Mecanismo Anti-*Spam* Baseado em Autenticação e Reputação

Danilo M. Taveira[1] **e Otto Carlos M. B. Duarte**[1*]

[1]Grupo de Teleinformática e Automação
Universidade Federal do Rio de Janeiro

Abstract. *False positives of an anti-spam mechanism can have a great impact for users when a legitimate message is classified as spam. This work proposes a mechanism that aims to reduce false positive rates. The key idea is to take into account the history of the user behavior on the message filtering process. An authentication and reputation mechanism is used to monitor and evaluate the user behavior. To evaluate the mechanism performance an analytical model was derived and a simulator was developed. The results show the efficiency of the proposed mechanism in reducing the false positives without increasing the false negatives.*

Resumo. *Os falsos positivos de um sistema anti-spam têm um impacto muito grande para os usuários, já que uma mensagem legítima pode ser identificada como spam e ser filtrada, causando grandes prejuízos. Dessa forma, este trabalho propõe um mecanismo anti-spam que tem como foco principal a redução dos falsos positivos. A idéia chave é decidir se a mensagem é legítima ou não a partir do histórico de comportamento dos usuários, utilizando um mecanismo de autenticação e reputação. Para avaliar o mecanismo proposto foi derivado um modelo matemático e desenvolvido um simulador. Os resultados mostram a eficiência do mecanismo proposto na redução da taxa de falsos positivos sem prejudicar a classificação dos spams.*

1. Introdução

O combate aos *spams* é um dos grandes desafios na Internet [Taveira et al., 2006]. O *spam*, de forma simplificada, é toda mensagem eletrônica não solicitada pelo destinatário. Devido à simplicidade do protocolo SMTP (*Simple Mail Transfer Protocol*), o correio eletrônico é a aplicação mais afetada pelos *spams*. As estatísticas mostram que os *spams* já correspondem a pelo menos dois terços de todo o tráfego de correio eletrônico transportado pelos provedores de serviço, causando prejuízos da ordem de milhões de dólares [Pfleeger e Bloom, 2005].

Os sistemas anti-*spam* são as principais contramedidas ao envio de mensagens não solicitadas. No entanto, estes sistemas precisam estar em constante evolução, pois para cada novo sistema desenvolvido, surgem novas técnicas para tentar burlá-lo. As principais propriedades de um mecanismo anti-*spam* são a sua taxa de falsos positivos e de falsos negativos. Entendem-se como falsos positivos todas as mensagens legítimas que são classificadas como *spams* e como falsos negativos todos os *spams* que são classificados como mensagens legítimas. Os falsos negativos têm um impacto menor, pois o usuário receberá o *spam* e poderá apagá-lo. Dessa forma, os custos com os falsos negativos estão relacionados com a perda de produtividade e a perda de foco na realização de atividades. Já a taxa de falsos positivos tem um impacto muito maior, pois uma mensagem legítima acaba

*Este trabalho foi realizado com recursos do CNPq, FINEP, RNP, FAPERJ e CAPES.

sendo perdida, gerando grandes transtornos e atrasos no processo de comunicação. Sendo assim, os custos com os falsos positivos tendem a ser altos, uma vez que informações estratégicas e oportunidades podem ser perdidas, gerando graves conseqüências profissionais e pessoais.

O mecanismo anti-*spam* proposto nesse trabalho tem como principal objetivo a redução dos falsos positivos. A idéia chave é considerar o histórico de comportamento dos usuários para decidir se a mensagem é legítima ou não. Portanto, um usuário que já enviou várias mensagens legítimas passa a ter uma probabilidade muito menor de ter suas mensagens classificadas como *spam*. Para identificar os usuários é utilizado um mecanismo de autenticação, porém sem utilizar informações pessoais, por questões de privacidade. Após a identificação do usuário, o seu histórico de comportamento pode ser monitorado. Para uma maior eficácia do mecanismo proposto, o histórico de comportamento de um usuário é determinado por diversos servidores de correio eletrônico, que trocam informações sobre esse histórico de comportamento. Na troca de informações é utilizado um mecanismo de reputação, uma vez que se deve levar em consideração a reputação de cada servidor para avaliar qual será o grau de confiança dessa informação. Os resultados mostram que o mecanismo proposto é eficiente na redução dos falsos positivos e também é robusto às contramedidas que podem ser adotadas pelos *spammers*.

Este trabalho está organizado da seguinte forma. Na Seção 2 são apresentados alguns trabalhos relacionados. Na Seção 3 é apresentado o mecanismo anti-*spam* baseado no histórico de comportamento dos usuários proposto nesse trabalho. A Seção 4 apresenta o modelo matemático do mecanismo proposto. O ambiente de simulação e os parâmetros utilizados são descritos na Seção 5. As análises dos resultados são apresentadas na Seção 6. Por fim, na Seção 7, são apresentadas as conclusões deste trabalho.

2. Trabalhos Relacionados

Atualmente vários mecanismos anti-*spam* são utilizados tais como as listas negras, listas cinza, filtros bayesianos e mecanismos baseados em pesos e regras [Taveira et al., 2006]. No entanto, esses mecanismos geralmente apresentam taxas de falsos positivos relativamente altas. Através da avaliação dos mecanismos tradicionais utilizando uma ferramenta de avaliação de mecanismos anti-*spam* observa-se que mesmo o mecanismo mais preciso possui uma taxa de falsos positivos de 2,3%, que é relativamente elevada [Taveira et al., 2008]. Assim, novas técnicas que se servem de reputação ou redes sociais vêm sendo adotadas, em adição aos mecanismos convencionais, para tomar a decisão final, aumentando a precisão da classificação.

Golbeck e Hendler propõem um mecanismo de reputação baseado em redes sociais [Golbeck e Hendler, 2004]. As redes sociais correspondem ao grafo que representa a comunicação entre usuários. Cada usuário representa um nó do grafo e uma aresta entre dois nós significa que os usuários já trocaram mensagens. Nesse mecanismo, cada usuário define a reputação dos usuários com os quais ele troca mensagens. Quando uma mensagem de um usuário desconhecido é recebida, procura-se no grafo da rede social se existe algum caminho entre o usuário que recebeu a mensagem e o usuário que enviou. Caso exista o caminho, a reputação de cada nó é levada em consideração para determinar a reputação do usuário que era desconhecido. O inconveniente desse mecanismo é que cada usuário deve atribuir manualmente a reputação dos outros usuários.

Chirita *et al.* definem um mecanismo chamado MailRank que também se baseia em redes sociais e reputação para classificar as mensagens [Chirita et al., 2005]. Toda vez que um usuário A envia uma mensagem para um usuário B, considera-se que o usuário A confia no usuário B. Assim, através da rede social, é possível determinar a reputação de todos os usuários. A desvantagem desse mecanismo é que a reputação dos usuários é calculada de forma centralizada por um servidor que possui acesso à rede social.

Balasubramaniyan *et al.* propõem um mecanismo baseado em redes sociais para combater *spams* em VoIP [Balasubramaniyan et al., 2007]. No entanto, a reputação de cada usuário é determinada através do tempo das ligações VoIP entre os usuários. Os dois mecanismos baseados em redes sociais se baseiam na avaliação da rede social dos usuários de diferentes servidores. Esta proposta elimina a privacidade dos usuários, pois a rede social revela com quem todos os usuários se comunicam.

Seigneur *et al.* propõem um mecanismo de reputação baseado em redes sociais com um mecanismo de autenticação [Seigneur et al., 2004]. O mecanismo de autenticação envia para o destinatário a nova mensagem juntamente com o resumo das mensagens enviadas anteriormente. Dessa forma, é possível identificar se o usuário que está enviando a nova mensagem é o mesmo usuário que já enviou mensagens passadas. No entanto, mensagens enviadas podem chegar atrasadas ou serem perdidas, prejudicando a comparação das mensagens enviadas pelo remetente e recebidas pelo destinatário.

McGibney e Botvich definem um mecanismo chamado TOPAS que se baseia na reputação dos servidores [McGibney e Botvich, 2007]. A reputação dos servidores é determinada a partir da classificação das mensagens através de um mecanismo anti-*spam* auxiliar. O valor da reputação é utilizado para ajustar o limiar de classificação do mecanismo auxiliar e tomar a decisão final se a mensagem será aceita ou não. A reputação do servidor é calcula utilizando uma média móvel exponencial. A cada mensagem recebida, a média é recalculada utilizando o valor um, caso a mensagem seja legítima, ou zero, caso seja *spam*. O valor da reputação de cada servidor inicia em 0,5. O limiar final é calculado como sendo 10 vezes o valor da reputação. Para determinar a reputação dos servidores são utilizadas informações locais e recomendações recebidas de outros servidores. Cada servidor também pode requisitar as reputações observadas por outros servidores. Esse mecanismo, no entanto, tem a desvantagem apenas considerar a reputação dos servidores, permitindo que *spammers* se aproveitem da reputação de servidores legítimos.

O mecanismo proposto neste artigo objetiva a determinação da reputação dos remetentes das mensagens sem violar sua privacidade e também não permitir que os *spammers* se aproveitem de servidores de usuários legítimos com boa reputação.

3. Mecanismo Proposto

O mecanismo anti-*spam* proposto utiliza um mecanismo de autenticação dos remetentes e um mecanismo de reputação para a troca de informações entre os servidores sobre os usuários autenticados. A reputação dos usuários é determinada a partir do histórico de mensagens enviadas e, quanto melhor for a reputação, menor será a probabilidade das mensagens serem classificadas como *spam*, diminuindo os falsos positivos.

Para avaliar o histórico de comportamento dos usuários, o mecanismo de reputação deve ser capaz de identificá-los. Para isso é necessário um mecanismo de autenticação dos remetentes, pois o protocolo SMTP não possui nenhum mecanismo de

autenticação dos remetentes, que podem ser facilmente forjados [Taveira et al., 2006]. A técnica de autenticação convencional utiliza certificados digitais emitidos por autoridades certificadoras. No entanto, a certificação é dispendiosa e elimina a privacidade dos usuários, pois os certificados possuem dados pessoais dos usuários. Assim, este artigo propõe um mecanismo de autenticação dos usuários sem utilizar informações pessoais, garantindo a privacidade. Para isso, o mecanismo de autenticação proposto baseia-se em pseudônimos atribuídos a cada usuário. Os usuários podem possuir um ou mais pseudônimos para se comunicarem com diferentes grupos de pessoas. No caso de listas de mensagens, cada lista possui um pseudônimo e quando algum usuário enviar uma mensagem para a lista, essa mensagem será enviada para os destinatários da lista utilizando o pseudônimo da lista. Os pseudônimos utilizados no processo de autenticação são compostos por um par de chaves assimétricas e podem ser gerados pelo próprio usuário.

O processo de autenticação é baseado no modelo desafio-resposta e não requer alterações no protocolo SMTP. O mecanismo de autenticação proposto é feito de forma totalmente automatizada e, portanto, sem a intervenção do usuário. Para iniciar o processo de autenticação, o cliente antes de enviar a mensagem ao seu servidor de correio eletrônico envia um pedido de autenticação ao servidor de correio eletrônico de cada destinatário da mensagem. Cada um dos servidores de destino envia para o cliente um desafio que é simplesmente uma seqüência de caracteres. O servidor armazena em uma tabela o desafio enviado para, posteriormente, verificar se a resposta corresponde a um desafio enviado pelo servidor. O cliente responde aos desafios assinando digitalmente cada um dos desafios com a sua chave privada e adiciona no cabeçalho da mensagem a ser enviada uma linha contendo sua chave pública e uma linha para cada uma das respostas dos desafios. Utilizando o protocolo SMTP, a mensagem é enviada para o servidor de correio eletrônico do remetente, que a encaminha até os servidores dos destinatários da mensagem. A partir desse ponto, cada servidor dos destinatários verifica se alguma das respostas é de um desafio enviado por ele. Caso exista a resposta para o desafio, o pseudônimo só é autenticado se a assinatura digital for verificada. Caso não ocorra sucesso nas verificações da resposta e da assinatura digital, o processo de autenticação falhará. Portanto, depois do processo de autenticação, a chave pública do pseudônimo pode ser utilizada como um identificador do pseudônimo. Utilizando esse identificador, o servidor pode buscar informações tanto localmente quanto remotamente através do mecanismo de reputação para determinar o histórico de comportamento do pseudônimo e assim determinar a reputação.

O mecanismo de reputação proposto troca informações entre os servidores sobre os pseudônimos já autenticados. Para realizar essa troca de informações, cada servidor possui uma determinada reputação em outros servidores, o que vai determinar a confiança na informação vinda desses outros servidores. A reputação de um servidor é determinada a partir do total de mensagens legítimas e *spams* que foram enviadas por esse servidor. Nessa etapa, para classificar as mensagens, é utilizado um mecanismo anti-*spam* convencional (listas, filtros bayesianos, pesos e regras etc.) como mecanismo auxiliar. Caso a mensagem seja classificada como legítima, a média do servidor, que representa sua reputação, é atualizada com o valor 1. Caso a mensagem seja classificada como *spam*, a média é atualizada com o valor -1. Para o cálculo da reputação utiliza-se uma média móvel exponencial, que considera pesos maiores aos valores observados mais recentemente e os pesos dos elementos mais antigos decrescem exponencialmente. Dessa forma, caso um servidor mude de comportamento, a média mudará de valor mais rapidamente. A

média móvel exponencial $M(t)$ de um valor $V(t)$ é expressa pela Equação 1. O Parâmetro N é o período da média que representa o número de amostras mais significativas[1].

$$M(t) = \frac{2}{N+1} \cdot V(t) + \left(1 - \frac{2}{N+1}\right) \cdot M(t-1) \quad (1)$$

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

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

O mecanismo proposto utiliza duas tabelas para armazenar as reputações dos pseudônimos e dos servidores que são mostradas nas Tabelas 1(a) e 1(b), com o tamanho em bits de cada campo. A identificação dos pseudônimos é através da sua chave pública. Já os servidores são identificados através do endereço IP. O campo de tempo é utilizado para remover entradas que não foram atualizadas por um longo período. O limiar para remover as entradas antigas pode ser escolhido de acordo com a capacidade de armazenamento do servidor. O tamanho total das tabelas é dado por $576N_P + 96N_S$, onde N_P e N_S são o número de entradas na tabela de pseudônimos e servidores, respectivamente. Para armazenar a reputação de 10 milhões de usuários e 10 milhões de servidores são necessários apenas 840 Megabytes, que não é uma capacidade de armazenamento alta para um servidor. Além disso, como o mecanismo proposto pode reduzir a quantidade de *spams* e a taxa de falsos positivos, a redução dos gastos tende a ser maior do que o gasto adicional devido ao mecanismo proposto.

Tabela 1. Tabela de reputação de pseudônimos e servidores.

(a) Reputação dos pseudônimos.

Chave pública	Reputação	Tempo
512 bits	32 bits	32 bits

(b) Reputação dos servidores.

Endereço IP	Reputação	Tempo
32 bits	32 bits	32 bits

Para avaliar a reputação de um pseudônimo, o servidor consulta outros servidores que informam a reputação do pseudônimo observada por eles. A informação de cada servidor sobre a reputação do pseudônimo é multiplicada pela reputação do servidor que foi observada localmente. A única exceção é quando ambas as reputações, do servidor e do pseudônimo, são negativas. Nesse caso, a multiplicação desses dois valores resulta em um valor positivo, tendo que ser multiplicado por -1 para resultar novamente em um valor negativo. Assume-se que um servidor malicioso não é utilizado por usuários legítimos para enviar mensagens legítimas, o que fará com que o servidor tenha uma baixa reputação. Dessa forma, um servidor malicioso que responda que todos os pseudônimos possuem reputação máxima não será considerado, pois a sua reputação será baixa.

Na troca de informações, não é viável consultar todos os servidores da Internet devido ao grande número de mensagens que seriam necessárias. Assim, apenas um con-

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

junto pequeno de servidores é consultado. O número de servidores consultados é definido pelo parâmetro N_c. A estratégia para escolher os servidores que são consultados pode variar. Podem ser escolhidos, por exemplo, os servidores com maior reputação ou então os últimos servidores que enviaram mensagens. Uma característica importante do mecanismo proposto é que a adoção pode ser incremental, o que é fundamental na Internet. Mesmo que apenas alguns servidores adotem o mecanismo, já é possível a troca de informações entre eles. Como são consultados N_c servidores além de considerar a reputação observada localmente, o valor da reputação máximo será $N_c + 1$ e o valor mínimo será $-(N_c + 1)$. Assim, quanto maior for o número de servidores consultados maior será a variação do valor da reputação, já que mais informações são utilizadas para calcular a reputação e o valor da reputação não é normalizado.

Após consultar os servidores, o servidor i calcula a reputação final R_f do pseudônimo j através da Equação 2, onde S_c é o conjunto de servidores consultados, $R_S(a, b)$ representa a reputação que o servidor a observou do servidor b, $R_P(c, d)$ representa a reputação que o servidor c observou do pseudônimo d e a função $f_a(x, y)$ é igual a -1 caso x e y sejam negativos e igual 1 caso contrário. Caso nenhum dos servidores consultados possua informações sobre o pseudônimo, a reputação será zero.

$$R_f(j) = R_P(i, j) + \sum_{l \in S_c} R_S(i, l) \cdot R_P(l, j) \cdot f_a(R_S(i, l), R_P(l, j)) \quad (2)$$

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

O mecanismo anti-*spam* auxiliar que é utilizado para determinar a reputação pode ser o mesmo mecanismo de pesos e regras só que utilizando o limiar padrão. Depois da reputação do remetente da mensagem ser calculada, avalia-se novamente se a mensagem deve ser classificada como *spam* ou não, de acordo com o novo limiar calculado com base na reputação. Dessa forma, um *spammer* que não utiliza um pseudônimo não poderá se aproveitar da reputação dos servidores legítimos. Se o *spammer* utilizar um pseudônimo p diferente para cada mensagem, $R_P(j, p)$ será sempre zero para todos os servidores j e a reputação de acordo com a Equação 2 também será zero. Assim, a decisão tomada é a mesma do mecanismo auxiliar. Se fosse utilizada a reputação do servidor como acontece no TOPAS, o *spammer* conseguiria se aproveitar da reputação dos servidores legítimos.

A proposta também é robusta ao roubo de pseudônimos de usuários legítimos. Caso a máquina de um usuário seja invadida por alguma praga digital, o pseudônimo do usuário pode ser utilizado por *spammers*, que se aproveitam da boa reputação do pseudônimo para enviar *spams*. Porém, a reputação do pseudônimo irá diminuir com o tempo, tornando-se inútil para os *spammers* e também para o usuário legítimo, que deverá realizar a troca do pseudônimo. Para realizar essa troca de forma automática, é utilizado um mecanismo de proteção. Esse mecanismo consulta a reputação do pseudônimo

no servidor para o qual o usuário está enviando uma mensagem. O servidor realiza todos os procedimentos para o cálculo da reputação do pseudônimo e retorna esse valor para o cliente. Caso o valor da reputação seja baixo, o usuário troca automaticamente seu pseudônimo. Caso a máquina do usuário ainda continue invadida por alguma praga digital, o novo pseudônimo obtido poderá ser utilizado. No entanto, como foram enviadas poucas mensagens com o pseudônimo novo, a reputação desse pseudônimo diminuirá ainda mais rápido do que a do pseudônimo original, não trazendo benefício para os *spammers*. Mesmo se um *spammer* realizar um ataque de força bruta consultando vários pseudônimos até descobrir algum com boa reputação, ele não conseguirá a chave privada do pseudônimo, tornando inútil esse tipo de ataque.

4. Modelo do Mecanismo

Para avaliar analiticamente a eficiência do mecanismo proposto, essa seção propõe um modelo simplificado do mecanismo proposto e do mecanismo TOPAS. Na análise, será considerado que o sistema está em estado estacionário, ou seja, todos os servidores já receberam um número de mensagens suficiente para avaliar a reputação dos outros servidores e dos pseudônimos.

O mecanismo anti-*spam* auxiliar considerado na análise é um mecanismo baseado em pesos e regras, similar ao mecanismo anti-*spam* mais utilizado na Internet, chamado *SpamAssassin*. Nesse mecanismo, existem regras que testam se características de mensagens legítimas e *spams* estão presentes na mensagem. Cada regra possui um peso associado, que pode ser positivo ou negativo. O somatório dos pesos das regras que correspondem às características da mensagem define o grau da mensagem ser *spam* ou não. Um limiar ρ é utilizado para determinar que mensagens com grau acima do limiar sejam classificadas como *spam*. Neste trabalho, o grau das mensagens legítimas e *spams* é modelado de acordo com uma distribuição normal com desvio padrão σ. Dessa forma, fixando-se o limiar e alterando a média dos graus das mensagens legítimas e *spams* pode-se alterar a taxa de falsos positivos e falsos negativos do mecanismo auxiliar. Assim, variando a média do grau das mensagens legítimas (μ_L) e *spams* (μ_S), diferentes taxas de falsos positivos e falsos negativos podem ser obtidas. Na prática, a variação da média é provocada pela utilização de regras com maior ou menor precisão. A probabilidade de ocorrência de falsos positivos é igual à probabilidade do grau da mensagem avaliada ser maior do que o limiar e pode ser expressa pela Equação 3, onde $erf(x)$ é a função erro da distribuição normal padrão. De forma similar, a taxa de falsos negativos pode ser expressa pela Equação 4.

$$FP(\mu_L, \rho, \sigma) = \frac{1}{2} - \frac{1}{2} erf\left(\frac{\rho - \mu_L}{\sigma\sqrt{2}}\right) \tag{3}$$

$$FN(\mu_S, \rho, \sigma) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\rho - \mu_S}{\sigma\sqrt{2}}\right) \tag{4}$$

Considerando as probabilidades de ocorrência de falsos positivos e falsos negativos do mecanismo auxiliar p_{fp} e p_{fn}, a probabilidade de uma mensagem enviada por um usuário com um pseudônimo P_i ser classificada como legítima pelo mecanismo auxiliar é dada pela Equação 5. Nessa equação, el_{P_i} é a probabilidade do usuário que possui o pseudônimo P_i enviar uma mensagem legítima e es_{P_i} é a probabilidade do usuário enviar um *spam*. De forma análoga, a probabilidade de uma mensagem enviada pelo usuário que

possui o pseudônimo P_i ser classificada como *spam* é dada pela Equação 6. A reputação do pseudônimo P_i é dada por $R_{P_i} = pl_{_{P_i}} - ps_{_{P_i}}$.

$$pl_{_{P_i}} = (1 - p_{fp}) \cdot el_{_{P_i}} + p_{fn} \cdot es_{_{P_i}} \tag{5}$$

$$ps_{_{P_i}} = p_{fp} \cdot el_{_{P_i}} + (1 - p_{fn}) \cdot es_{_{P_i}} \tag{6}$$

As probabilidades de um servidor S_i enviar mensagens legítimas e *spams* são dadas pelas Equações 7 e 8, onde U_{S_i} é o conjunto de usuários que utilizam o servidor S_i para enviar mensagens, λ_i é a taxa de mensagens enviadas pelo usuário i e N é o número de elementos do conjunto U_{S_i}.

$$el_{_{S_i}} = \frac{1}{N \sum_{k \in U_{S_i}} \lambda_k} \sum_{j \in U_{S_i}} \lambda_j \cdot el_{_{P_j}} \tag{7}$$

$$es_{_{S_i}} = \frac{1}{N \sum_{k \in U_{S_i}} \lambda_k} \sum_{j \in U_{S_i}} \lambda_j \cdot es_{_{P_j}} \tag{8}$$

As probabilidades das mensagens do servidor S_i serem classificadas como legítimas e *spams* são dadas pelas Equações 9 e 10. A reputação do servidor S_i é então dada por $R_{S_i} = pl_{_{S_i}} - ps_{_{S_i}}$.

$$pl_{_{S_i}} = (1 - p_{fp}) \cdot el_{_{S_i}} + p_{fn} \cdot es_{_{S_i}} \tag{9}$$

$$ps_{_{S_i}} = p_{fp} \cdot el_{_{S_i}} + (1 - p_{fn}) \cdot es_{_{S_i}} \tag{10}$$

Para avaliar o impacto dos *spammers* que tentam se beneficiar da reputação dos servidores legítimos é considerado um percentual de *spammers* p_{sl} que envia mensagens através desses servidores. O número de usuários legítimos N_L será considerado igual a metade do número de *spammers* N_S. A taxa de mensagens que os usuários legítimos enviam será considerada igual à taxa de mensagens enviada pelos *spammers*. Nesse caso as Equações 7 e 8 podem ser simplificadas para $el_{_S} = \frac{p_{sl} \cdot N_S}{N_L + p_{sl} \cdot N_S}$ e $es_{_S} = \frac{N_L}{N_L + p_{sl} \cdot N_S}$ para os servidores legítimos.

A reputação dos servidores e dos pseudônimos legítimos é dada pelas Equações 11 e 12, considerando $el_{_{P_i}} = 1$ e $es_{_{P_i}} = 0$. Como está sendo considerado que todos os servidores já receberam mensagens suficientes para avaliar os pseudônimos e os outros servidores, a reputação de um dado pseudônimo e um dado servidor é a mesma em todos os servidores.

$$R_{S_{leg}} = ((1 - p_{fp}) \cdot el_{_S} + p_{fn} \cdot es_{_S}) - (p_{fp} \cdot el_{_S} + (1 - p_{fn}) \cdot es_{_S}) \tag{11}$$

$$R_P = 1 - 2p_{fp} \tag{12}$$

Considerando que a taxa de falsos positivos p_{fp} é menor do que 0,5, a reputação do pseudônimo é sempre maior do que zero. Dessa forma, nunca acontece, para o usuário legítimo, o caso em que as reputações do servidor e do pseudônimo são negativas, que deve ser tratado de forma diferente.

A reputação final dos pseudônimos dos usuários legítimos é dada de acordo com a Equação 2. O valor final da reputação dos pseudônimos de usuários legítimos é dado por $R_f = R_P + N_c R_{S_{leg}} R_P$. Portanto, a probabilidade da ocorrência de falsos positivos no mecanismo proposto é dada pela Equação 3, utilizando o limiar que classifica as mensagens, dado pela Equação 13.

$$\rho_{leg} = (R_P + N_c R_{S_{leg}} R_P + 1)\rho_p \tag{13}$$

O mecanismo proposto foi comparado com o TOPAS. No TOPAS é considerada apenas a reputação do servidor que envia as mensagens. A reputação do servidor legítimo é dada pela Equação 14. Assim, o limiar para o mecanismo TOPAS é dado pela Equação 15 de acordo com a descrição feita na Seção 2.

$$R_{S_{leg}} = (1 - p_{fp}) \cdot el_{_S} + p_{fn} \cdot es_{_S} \quad (14)$$

$$\rho_{leg} = 10\left((1 - p_{fp}) \cdot el_{_S} + p_{fn} \cdot es_{_S}\right) \quad (15)$$

Assumindo que os *spammers* não utilizam pseudônimos, a reputação é zero e o limiar padrão no mecanismo proposto é utilizado na classificação. A taxa de falsos negativos do mecanismo proposto é dada pela Equação 4 utilizando o limiar padrão.

Para o mecanismo TOPAS, o limiar adotado para os *spams* enviados através de servidores legítimos é dado pela Equação 15. Quando as mensagens são enviadas através de servidores que enviam apenas *spam*, a reputação desses servidores é dada por $R_{S_{spam}} = p_{fn}$. O limiar adotado nesse caso é dado pela Equação 16.

$$\rho_{spam} = 10p_{fn} \quad (16)$$

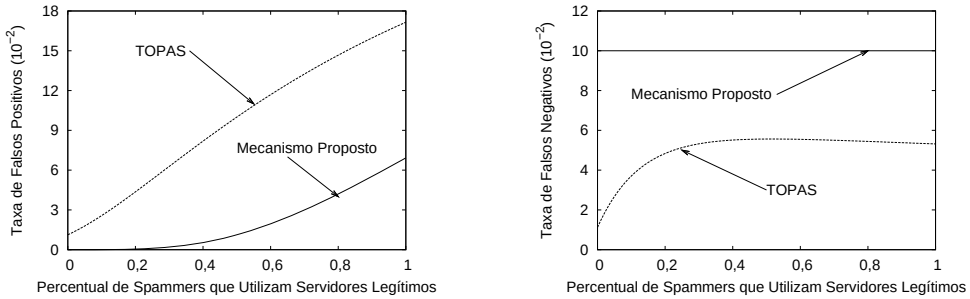
A taxa de falsos negativos do mecanismo TOPAS é diferente se o *spam* é enviado através de um servidor legítimo ou através de um servidor que envia apenas *spams*. Portanto, se a mensagem é enviada através de um servidor legítimo, a taxa de falsos negativos FN/L é dada pela Equação 4, utilizando o limiar da Equação 15. Caso o *spam* seja enviado através de um servidor que envia apenas *spam*, a taxa de falsos negativos FN/S é dada pela Equação 4, utilizando o limiar da Equação 16. A taxa final de falsos negativos é dada por $FN/L \cdot p_{sl} + FN/S \cdot (1 - p_{sl})$.

A Figura 1 mostra a taxa de falsos positivos e falsos negativos em função do percentual de *spammers* que utilizam servidores legítimos. As taxas de falsos positivos e falsos negativos do mecanismo auxiliar são de 10%, representando um mecanismo auxiliar que possui uma precisão relativamente baixa. No mecanismo proposto são consultados três servidores a cada vez. Pode-se observar a maior eficiência do mecanismo proposto em relação à taxa de falsos positivos que é sempre menor do que a do mecanismo TOPAS. A taxa de falsos negativos do mecanismo TOPAS é menor do que o mecanismo proposto, já que no mecanismo proposto a taxa de falsos negativos acaba sendo a taxa de falsos negativos do mecanismo auxiliar e, no mecanismo TOPAS a reputação dos servidores é levada em consideração. Inicialmente, a taxa de falsos negativos do mecanismo TOPAS aumenta, porém a taxa diminui depois. Isto ocorre porque, inicialmente, os *spammers* conseguem se aproveitar da reputação dos servidores legítimos, mas com o aumento da utilização por *spammers*, a reputação dos servidores legítimos acaba sendo reduzida e os falsos negativos diminuem e os falsos positivos aumentam.

5. Ambiente de Simulação

Para avaliar o mecanismo proposto foi desenvolvido um simulador de eventos discretos na linguagem C++. Para realizar a comparação com mecanismos que levam em conta apenas a reputação no servidor também foi implementado o mecanismo TOPAS descrito na Seção 2.

Nos testes foram usados 50 usuários legítimos e 100 *spammers* que enviam as mensagens. Todos os dois tipos de usuários enviam mensagens na mesma taxa média.



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

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

Figura 1. Influência da utilização de servidores legítimos por *spammers* avaliada através do modelo.

Com isso o percentual médio de *spams* é de $2/3$ das mensagens, que é aproximadamente o observado na prática [Pfleeger e Bloom, 2005]. O intervalo entre o envio de cada mensagem pelos usuários possui uma distribuição exponencial com média de 0,5 unidades de tempo. O tempo total de simulação é de 60.000 unidades de tempo. A distribuição e a taxa de mensagens que os usuários geram não têm um impacto grande nos resultados da simulação, uma vez que o mecanismo proposto é baseado no percentual de mensagens legítimas e *spams* e não no volume ou taxa das mensagens. O valor das taxas de mensagem e o tempo de simulação foram escolhidos apenas de tal forma que o número de mensagens recebidas por cada servidor seja grande, para que a reputação seja avaliada de forma mais precisa.

Nas simulações são utilizados 50 servidores de usuários legítimos e 100 servidores de usuários que enviam *spam*. Os servidores são separados em dois tipos, pois os servidores legítimos geralmente possuem medidas para evitar que *spammers* os utilizem. Na prática, a distribuição dos usuários legítimos não é igual para cada servidor, pois poucos servidores possuem muitos usuários e muitos servidores possuem poucos usuários. Dessa forma, para modelar essa distribuição dos usuários nos servidores foi utilizada uma distribuição Zipf. Com essa distribuição, a probabilidade de um usuário estar em um servidor i é dada pela Equação 17, onde N é o número total de servidores e s é um parâmetro da distribuição. Quanto maior o valor de s maior será a probabilidade dos usuários se concentrarem em poucos servidores. Dessa forma, a concentração dos usuários nos primeiros servidores será maior, simulando a distribuição dos usuários na prática. Nas simulações é utilizada uma distribuição Zipf com parâmetro $s = 1$. Assim, alguns servidores não possuem usuários que enviam mensagens. No entanto, esses servidores recebem mensagens dos usuários que enviam mensagens dos outros servidores. A maioria dos *spammers* utiliza servidores diferentes, que podem ser servidores invadidos, mal-configurados ou máquinas zumbis [Ramachandran e Feamster, 2006]. Devido a essa diferença, os *spammers* são distribuídos nos servidores que enviam *spam* através de uma distribuição uniforme. Como os *spammers* também utilizam servidores que possuem usuários legítimos, foi definido um parâmetro que determina o percentual de *spammers* que utilizam os servidores legítimos. Os *spammers* que utilizam os servidores legítimos são distribuídos de acordo com a mesma distribuição Zipf utilizada para a distribuição dos usuários legítimos.

$$\frac{1/i^s}{\sum_{k=1}^{N} 1/k^s} \tag{17}$$

A reputação local dos pseudônimos calculada pelo servidor é determinada conforme descrito na Seção 3, utilizando a média móvel exponencial com parâmetro $N = 50$. Já para a reputação dos servidores, é utilizado o parâmetro $N = 500$. O valor é diferente, pois a média do servidor é influenciada pelas mensagens enviadas por todos os usuários do servidor. Dessa forma, adota-se um período maior para a média do servidor, para considerar mais mensagens no cálculo da média.

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

O mecanismo auxiliar usado na simulação é o mesmo descrito na Seção 4. O valor do limiar ρ utilizado é igual a 5, que é o limiar padrão do mecanismo *SpamAssassin*, e o desvio padrão σ da distribuição normal utilizado é igual a 4.

6. Resultados

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

A Figura 2 mostra a taxa de falsos positivos do mecanismo proposto de acordo com o número de servidores que são consultados e de acordo com a estratégia de escolha dos servidores que são consultados. A estratégia de consulta apenas localmente possui o pior resultado, já que menos informações estão disponíveis. Nessa curva, independente do parâmetro de número de servidores consultados, a consulta é apenas local. A estratégia que obteve o melhor resultado foi a estratégia de consultar os servidores com maior reputação, já que no cálculo da reputação são considerados os maiores valores de reputação dos servidores. A estratégia de consultar os últimos servidores é pior que a anterior, mas a diferença diminui com o número de consultas, já que ocorre um aumento na quantidade de reputações consideradas. Por isso, foi escolhido um valor de três consultas por vez para as simulações mostradas a seguir. Para esse valor escolhido, a taxa de falsos positivos diminuiu de 10% para 0,018%, seguindo a estratégia de consultar os servidores com maior reputação. Nas simulações seguintes, a estratégia de consulta aos servidores com maior reputação sempre será adotada. A taxa de falsos negativos é sempre a mesma do mecanismo auxiliar, uma vez que os *spammers* não utilizam pseudônimos.

A Figura 3 mostra a taxa de falsos positivos e falsos negativos do mecanismo proposto e do mecanismo TOPAS, em função da taxa de falsos positivos e falsos negativos do mecanismo auxiliar. O mecanismo proposto possui uma taxa de falsos positivos menor

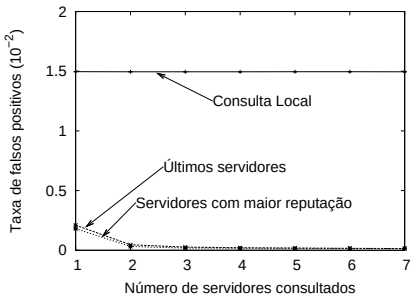
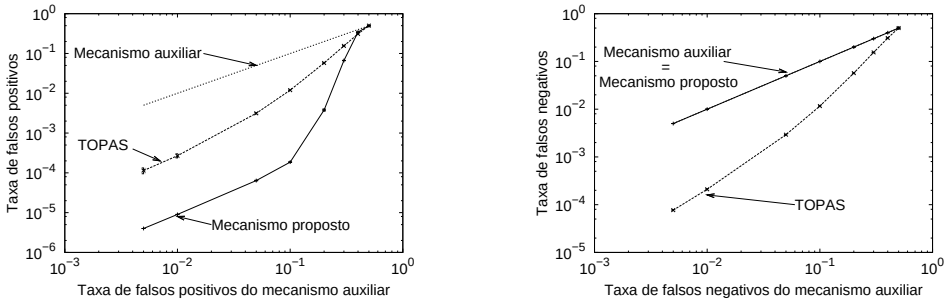


Figura 2. Relação entre falsos positivos e número de servidores consultados.

do que a do mecanismo TOPAS, até quando a taxa de falsos positivos do mecanismo auxiliar é menor do que 20%. Quando a taxa de falsos positivos é de 0,5%, o mecanismo proposto reduz essa taxa para 0,0004%, ou seja, ocorre uma redução de 1250 vezes. Já para a taxa de falsos positivos de 10%, a redução é de 537 vezes. Para valores de falsos positivos do mecanismo auxiliar acima de 30%, a eficiência dos mecanismos passa a ser pequena, uma vez que a avaliação do histórico de comportamento fica comprometida devido à alta taxa de falsos positivos. Conforme explicado anteriormente, o mecanismo proposto não diminui a taxa de falsos negativos do mecanismo auxiliar. Já o TOPAS possui uma taxa de falsos negativos menor, pois considera a reputação dos servidores.

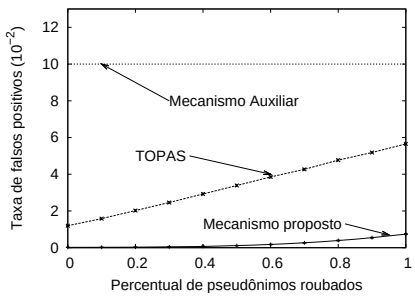


(a) Relação entre falsos positivos e falsos positivos do mecanismo auxiliar.

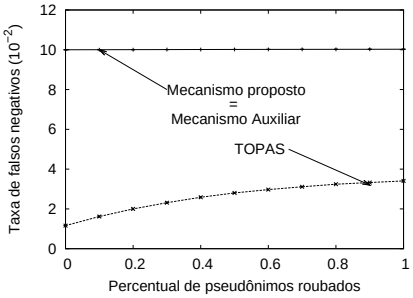
(b) Relação entre falsos negativos e falsos negativos do mecanismo auxiliar.

Figura 3. Avaliação dos falsos positivos e falsos negativos.

A Figura 4 mostra a relação entre o percentual de pseudônimos roubados e as taxas de falsos positivos e falsos negativos. O roubo dos pseudônimos acontece na metade do tempo de simulação e os usuários legítimos continuam a usar os pseudônimos roubados até que o mecanismo de proteção troque-os. O mecanismo proposto obteve a menor taxa de falsos positivos, até mesmo quando todos os pseudônimos legítimos foram roubados. A taxa de falsos positivos do mecanismo proposto aumentou para apenas 0,74% quando todos os pseudônimos foram roubados. Já o mecanismo TOPAS teve um aumento da taxa de falsos positivos para 5,65%. Apesar do mecanismo TOPAS não utilizar pseudônimos, a taxa de falsos positivos aumenta, pois quando ocorre o roubo do pseudônimo, o *spammer* utiliza o servidor legítimo que estava sendo usado pelo usuário legítimo. Dessa forma, os servidores legítimos terão sua reputação reduzida no mecanismo TOPAS. O mecanismo TOPAS alcançou uma taxa de falsos negativos menor do que o mecanismo proposto. No entanto, no mecanismo proposto a taxa de falsos negativos não aumenta, já que os *spammers* não conseguem se aproveitar da reputação dos servidores legítimos.



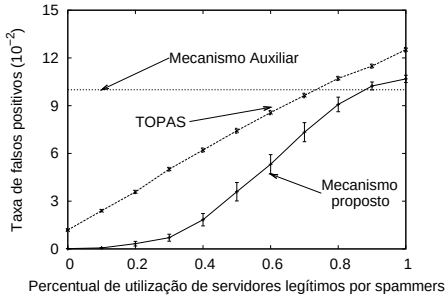
(a) Relação entre falsos positivos e pseudônimos roubados.



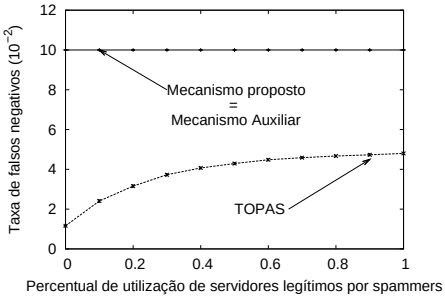
(b) Relação entre falsos negativos e pseudônimos roubados.

Figura 4. Influência dos pseudônimos roubados.

A Figura 5 mostra a influência do percentual de *spammers* que utilizam os servidores legítimos nas taxas de falsos positivos e falsos negativos. O comportamento observado na simulação foi o mesmo observado através do modelo descrito na Seção 4. O mecanismo TOPAS possui uma taxa de falsos positivos maior do que o mecanismo proposto. É importante observar que o mecanismo TOPAS possui uma taxa de falsos positivos superior à do mecanismo auxiliar quando o percentual de *spammers* que utilizam servidores legítimos é maior do que 60%. Já o mecanismo proposto, além de possuir uma taxa de falsos positivos menor, não supera a taxa de falsos positivos do mecanismo auxiliar, mesmo quando 80% dos *spammers* utilizam os servidores legítimos. A taxa de falsos negativos do mecanismo proposto é igual à taxa de falsos negativos do mecanismo auxiliar, porém maior do que a do mecanismo TOPAS. A taxa de falsos negativos do TOPAS possui o mesmo comportamento de aumentar inicialmente e depois diminuir que foi observado e explicado através do modelo na Seção 4.



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



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

Figura 5. Influência da utilização de servidores legítimos por *spammers*.

7. Conclusão

Neste artigo, é proposto um mecanismo anti-*spam* baseado em autenticação e reputação. O mecanismo de autenticação permite que usuários possam se autenticar através de pseudônimos, mantendo a privacidade do usuário. O mecanismo proposto utiliza as informações do histórico de comportamento dos pseudônimos para aprimorar a classificação das mensagens. O mecanismo de reputação permite que informações sobre os pseudônimos sejam trocadas entre os servidores, aumentando a quantidade de informação disponível para classificar a mensagem.

O modelo analítico derivado mostra que a utilização de mecanismos que consideram apenas a reputação do servidor não é eficiente, pois os *spammers* acabam se aproveitando da boa reputação dos servidores legítimos para enviar *spams* e tornar mais difícil a classificação. A utilização de pseudônimos no mecanismo proposto garante que apenas os usuários com boa reputação possuam uma menor taxa de falsos positivos. Além disso, no mecanismo proposto, os *spammers* que não possuem pseudônimos com boa reputação não conseguem se aproveitar da boa reputação dos servidores legítimos.

As simulações mostram a eficiência do mecanismo proposto, que reduz em 537 vezes a taxa de falsos positivos quando a taxa de falsos positivos do mecanismo auxiliar é de 10%, ou seja, uma redução de 99,81%. As simulações também mostram a robustez do mecanismo proposto em relação ao roubo de pseudônimos de usuários legítimos. Mesmo com altas taxas de roubo de pseudônimos, a taxa de falsos positivos não aumentou consideravelmente e a taxa de falsos negativos também não aumentou, já que a reputação do pseudônimo diminui muito rapidamente quando o *spammer* começa a enviar *spams* com o pseudônimo roubado, tornando-o inútil.

Referências

Balasubramaniyan, V. A., Ahamad, M. e Park, H. (2007). CallRank: Combating SPIT using call duration, social networks and global reputation. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)*.

Chirita, P.-A., Diederich, J. e Nejdl, W. (2005). Mailrank: using ranking for spam detection. Em *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, páginas 373–380, New York, NY, USA. ACM.

Golbeck, J. e Hendler, J. (2004). Reputation network analysis for email filtering. Em *Proceedings of the Conference on Email and Anti-Spam (CEAS)*.

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

Pfleeger, S. L. e Bloom, G. (2005). Canning spam: Proposed solutions to unwanted email. *IEEE Security & Privacy Magazine*, 3(2):40–47.

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.

Seigneur, J.-M., Dimmock, N., Bryce, C. e Jensen, C. D. (2004). Combating spam with TEA. Em *Conference on Privacy, Security and Trust*.

Taveira, D. M., Mattos, D. M. F. e Duarte, O. C. M. B. (2008). Ferramenta para análise de características de spams e mecanismos anti-spam. Em *Salão de Ferramentas do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'08)*.

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.