erros, seja necessário, periodicamente, a aplicação de um conjunto de treino obtido por métodos externos (como a primeira estratégia proposta), reiniciando um novo ciclo de iterações. A frequência com que isto deve ser feito é objeto de trabalho futuro, pois dependerá da disponibilidade de uma série temporal mais longa para avaliação.

4. Avaliação e Resultados

4.1. Bases de Dados

Este trabalho utiliza 4 bases de dados diferentes, sendo que duas delas refletem o tráfego SMTP de um provedor de Internet de banda larga e duas contêm listas de usuários daquele provedor que foram denunciados como *spammers* através do endereço `abuse` daquele provedor durante o período considerado.

Cada base de dados de tráfego contém um log de tráfego e um log do serviço DHCP do provedor, ambos cobrindo um mesmo período. Os logs de tráfego são formados por *transações*. Cada transação representa uma conexão TCP ou um fluxo de dados UDP, contendo informações como endereços IP de origem e de destino, serviço/protocolo utilizado, data/hora inicial, duração e volume de bytes enviados e recebidos. Os logs do serviço DHCP permitem associar transações e usuários através do mapeamento dos endereços físicos de suas máquinas (*MAC addresses*) para os endereços IP fornecidos pelo provedor, com base nas informações de data e hora presentes nos dois logs. Vale ressaltar que os dados dos usuários foram anonimizados, por questões de segurança e privacidade.

As bases de dados de tráfego cobrem os períodos de 01 a 28 de março de 2009 e 12 de junho a 09 de julho de 2010. A base de 2009 contém 40,6 milhões de transações associadas a 44,2 mil usuários. Já a de 2010 contém 45,6 milhões de transações associadas a 48 mil usuários. Cada base passou por um processo de filtragem, sendo removidas transações: (1) que não usavam SMTP; (2) com duração, número de bytes enviados e/ou recebidos iguais a zero, consideradas erros de coleta; (3) que enviaram menos de 160 bytes ou receberam menos de 80 bytes. Estes últimos limiares foram definidos por corresponderem ao número mínimo de bytes para se estabelecer e encerrar uma conexão TCP, considerando 40 bytes para os cabeçalhos IP e TCP nos pacotes de *three-way handshake* e de finalização. Após filtragem, restaram 6,3 milhões de transações SMTP associadas a 5.479 usuários na base de 2009, e 5 milhões de transações SMTP associadas a 5.389 usuários na base de 2010.

As duas outras bases de dados contêm denúncias recebidas pelo endereço `abuse` do provedor durante os períodos das bases de tráfego, identificando certos usuários como *spammers*. Os e-mails de denúncia informam o endereço IP de origem do *spam* e a data/hora do seu recebimento e estão no formato ARF (*Abuse Reporting Format*), utilizado para mensagens desse tipo [Shafranovich et al. 2010]. Foi desenvolvida uma ferramenta de extração para processar essas mensagens e realizar a junção das mesmas com as transações SMTP, possibilitando a identificação de usuários denunciados. Dessa forma, foram identificados 67 e 93 *spammers* nas bases de 2009 e 2010, respectivamente. Para todos esses usuários, os endereços IPs e as datas/horas listados nas denúncias coincidiram com dados de transações realizadas, listadas nas bases de tráfego utilizadas.

4.2. Procedimento de Avaliação

A avaliação consistiu de dois experimentos de classificação, um com as bases de 2009 e outro com as bases de 2010. Para a classificação da base de 2009, foram selecionados M=30 usuários mais próximos do centróide de cada uma das classes de usuários legítimos (Seção 3.1). Além disso, dos 67 usuários denunciados como *spammers* identificados na base de 2009, 40 são da classe 3 e 27 da classe 4. Assim, essas duas classes são formadas principalmente por *spammers* e são o principal alvo do método de detecção. O conjunto de treino foi então composto por esses 127 usuários, e para teste foram utilizados todos os usuários da base de 2009 *que não estavam no conjunto treino*.

Para o segundo experimento de classificação, realizado sobre a base de 2010, foram utilizados como treino usuários selecionados pelo algoritmo 1 considerando o resultado da classificação da base de 2009 (Seção 3.3). Foi utilizado α = 20%, o que resultou em uma confiança mínima para todos os usuários selecionados de 64%. No total, foram selecionados 787, 605, 621 e 52 usuários das classes 1, 2, 3 e 4, respectivamente.

As seções seguintes apresentam os principais resultados dos dois experimentos. Para o primeiro experimento, não foi possível quantificar a efetividade da classificação, uma vez que os dados do `abuse`, que poderiam servir como base de comparação, foram usados no treino. Optou-se nesse caso por analisar os padrões de comportamento dos usuários classificados em cada classe. A principal validação quantitativa é feita sobre os resultados do segundo experimento: além da avaliação dos padrões identificados, utilizou-se a lista de usuários denunciados como *spammers* em 2010 e fez uma inspeção manual de uma amostra dos demais usuários para estimar a efetividade da classificação.

4.3. Classificação da Base de Dados de 2009

A Tabela 1 mostra um sumário dos usuários selecionados para compor o conjunto de treino e que representam as quatro classes identificadas anteriormente [Castilho et al. 2010]. Ela mostra, para cada métrica analisada, a média e o coeficiente de variação CV (razão entre desvio padrão e a média), computados para os usuários de cada classe. As classes 1 e 2 apresentam comportamentos razoavelmente bem uniformes, principalmente com relação aos números de transações, número de remetentes distintos, número de servidores contactados e IAT das transações. A classe 1 compreende usuários que fazem muito baixo uso do correio eletrônico, com apenas uma transação no período coberto pela base. Já a classe 2 compreende usuários com um nível de atividade um pouco mais alto, realizando tipicamente uma transação SMTP a cada 3 dias (IAT médio igual a 70,43 horas). Para ambas as classes, os números de remetentes identificados

(1 e 1,70, em média) e de servidores SMTP distintos acessados (1 e 1,53, em média) são baixos, demonstrando um comportamento esperado de usuários legítimos. O mesmo vale para a distância geodésica, relativamente baixa.

Tabela 1. Conjunto de Treino para a Classificação da Base de 2009.

	Classe 1	Classe 2	Classe 3	Classe 4
Número de usuários	30	30	40	27
	Média (CV)	**Média (CV)**	**Média (CV)**	**Média (CV)**
Número de transações SMTP	1 (0)	4,53 (0,46)	5.304,30 (0,85)	37.841,48 (0,87)
Número de remetentes distintos	1 (0)	1,70 (0,62)	2.718,32 (0,71)	21.660,22 (1,10)
Número de servidores SMTP distintos	1 (0)	1,53 (0,62)	2.770,90 (0,63)	14.704,89 (0,58)
Tamanho das transações SMTP (KB)	388,81 (3,61)	71,24 (2,07)	2,53 (1,66)	1,84 (0,44)
Distância geodésica entre os IPs (km)	1.833 (1,65)	3.443 (0,90)	8.833 (0,05)	8.216 (0,05)
IAT das transações SMTP (h)	672 (0)	70,43 (0,01)	0,19 (2,62)	0,01 (0,54)

Tabela 2. Classificação dos Usuários da Base de 2009.

	Classe 1	Classe 2	Classe 3	Classe 4
Número de usuários	1.422	2.938	927	65
	Média (CV)	**Média (CV)**	**Média (CV)**	**Média (CV)**
Número de transações SMTP	1 (0)	25,71 (9,61)	2.697,54 (1,53)	39.108,21 (1,05)
Número de remetentes distintos	1 (0)	4,14 (1,93)	1.202,50 (1,39)	16.032,23 (0,98)
Número de servidores SMTP distintos	1 (0)	5,33 (2,32)	1.261,37 (1,17)	11.766,04 (0,60)
Tamanho das transações SMTP (KB)	646,81 (4,34)	611,40 (7,51)	9,64 (12,83)	1,67 (0,64)
Distância geodésica entre os IPs (km)	2.657 (1,32)	3.133 (1,1)	8.081 (0,22)	8.352 (0,16)
IAT das transações SMTP (h)	672 (0)	55,34 (1,02)	0,57 (2,45)	0,01 (0,55)

Já as classes 3 e 4, representativas de *spammers* e definidas pelos usuários denunciados, revelam padrões bem distintos. Embora ambas apresentem números de transações, remetentes e servidores SMTP superiores aos das classes 1 e 2, a classe 4 revela um padrão muito mais abusivo, com cada usuário enviando 37.841 transações SMTP (uma a cada 36 segundos), utilizando 21.660 remetentes distintos e acessando 14.704 servidores SMTP distintos, em média. De fato, essas classes revelam dois tipos de *spammers* distintos: um envia o maior número de mensagens possível (classe 4) e o outro (classe 3) envia mensagens utilizando, possivelmente, um controle de fluxo com o objetivo de disfarçar sua presença [John et al. 2009]. É interessante notar também os valores médios de IAT muito baixos e as distâncias geodésicas muito mais altas, para ambas as classes, padrões esperados para *spammers* [Kim e Choi 2008, Hao et al. 2009]. Por fim, vale também ressaltar os tamanhos de transações muito menores que os das classes 1 e 2, em consistência com estudos anteriores que demonstraram que *spams* tendem a ser menores que mensagens legítimas [Gomes et al. 2007].

Os resultados da classificação são apresentados na Tabela 2, que sumariza as principais características dos usuários em cada classe. A classe 1 é formada por 1.422 usuários, que representam 26,56% do total de usuários na base de 2009, mas são responsáveis por menos de 1% do total de transações SMTP enviadas. Em consistência com o conjunto de treino, esta classe se mostrou a mais uniforme: todos os usuários fizeram apenas uma transação SMTP, utilizando portanto apenas um remetente e acessando apenas um servidor SMTP no período de 28 dias. A distância geodésica média entre os IPs origem e destino e, principalmente, o tamanho médio das transações SMTP possuem uma variabilidade maior, mas ainda assim refletem padrões que podem ser esperados de usuários legítimos. Por exemplo, 60% dos usuários tiveram distância geodésica média igual a 0 km, indicando transações SMTP realizadas dentro do Brasil. Mais ainda, 98% dos usuários possuem distância geodésica média menor que 7.000 km, possivelmente refletindo

acesso aos principais servidores globais de e-mail, tais como Gmail, Hotmail e Yahoo. Com relação ao tamanho médio das transações, observou-se valores entre 1 KB e 45 MB. O limite superior pode refletir o envio de mensagens com anexos volumosos, como fotos e vídeos, outro indício de comportamento legítimo.

A classe 2, composta por 2.938 usuários (54,9% do total), apresenta uma variação muito grande em seus usuários, principalmente com relação ao número de transações SMTP, que varia de 2 a 11.804. Apesar deste limite superior muito alto, nota-se que apenas 2% dos usuários desta classe realizaram mais que 100 transações no período. Além disto, cerca de 90% dos usuários acessaram apenas 10 servidores SMTP distintos (5,33, em média) e utilizaram menos de 10 remetentes distintos (4,14, em média). O IAT tende a ser alto, com, em média, 55 horas de inatividade entre transações, e o tamanho das transações possui valores muito altos também, com um máximo de 165 MB. Todas essas características indicam que grande parte dos usuários que compõem esta classe são legítimos. Alguns poucos usuários com um número de transações, número de remetentes e/ou número de servidores SMTP muito altos podem representar *spammers* erroneamente classificados como legítimos (falsos negativos), ou usuários com redes locais com diversos usuários.

A classe 3, formada por 927 usuários (17,32% do total), também apresenta grande variabilidade entre os usuários. O número de transações SMTP, por exemplo, varia entre 2 e 41.227, embora a média seja bastante alta (2.697 transações). De fato, mais de 50% dos usuários realizaram mais de 1.000 transações SMTP, utilizaram mais de 500 remetentes distintos e acessaram mais de 700 servidores SMTP. Além disso, o IAT médio apresentado por 90% dos usuários é menor que 10 minutos, o que indica que a cada 10 minutos o usuário realiza uma transação SMTP. Com base nestes dados, pode-se supor que esses usuários estejam infectados por *malwares*, agindo como *bots* que utilizam controle de fluxo [John et al. 2009], ou seja, enviam *spams* com uma frequência relativamente baixa (do ponto de vista de ferramentas automatizadas) para dificultar sua detecção por sistemas *anti-spam*. É importante ressaltar que essa maior variabilidade dos usuários nas classes 2 e 3 era esperada, uma vez que elas representam comportamentos de fronteira, que podem ser difíceis de serem distinguidos pelo classificador. Assim como discutido para a classe 2, conjectura-se que alguns usuários classificados como da classe 3 podem de fato ser falsos positivos.

Assim como a classe 1, a classe 4 apresenta pouca variabilidade, com CVs variando entre 0,15 e 1,05. Consistente com o conjunto de treino, foram identificados 65 novos usuários com um padrão muito abusivo de uso do SMTP, claramente relacionados à atividade de envio de *spam*. Embora representem apenas 1,21% dos usuários, eles são responsáveis por quase 40% de todas as transações SMTP realizadas.

4.4. Classificação da Base de Dados de 2010

A Tabela 3 mostra os resultados da classificação da base de dados de 2010, utilizando como conjunto de treino usuários selecionados a partir do resultado da classificação da base de 2009, conforme discutido na Seção 4.2. Note que, em termos gerais, os usuários de cada classe mantêm padrões de comportamento bastante semelhantes aos usuários da mesma classe na base de 2009. Por exemplo, as classes 1 e 2, que contabilizam 83% de todos usuários mas apenas 1,6% das transações SMTP realizadas, revelam padrões bastantes consistentes com usuários legítimos: números pequenos de transações, remetentes

Tabela 3. Classificação dos Usuários da Base de 2010.

	Classe 1	Classe 2	Classe 3	Classe 4
Número de usuários	1.821	2.656	836	76
	Média (CV)	**Média (CV)**	**Média (CV)**	**Média (CV)**
Número de transações SMTP	2,78 (9,82)	27,99 (8,79)	2.892.91 (1,53)	34.018,17 (0,70)
Número de remetentes distintos	1 (0)	3,32 (1,37)	1.341,38 (1,73)	22.064,07 (0,73)
Número de servidores SMTP distintos	1 (0)	3,04 (1,81)	1.192,24 (1,52)	14.234,76 (0,50)
Tamanho das transações SMTP (KB)	891,55 (4,53)	826,45 (4,29)	22,64 (5,49)	2,23 (0,78)
Distância geodésica entre os IPs (km)	3.968 (0,91)	4.136 (0,82)	7.653 (0,29)	8.572 (0,05)
IAT das transações SMTP (h)	535,51 (0,49)	58,04 (0,96)	1,00 (2,05)	0,02 (0,50)

e de servidores distintos e longos períodos de inatividade. Além disso, 43% dos usuários da classe 1 e 35% da classe 2 têm distância geodésica média igual a 0 km, enquanto que 98% e 92%, respectivamente, têm distância inferior a 7.000 km. Uma diferença com relação aos resultados da base de 2009 é a maior variabilidade no número de transações (e consequentemente no IAT) entre usuários da classe 1. De fato, o número de transações de usuários da classe 1 variou de 2 a 1.112, embora 95% deles tenha efetuado menos que 10 transações e apenas 1% tenha feito mais de 50 transações. Uma conjectura possível é que a contínua popularização do uso do correio eletrônico possa ser responsável pela maior frequência de uso pelos usuários.

Já as classes 3 e 4, que representam 17% dos usuários e 98,4% das transações, mais uma vez demonstram padrões muito abusivos. Usuários da classe 4 fazem um uso muito mais intenso de tráfego SMTP que os da classe 3. Ainda assim, a classe 3 possui características muito pouco prováveis para usuários legítimos. Por exemplo, dificilmente um usuário legítimo realizaria 2.800 transações em um período de 28 dias (uma média de 100 transações por dia), utilizando 1.300 remetentes distintos e acessando 1.100 servidores SMTP distintos (pouco mais de 2,5 transações para cada servidor).

Como observado para a base de 2009, a classe 3 apresenta grande variabilidade entre usuários. Por exemplo, observa-se uma variação do número de transações SMTP entre 2 e 37.286, do número de remetentes entre 1 e 21.866 e do número de servidores SMTP entre 1 e 9.248. Uma possível explicação para que usuários com baixo números de transações, remetentes e servidores tenham sido classificados como da classe 3 é que, a despeito destas características, eles apresentam valores de IAT muito baixos (no máximo 16 horas e meia) e, consequentemente, muito inferiores aos valores típicos de usuários das classes 1 e 2. Este atributo, com baixa variabilidade nas classes 1 e 2, poderia levar a classificação destes usuários com baixa atividade como sendo da classe 3.

No geral, pode-se concluir que, consistentemente nas duas bases analisadas (2009 e 2010), as classes de usuários legítimos: (i) realizam poucas transações SMTP; (ii) utilizam poucos remetentes distintos; (iii) acessam poucos servidores distintos; (iv) possuem alto intervalo de inatividade entre as transações; (v) possuem alta variabilidadade do tamanho médio das transações SMTP, uma vez que usuários legítimos podem enviar tanto mensagens apenas de texto, quanto mensagens com anexos extensos, contendo vídeos e imagens; e (vi) realizam suas transações principalmente para servidores brasileiros ou servidores localizados nos Estados Unidos.

Em contraste, as classes de *spammers*: (i) realizam um número alto de transações SMTP; (ii) utilizam um número elevado de remetentes distintos; (iii) acessam vários servidores SMTP distintos; (iv) efetuam as transações com um período de inatividade

muito baixo, sendo muitas vezes de apenas segundos; (v) possuem tamanho médio das transações SMTP baixos, uma vez que, normalmente, *spams* possuem apenas texto; e (vi) tendem a apresentar distância geodésica média maior que usuários legítimos.

Como discutido na Seção 4.2, além da avaliação dos padrões de comportamento detectados, também foi feita uma validação da classificação da base de 2010, utilizando a lista de usuários denunciados em 2010 (não utilizada como parte do treino) e inspecionando manualmente uma amostra aleatória de 5% dos usuários de cada classe (exceto classe 4). Essa taxa de amostragem garante, com confiança de 90%, um erro inferior a 12% nas estimativas [Jain 1991]. Como trabalho futuro, pretende-se realizar a inspeção manual com uma fração maior dos usuários.

Tabela 4. Eficácia da Classificação: Estimativas de Taxa de Acerto, Falsos Positivos e Falsos Negativos (Base de Dados 2010).

	Classe Predita			
Classe Real	**1**	**2**	**3**	**4**
1 (legítimo)	100%			
2 (legítimo)	0,7%	95,5%	3,8%	
3 (*spammer*)		15,9%	84,1%	
4 (*spammer*)				100%

Dos 93 usuários denunciados, 31 se encontram na classe 4 e 62 na classe 3. Ou seja, 40,78% e 7,41% dos usuários classificados nas classes 4 e 3, respectivamente, foram de fato denunciados como *spammers*. Foram inspecionados todos os 45 usuários restantes da classe 4, concluindo que todos apresentam um comportamento consistente com *spammers* bastante abusivos. Logo, todos os usuários classificados na classe 4 foram corretamente identificados como *spammers*. Além disto, verificou-se que 35 de 44 usuários selecionados da classe 3 para inspeção manual foram corretamente classificados, enquanto os 7 restantes apresentavam um comportamento aceitável para usuários legítimos da classe 2, sugerindo assim falsos positivos.

Quanto à classificação de usuários em usuários legítimos, apenas 1 de todos os usuários selecionados da classe 1 foi classificado erroneamente, apresentando um comportamento mais condizente com a classe 2. Note que, ainda assim, esse usuário foi classificado como legítimo. Já para a classe 2, 127 dos 132 usuários selecionados foram corretamente classificados, enquanto 5 tinham um padrão mais próximo de *spammers* da classe 3, sendo assim considerados falsos negativos.

Esses resultados, computados sobre os usuários amostrados das classes 1, 2 e 3 e sobre todos os usuários da classe 4, são sumarizados na Tabela 4. Cada linha representa uma classe atribuída aos usuários por inspeção ou pelo `abuse` (*classe real*), enquanto as colunas representam as classes assinaladas pelo SpaDeS (classes preditas). Os valores indicam as porcentagens das amostras de uma classe real que foram atribuídas à classe predita indicada. Dessa forma, a diagonal indica a taxa de acerto para cada classe e as demais posições indicam predições incorretas. Considerando a classificação de todos usuários nas super-classes "legítimos" (classes 1 e 2) e "*spammers*" (classes 3 e 4), o método SpaDeS apresentou uma excelente taxa de acerto, identificando corretamente 98% e 94% dos usuários legítimos e *spammers*, respectivamente, apresentando assim taxas de falsos positivos e de falsos negativos de somente 2% e 6%, respectivamente.

Os resultados acima foram obtidos para α=20% (parâmetro utilizado na geração

de treino — vide algoritmo 1, Seção 3.3). Foram também realizados experimentos com α igual a 10% e 30%. Para α=10%, os resultados se mantiveram bastante semelhantes: estima-se que 96% e 94% dos usuários legítimos e *spammers*, respectivamente, foram corretamente identificados, conforme inspeção em amostras de 5% dos usuários e lista de usuários denunciados. Isso ocorre porque a confiança na classificação dos usuários variou muito pouco: com α=10% a confiança mínima dos usuários utilizados como treino é de 65%, muito próxima daquela para α=20%. Como consequência, o número de usuários selecionados para treino é apenas ligeiramente menor. Já para α=30%, a confiança mínima cai para 58%, o que prejudica a efetividade do método: as taxas de acertos estimadas para usuários legítimos e *spammers* caem para 94% e 91%, respectivamente.

Em suma, o SpaDeS apresentou uma excelente efetividade na detecção de *spammers* na rede de origem. Não se conhece nenhum outro método que se proponha a fazer esta detecção tão próximo à fonte dos *spams*, reduzindo assim o tráfego na rede. Uma comparação do SpaDeS com outros métodos de detecção de *spammers* disponíveis na literatura e que abordam o problema em outros pontos do sistema é bastante difícil considerando as bases de dados disponíveis para validação do método proposto, uma vez que tais métodos utilizam informações diferentes das exploradas pelo SpaDeS.

5. Conclusões

Neste trabalho foi apresentado, aplicado e validado com dados reais um método para identificação e detecção de *spammers* na rede origem, mais especificamente em uma rede de acesso à Internet de banda larga. Este método, denominado SpaDeS (*Spammer Detection at the Source*), tem como principal componente um algoritmo de classificação supervisionada – *Lazy Associative Classification* (LAC) – e utiliza métricas que não requerem a inspeção do conteúdo da mensagem enviada, para classificar os usuários como sendo legítimos ou *spammers*. O SpaDeS apresentou uma excelente efetividade, com taxa de acerto de 98% para usuários legítimos e 94% na classificação de *spammers*. Como trabalho futuro, propõe-se o aprimoramento do método e a construção e validação de um sistema que viabilize o uso do SpaDeS em tempo real.

Agradecimentos

Esta pesquisa é parcialmente financiada pelo Instituto Nacional de Ciência e Tecnologia para a Web - INCTWeb (MCT/CNPq 573871/2008-6), CNPq, FAPEMIG, FAPERJ e pelo PROBIC/PUC-Minas.

Referências

Castilho, L. H. D., Las-Casas, P. H. B., Dutra, M. D., Ricci, S. M. R., Marques-Neto, H. T., Ziviani, A., Almeida, J. M., e Almeida, V. (2010). Caracterização de tráfego SMTP na Rede de Origem. Em *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Gramado, Brasil.

Gomes, L. H., Almeida, V. A. F., Almeida, J. M., Castro, F. D. O., , e Bettencourt, L. M. A. (2009). Quantifying Social And Opportunistic Behavior In Email Networks. *Advances in Complex Systems*, 12(1):99–112.

Gomes, L. H., Cazita, C., Almeida, J. M., Almeida, V., e Jr., W. M. (2007). Workload Models of Spam and Legitimate E-mails. *Performance Evaluation*, 64(7-8):690–714.

Guerra, P. H. C., Pires, D. E. V., Guedes, D., Jr., W. M., Hoepers, C., e Steding-Jessen, K. (2008). A Campaign-based Characterization of Spamming Strategies. Em *Proceedings of the Fifth Conference on Email and Anti-Spam*, pág. 1–10, Mountain View, CA, USA.

Guerra, P. H. C., Pires, D. E. V., Guedes, D., Jr., W. M., Hoepers, C., Steding-Jessen, K., e Chaves, M. (2009). Caracterização de Encadeamento de Conexões para Envio de Spams. Em *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Recife, Brasil.

Hao, S., Syed, N. A., Feamster, N., Gray, A., e Krasser, S. (2009). Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine. Em *Usenix Security*, Montreal, Canadá.

IronPort (2009). 2009 Internet Security Trends. Online. http://www.ironport.com/.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, Inc., 1st edition.

John, J., Moshchuk, A., Gribble, S. D., e Krishnamurthy, A. (2009). Studying Spamming Botnets Using Botlab. Em *6th USENIX Symp. on Networked Systems Design and Implementation*, Boston, EUA.

Kim, J. e Choi, H. (2008). Spam Traffic Characterization. Em *Int'l Technical Conference on Circuits/Systems, Computers and Communications*, Shimonoseki City, Japão.

Kolcz, A. e Alspector, J. (2001). SVM-based Filtering of E-mail Spam with Content-specifcs Misclassification Costs. Em *Workshop on Text Mining*, San Jose, EUA.

MessageLabs (2010). MessageLabs Intelligence: November 2010. Online.

Newman, M. E. J., Forrest, S., e Balthrop, J. (2002). Email Networks and the Spread of Computer Viruses. *Physical Review E*, 66(3):035101.

Pelleg, D. e Moore (2000). X-means: Extending K-means with efficient estimation of the number of clusters. Em *17th Int'l Conf. on Machine Learning*, San Francisco, USA.

Ramachandran, A. e Feamster, N. (2006). Understanding the Network-Level Behavior of Spammers. *SIGCOMM Comput. Commun. Rev.*, 36(4):291–302.

Schatzmann, D., Burkhart, M., e Spyropoulos, T. (2009). Inferring Spammers in the Network Core. Em *10th Int'l Conf. on Passive and Active Network Measurement*, Berlin, Heidelberg.

Shafranovich, Y., Levine, J., e Kucherawy, M. (2010). An Extensible Format for Email Feedback Reports. RFC 5965.

Sperotto, A., Vliek, G., Sadre, R., e Pras, A. (2009). Detecting Spam at the Network Level. Em *EUNICE'09: 15th Open European Summer School and IFIP TC6.6 Workshop on The Internet of the Future*, Barcelona, Spain.

Taveira, D. e Duarte, O. (2008). Mecanismo Anti-Spam Baseado em Autenticação e Reputação. Em *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Rio de Janeiro.

Veloso, A., Meira, W., e Zakib, M. J. (2006). Lazy associative classification. Em *Sixth International Conference on Data Mining*, Hong Kong, China.