# Capítulo 4

# Abordagens para Detecção de Spam de E-mail

Cleber K. Olivo, Altair O. Santin e Luiz Eduardo S. Oliveira

***Abstract***

*The e-mail, one of the oldest and widely used services on the Internet, is the more used tool to send an indiscriminate number of unsolicited message, known as spam. Given the wide variety of techniques used for sending spam, this type of e-mail is a problem still far from being solved. This work aims to present works and techniques relating of spam detection from a new perspective. Instead of classifying the works by type of detection technique, as is usually made in the literature, each one will be organized using the technique applied in spam dissemination as entry key. Then, it will addressed the detection techniques for each case and it will make consideration about its efficiency.*

***Resumo***

*O e-mail, um dos serviços mais antigos e mais utilizados na Internet, é o meio mais utilizado para o envio indiscriminado de mensagens não solicitadas, conhecidas como spam. Devido à grande variedade de técnicas utilizadas para o envio de spam, esse tipo de e-mail é um problema que ainda está longe de ser solucionado. Este trabalho tem como objetivo apresentar as principais técnicas e trabalhos relacionados a detecção de spam sob uma nova perspectiva. Ao invés de classificar os trabalhos pelo tipo de técnica de detecção do spam, como normalmente é feito na literatura, as abordagens serão organizadas a partir da técnica utilizada na disseminação do spam. Então, serão abordadas as técnicas de detecção para cada caso e feitas considerações acerca de sua eficiência.*

## 4.1. Introdução

### 4.1.1 Spam

O SMTP (*Simple Mail Transfer Protocol*) é o protocolo padrão utilizado para transferência de e-mails [1]. Nas últimas décadas, o e-mail tem sido um dos serviços mais utilizados na Internet, sendo o primeiro da lista quando o assunto é comunicação entre usuários na rede mundial. Por ser um dos serviços mais antigos e mais utilizados na Internet, o e-mail tornou-se o favorito para o envio de mensagens de marketing e, em alguns casos, até mesmo para o envio de mensagens fraudulentas ou contendo códigos maliciosos anexados. O envio indiscriminado de e-mails sem o consentimento de seus destinatários é conhecido como *spam.* O envio de *spam* pela Internet causa vários problemas, desde o aborrecimento dos usuários - por receberem mensagens indesejadas, até problemas de

sobrecarga dos servidores SMTP - devido ao grande volume de *spam* recebido. Em alguns casos mais graves, como o *phishing* (considerado uma subcategoria de *spam*), o dano causado pode ir muito além de um simples aborrecimento, levando o usuário a ter a segurança de seus computadores comprometida ou a ter prejuízos financeiros [2].

O *spam*, na sua forma virtual, é um problema cada vez mais presente na vida dos usuários e administradores de sistemas. O primeiro envio de *spam* por e-mail ocorreu em 1978, quando foi enviado para 393 usuários da ARPANET [3]. Desde então, as estatísticas sobre o envio de *spam* apresentam-se cada vez piores. Em 1997, a AOL estimou que de 5 a 30% de seus 10 milhões de e-mails recebidos por dia eram *spam* [4]. Em 2004, um estudo apresentou uma previsão de que esse percentual chegará em 95%, numa escala global, até 2015 [5]. Mais recentemente, estatísticas divulgadas pela Symantec revelaram que um percentual de 89,1% já foi atingido em 2010 [6], sendo o maior percentual até então [7]. Para se ter uma ideia, naquele ano a quantidade de e-mail enviado por dia foi de 62 bilhões. Nos últimos anos o volume que esses percentuais representam vem diminuindo para 60% no ano de 2013, para um volume global de spam estimado em 29 bilhões de mensagens por dia [8].

O *spamming* (envio de *spam*) vai muito além do aborrecimento do usuário. Enquanto o custo computacional para envio do *spam* é relativamente baixo, os provedores de serviços na Internet e seus usuários tem um custo alto, causado pelo desperdício de banda e pelos custos das tecnologias empregadas para a sua detecção [3].

Os *spammers* (termo utilizado para definir quem envia *spam),* a fim de dificultar a detecção de suas mensagens, fazem uso de subterfúgios técnicos que vão desde a sua forma de envio até informações inseridas propositalmente no corpo da mensagem, com o objetivo de confundir os mecanismos de detecção de spam (e.g. filtros *antispam*). O SPF (*Sender Policy Framework*), concebido para ser uma técnica *antispam*, embora se mostrasse promissor no começo, acabou sendo pouco eficiente, visto que os próprios *spammers* começaram a publicar seus registros SPF [9]. Isso mostra que protocolos de autenticação de e-mail por si só não são suficientes, tendo mais utilidade em casos de *phishing* ou e-mail *spoofing*. Um outro exemplo de técnica para burlar os mecanismos de detecção é o uso de imagens no envio das mensagens de e-mail. Isto aconteceu porque a eficiência dos mecanismos *antispam* na forma textual aumentou e então, em 2006, os *spammers* começaram a converter as mensagens em imagem [10]. De acordo com a *Ironport*, naquele mesmo ano, a quantidade de *spam* de imagem quadruplicou, representando entre 25% e 45% de todo *spam,* em alguns dias [11].

As técnicas para driblar os mecanismos *antispam* também são utilizadas para tornar ineficazes as técnicas baseadas em reconhecimento de texto. Palavras como 'viagra', por exemplo, podem se apresentar de diversas formas (e.g. V1AGR4, v.i.a.g.r.a etc). Um estudo revelou que essa mesma palavra pode ter mais de um quintilhão de variações [12], sendo praticamente impossível que os mecanismos de detecção baseados em análise de texto tenham conhecimento de todas as variações possíveis de uma única palavra. O problema atinge uma escala muito maior se for considerado que há várias palavras que podem produzir estes números significativos de variações num mesmo texto.

A criação de novas técnicas de detecção de *spam* levou os *spammers* a criarem novas técnicas de disseminação, geralmente baseadas na alteração constante do conteúdo da mensagem. De um modo geral, algumas das técnicas de detecção de *spam* existentes, envolvendo a área de reconhecimento de padrões, conseguem taxas razoáveis de classificação correta de e-mails. Entretanto, devido à grande variedade de técnicas utilizadas para o envio de *spam,* haverá detalhes específicos na mensagem que dificultarão a sua correta identificação. Assim, o *spam* é um problema que ainda está longe de ser solucionado,

visto que não existe uma técnica de detecção que seja eficiente para todas as técnicas de disseminação existentes.

Muitas ferramentas para classificação de e-mails, assim como a maioria das ferramentas de navegador, utilizam listas de remetentes "bons" (*whitelists*) e "maus" (*blacklists*). Normalmente, as *blacklists* bloqueiam o endereço IP do servidor de e-mail de origem das mensagens *spam*, ou ainda o próprio endereço de e-mail do remetente. O bloqueio do endereço IP ou domínio pode causar problemas quando o remetente utiliza o servidor de SMTP de algum provedor (e.g Yahoo, Gmail, etc), pois acaba por bloquear todos os remetentes que o utilizam. Já o bloqueio do e-mail do remetente pode ser ineficiente, visto que o mesmo pode ter sido forjado, sequestrado ou roubado de um usuário legítimo [13].

Algumas abordagens sugerem, também, a colaboração dos usuários para auxiliar o mecanismo *antispam* a classificar mensagens [14, 15]. Isto pode confundir o mecanismo de classificação de e-mails, pois um usuário pode considerar algo como *spam,* quando na verdade apenas não gosta de receber aquele tipo de e-mail por uma questão pessoal, enquanto outros usuários gostariam de recebê-lo.

As ferramentas baseadas em aprendizagem de máquina, em sua maioria acabam falhando quando um novo tipo de *spam* é recebido, visto que um novo modelo precisa ser treinado para que o novo *spam* possa ser detectado. Ou seja, até que o classificador seja treinado novamente, este novo *spam* já atingiu vários usuários. Além disto, o uso de imagens na mensagem, como comentado anteriormente, se tornou muito comum, então esta técnica passou a ser uma das mais exploradas, de acordo com a literatura [16, 17, 18, 19, 20, 21] para burlar os mecanismos *antispam*.

Uma técnica de disseminação de *spam* baseada no conteúdo da mensagem (mensagem textual), e que se tornou um problema para os classificadores baseados na ocorrência de palavras, é o ofuscamento de palavras ou caracteres. A quantidade de caracteres que podem ser trocados, visando confundir os mecanismos *antispam*, aumenta drasticamente a quantidade de palavras que podem substituir os termos originais que são comuns em mensagens de spam [12], pois para essas técnicas de classificação, se um único caractere é trocado em uma palavra, já não é computacionalmente considerado a mesma *string* de caracteres.

Este capítulo tem como objetivo apresentar as principais técnicas e trabalhos relacionados à detecção de *spam*. Porém, ao invés de classificar os trabalhos existentes pela técnica de detecção/classificação utilizada, como normalmente é feito na literatura, as propostas serão classificadas de acordo com a técnica utilizada na disseminação de *spam.* Ou seja, para cada artimanha utilizada pelos *spammers* para driblar os mecanismos de detecção, são apresentadas as soluções propostas na literatura para mitigar o problema.

Um dos problemas de apresentar as técnicas de detecção de *spam* na forma tradicional (classificando pela técnica de detecção ao invés da técnica de envio) é que, para quem não conhece o assunto a fundo, esta abordagem não revela os problemas relacionados à detecção de *spam* que devem ser resolvidos. Após a apresentação dos principais tipos de *spam*, classificando a partir das técnicas utilizadas, são apresentadas as principais técnicas de detecção propostas na literatura. Deste modo, é mais fácil entender e questionar a eficiência das técnicas de detecção para cada tipo de abordagem utilizada na disseminação das mensagens de *spam*.

### 4.1.2 Organização

A seção 4.2 apresenta alguns conceitos básicos sobre o funcionamento do serviço de e-mail. O objetivo dessa seção é entender como funciona o envio do e-mail, alguns aspectos importantes em relação ao protocolo SMTP (*Simple Mail Transfer Protocol*) e os principais campos da mensagem utilizados na comunicação entre servidores, aplicativos e usuários.

A seção 4.3 apresenta as principais técnicas utilizadas para o envio de *spam*. Diferente da maioria dos trabalhos, os quais classificam as abordagens a partir das técnicas de detecção propostas [4, 22, 23, 24, 25, 26], esta proposta apresenta uma visão diferenciada, que trata das técnicas de detecção sob o ponto de vista de cada técnica de envio de *spam*. Ou seja, a partir das principais técnicas de envio de *spam* são apresentadas as técnicas de detecção mais indicadas. Essa visão facilita o entendimento do problema, identificando a técnica de disseminação de *spam* e a abordagem de classificação de e-mails mais apropriada. Desse modo, se um administrador de sistemas estiver com problemas no recebimento de *spam* que utilize uma técnica específica (e.g *spam* de imagem), poderá conhecer as principais técnicas para a detecção deste tipo específico de *spam*.

Após a apresentação dos principais tipos de *spam*, ao apresentar cada técnica de detecção ficará mais fácil questionar a sua eficiência para um ou outro tipo de mensagem. Havendo o entendimento dessas técnicas de disseminação, é possível seguir para a próxima etapa (seção 4.4), onde são discutidas as principais técnicas de detecção de *spam* encontradas na literatura.

Na seção 4.5 é apresentada uma relação entre os tipos de *spam* apresentados na seção 4.3 e as técnicas de detecção apresentadas seção 4.4. A partir deste ponto, será possível inferir que nenhuma das técnicas isoladamente é capaz de conseguir ótimos resultados, sendo necessária uma combinação das mesmas para obter um mecanismo (e.g. filtro ou classificador) robusto e eficiente. Para cada tipo de *spam*, será apresentada a técnica mais recomendada. Adicionalmente, serão apresentadas as principais limitações de cada técnica e algumas alternativas utilizadas.

A última seção (4.6) apresenta as considerações finais acerca do assunto, possibilitando uma visibilidade maior sobre a complexidade do problema que o *spam* representa, incluindo os esforços mais recentes em pesquisa que buscam mitigar suas causas.

## 4.2. O Serviço de E-mail

O envio e recebimento de e-mails envolve dois componentes principais: o MTA (*Mail Transfer Agent*) e o MUA (*Mail User Agent*). O MTA (e.g. Postfix [27], qmail [28], Exchange [29], etc.) é o servidor responsável pelo envio e recebimento dos e-mails. Em um envio de e-mail através da Internet, na origem da mensagem, o MUA (e.g. Thunderbird [30], Microsoft Outlook [31], etc.) tem a função de coletar o e-mail do remetente e encaminhar a mensagem, através do protocolo SMTP, a um MTA de origem, que é o servidor encarregado de encaminhar a mensagem ao MTA de destino. O MTA de destino é encarregado pela entrega do e-mail ao MUA do destinatário através de algum serviço de entrega (e.g. IMAP [32] ou POP3 [33]). A Figura 4.1 ilustra um cenário onde há troca de mensagens entre dois MTAs.

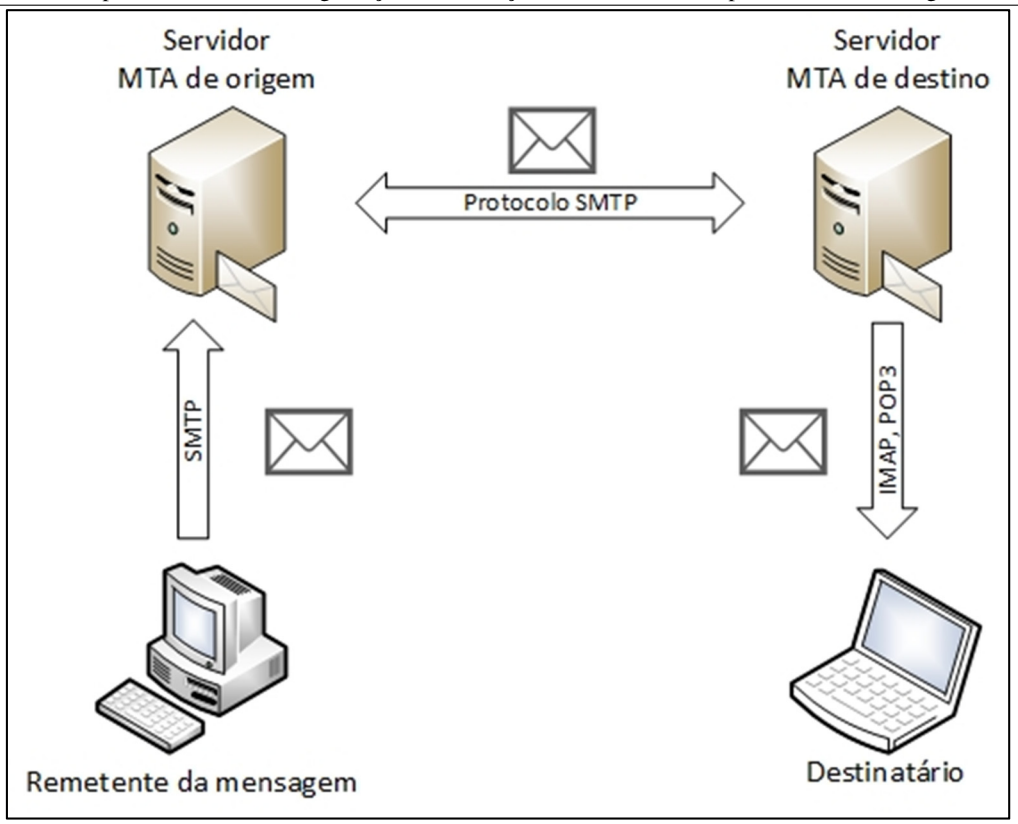



**Figura 4.1. Troca de mensagens entre dois MTAs.**

As regras para a troca de mensagens entre MTAs estão definidas na RFC 2821 (*Simple Mail Transfer Protocol*), que especifica que o e-mail é composto por um envelope (em inglês, *envelope)* e por um conteúdo *(content)* [1]. O envelope é composto por um endereço de origem (para onde os relatórios de erro são direcionados), um ou mais endereços de destino (destinatários) e outras informações opcionais do protocolo. O conteúdo é dividido em duas partes: cabeçalho (*header*) e o corpo *(body)*.

O cabeçalho sempre precede o corpo do e-mail, e contém informações obrigatórias, tais como os campos FROM (e-mail do remetente), TO (endereço do destinatário) e DATE (data), e outras informações opcionais, mas que geralmente são utilizadas, tais como os campos SUBJECT (assunto) e CC (do inglês, *carbon copy* – demais destinatários copiados no e-mail).

No que diz respeito ao *spam*, o próprio protocolo SMTP, por questões de projeto, possui limitações que são exploradas pelos *spammers* no envio de suas mensagens, possibilitando, por exemplo, que o endereço no campo remetente seja substituído por outro diferente do e-mail de quem está enviando a mensagem. Essas limitações são apresentadas com mais detalhes na seção 4.3.1.

O corpo contém a mensagem do e-mail, que pode ser composta por imagens e por texto com uso de recursos de hipertexto, como o HTML (*HyperText Markup Language*) - que serão interpretados pelo MUA. Embora o cabeçalho do e-mail seja essencialmente codificado no formato US-ASCII (*American Standard Code for Information Interchange*), o corpo do e-mail é estruturado conforme o formato MIME – *Multipurpose Internet Mail Extensions* [34].

A definição do MIME foi necessária pois o padrão que antecedia a RFC 2821 (RFC 822 – *Standard for the Format of ARPA Internet Text Messages*, 1982) [35], tinha a intenção de especificar apenas um formato para mensagens de texto. Mensagens em

outros formatos, tais como mensagens multimídia, que podem incluir áudio e vídeo, não foram mencionadas no padrão. Além disso, o padrão especificado pela RFC 822 é inadequado para as necessidades atuais dos usuários de e-mail, os quais utilizam idiomas que necessitam de um conjunto de caracteres mais amplo que o US-ASCII, como os caracteres de idiomas asiáticos e europeus [34]. O MIME redefine o formato das mensagens para permitir:

(1) Texto do corpo da mensagem utilizando conjuntos de caracteres diferentes do US-ASCII;

(2) Um conjunto extensível de diferentes formatos para corpos de texto em formato não-textual;

(3) Corpos de mensagem *multi-part*, ou seja, divididos em duas ou mais partes, cada uma com conjuntos de caracteres diferentes;

(4) Informação textual do cabeçalho em um conjunto de caracteres diferente do US-ASCII.

A possibilidade de uso de diversos conjuntos de caracteres, somada ao uso de imagens e recursos de hipertexto (HTML), permitiu que o corpo do e-mail se tornasse o campo da mensagem onde há o maior número de subterfúgios técnicos que podem ser utilizados para burlar os mecanismos a*ntispam*.

A seção 4.3 apresenta em detalhes as principais formas como o protocolo SMTP, o corpo do e-mail e o cabeçalho são utilizados na disseminação de *spam*, buscando a máxima eficiência, seja com o objetivo de alcançar o maior número de destinatários no menor tempo possível ou, ainda, driblando os mecanismos *antispam*.

## 4.3. Principais Tipos de Spam com Base na Técnica de Disseminação Utilizada

As técnicas utilizadas para a disseminação de *spam* sofreram mudanças significativas nos últimos anos. A criação de técnicas mais eficientes para a detecção de *spam* obrigou os *spammers* a criarem novas artimanhas para driblar os mecanismos de detecção. As técnicas de *spam* podem ser classificadas em duas categorias principais: i) técnicas baseadas no envio do *spam* e ii) técnicas baseadas no conteúdo do e-mail.

As técnicas baseadas no envio do *spam* correspondem à forma como o e-mail é enviado (e.g. através de uma *botnet* ou MTA legítimo). Já as técnicas baseadas no conteúdo normalmente estão associadas às artimanhas utilizadas no corpo do e-mail, para confundir os mecanismos de detecção baseados no conteúdo (textual ou não) da mensagem. Normalmente, o *spammer* utiliza uma combinação de técnicas presentes nas duas categorias (ou ainda várias técnicas da mesma categoria), maximizando as chances da mensagem passar pelos mecansimos *antispam* sem ser detectada. A possibilidade de inúmeras combinações dessas técnicas aumenta substancialmente o desafio de quem faz a classificação das mensagens (*spam* ou *não-spam*), i.e., os administradores de servidores de e-mail, que configuram os mecanismos *antispam* ou os pesquisadores que desenvolvem novas técnicas de classificação de mensagens.

A seguir será mostrado que cada uma dessas categorias principais pode abrigar várias subcategorias, como será apresentado nas próximas seções (4.3.1 – 4.3.3).

### 4.3.1. Técnicas Baseadas no Envio de Spam

As técnicas baseadas no envio de spam estão relacionadas com a forma como as mensagens são encaminhadas para os endereços de e-mail dos usuários. Ou seja, essas técnicas estão mais associadas com o "disparo" dos e-mails do que com o conteúdo da mensagem. Por exemplo, um *spam* pode ser encaminhado com ou sem a opção de verificação do recebimento da mensagem e fazer o reenvio em alguns casos de falha na entrega. Um *spam* poderia, também, ser encaminhado por um MTA que pertence a uma organização confiável - para fins publicitários, ou por uma origem fraudulenta, através de um mecanismo simples de envio – e.g. através de um *malware* (código malicioso desenvolvido para fins ilícitos e sem o consentimento do usuário).

Mais exemplos de técnicas baseadas no envio do *spam* e suas características são apresentadas nos itens de *a* até *e*.

#### a) Mecanismo simples de envio

Uma das formas típicas de enviar *spam* é através de um *mecanismo simples de envio* massivo de e-mails. Geralmente, este tipo de técnica é utilizado quando o objetivo é atingir o máximo de destinatários no menor tempo possível, sendo feito o envio de milhares de e-mails (ou mesmo dezenas ou centenas de milhares) sem se preocupar com erros de envio e confirmações de entrega das mensagens.

Em um MTA devidamente configurado, normalmente é verificado o recebimento da mensagem pelo MTA de destino. Além de problemas de rede, pode ocorrer, inclusive, verificações relacionadas à conta de e-mail do destinatário, retornando uma mensagem de erro caso a conta de usuário não exista ou esteja com a cota em disco esgotada. Em alguns casos de erro na entrega, a mensagem pode ser recolocada em uma fila para uma posterior tentativa de reenvio do e-mail. Esses controles e verificações que, entre vários propósitos, têm por objetivo a redução de erros de comunicação e maior eficiência do serviço, aumentam o custo computacional, o tempo de envio dos e-mails, e a complexidade dos softwares envolvidos no processo de envio massivo de mensagens.

Com a ausência de qualquer tipo de controle ou verificação na transmissão da mensagem (Figura 4.2), o mecanismo de envio massivo de e-mails torna-se muito mais simples, reduzindo consideravelmente o custo computacional, a complexidade do software e o tempo necessários para o envio do *spam*. O envio através de um *mecanismo simples* pode ser realizado através de um software com poucas linhas de código (inclusive algum *malware*), não sendo necessário a configuração de um servidor MTA completo. Neste caso não é feita uma nova tentativa de envio, ou seja, o servidor de origem (*spammer*) não precisa tratar os e-mails que retornam com erro, simplificando o sistema de envio do atacante.

A Figura 4.2 ilustra um exemplo no qual o *spammer* utiliza um computador pessoal para enviar *spam* para uma conta de e-mail que não existe naquele domínio (evento 1). O servidor SMTP ao receber a mensagem e constatar que não existe uma caixa de entrada para aquele endereço (usuario_inexistente@servidor_smtp.com), fornece uma mensagem de erro para o endereço do *spammer* (evento 2). A RFC 3463 mostra os códigos de estado para o serviço de e-mail [36]. O computador do *spammer*, por estar configurado como um mecanismo simples de envio de e-mails, não está preparado para tratar os e-mails de retorno, descartando a mensagem de erro (evento 3).

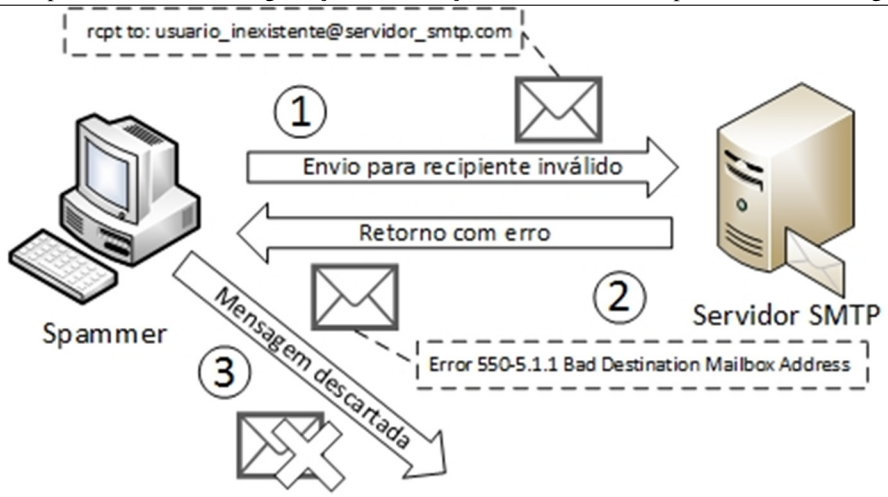



**Figura 4.2. Envio através de mecanismo simples, sem tratamento de erros.**

Em um *gateway* SMTP (ambiente de entrada e saída de e-mails onde pode haver um ou mais MTAs configurados) com grande volume de mensagens, é possível notar uma grande quantidade de e-mails destinados a endereços inexistentes (contas de usuário desativadas ou que nunca existiram). Isto ocorre porque os *spammers* muitas vezes se utilizam de listas de e-mails que possuem uma grande quantidade desses endereços inexistentes. Como não há tratamento desses e-mails com destinatários inválidos, o único tempo desperdiçado pelo *spammer*, nesses casos, é o do próprio envio da mensagem.

**b) Envio com tratamento de erros**

Neste caso, diferente do *mecanismo simples de envio*, os e-mails que retornam com erro são recolocados na fila de envio do MTA e são programados para serem retransmitidos mais tarde (Figura 4.3). Por exemplo, se ocorrer um erro temporário de rede ou no MTA de destino, o *spam* retorna ao MTA de origem e é reenviado mais tarde. O número de tentativas de reenvio e o tempo de espera para a retransmissão da mensagem é variável, podendo ser configurado no MTA de origem. Este tipo de envio se tornou mais comum à medida que foram criadas técnicas para a detecção de mensagens de *spam* enviadas através de *mecanismos simples de envio*, sendo também muito comum em servidores de e-mail que enviam *spam* com fins publicitários.

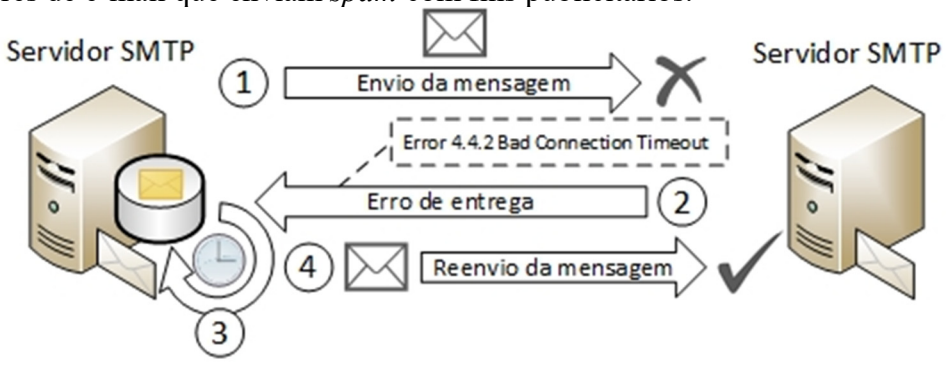



**Figura 4.3. Tratamento de erros entre MTAs completos.**

Na Figura 4.3, o servidor SMTP da origem da mensagem encaminha o e-mail para o servidor SMTP de destino e, por alguma razão (e.g. problemas de rede ou no servidor de destino) não consegue concluir a transferência do e-mail (evento 1). Ao constatar o erro na entrega do e-mail (evento 2), a mensagem é colocada em uma fila (evento 3) para posterior tentativa de reenvio. O tempo que a mensagem permanece na fila e o seu

descarte após várias tentativas é uma questão de configuração do MTA de origem. Na ilustração, a mensagem é entregue com sucesso após uma nova tentativa (evento 4).

Esta técnica de envio permite que o *spammer* valide e atualize os endereços de suas listas de e-mail, podendo remover aqueles e-mails que retornaram erro, evitando futuros envios sem sucesso.

Esta forma de envio também é adotada pela maioria das entidades legítimas em campanhas de marketing (e.g. sites de *e-commerce*), podendo ser de interesse parcial dos destinatários que recebem seus e-mails, dificultando ainda mais a tarefa de classificação da mensagem, uma vez que o que é *spam* para alguns usuários pode não ser considerado da mesma forma para outros. Detalhes sobre o conteúdo das mensagens de *e-mail marketing* são apresentados na seção 4.3.2.

**c) Envio com substituição de remetente ou transmissor**

O próprio protocolo SMTP por uma questão de projeto [1], permite que o endereço do remetente seja substituído. Assim, num ataque, além de forjar o remetente de um e-mail, também é comum que o atacante forje o endereço IP (*Internet Protocol*) do remetente (*IP Spoofing*). Essas técnicas são utilizadas para violar os mecanismos *antispam* baseados no endereço de e-mail ou IP do remetente. A Figura 4.4 ilustra este exemplo.

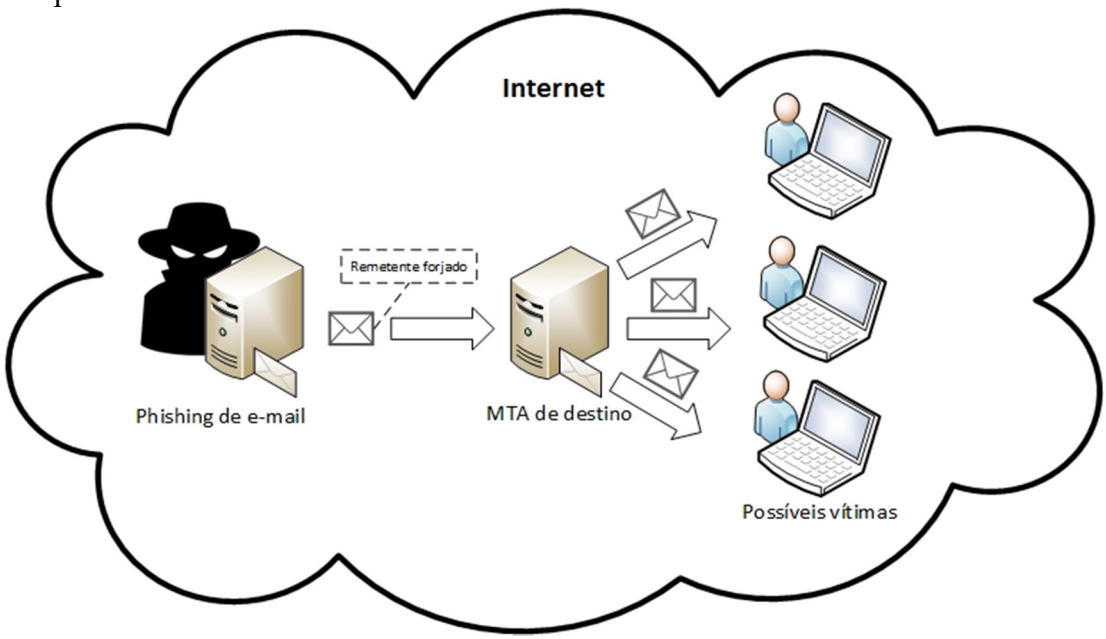



**Figura 4.4. Envio com forja do remetente.**

Além do intuito de burlar os mecanismos baseados no e-mail ou IP do remetente para envio de *spam*, em geral, esta prática é muito comum em casos de *phishing*[1] de e-mail. Em outras palavras, o *phishing* passa uma mensagem à vítima tentando convencê-la de que o remetente é uma fonte confiável (e.g. banco, site de *e-commerce*, órgãos governamentais etc.). Além disto, o *phishing* executa alguma ação que normalmente lhe causará algum tipo de prejuízo (e.g. cartão de crédito clonado ou senha bancária roubada)

[1] Phishing é uma forma de estelionato que usa engenharia social para fazer vítimas, enganando-as com o uso de recursos tecnológicos, normalmente com o objetivo de obter suas informações pessoais (geralmente de cunho financeiro) e causar-lhes prejuízos [2]. De acordo com o Código Penal Brasileiro, estelionato é "obter, para si ou para outrem, vantagem ilícita, em prejuízo alheio, induzindo ou mantendo alguém em erro, mediante artifício, ardil, ou qualquer outro meio fraudulento" [37].

ao usuário. Com o remetente forjado, ao abrir o e-mail, o usuário vítima do *phishing* acaba sendo convencido de que o remetente é de uma fonte confiável, sem perceber o golpe. Além de forjar o remetente, o atacante costuma usar um conteúdo da mensagem bem convincente, porém esta questão será explorada na seção 4.3.2.

**d) Envio através de botnets**

Uma *botnet* é uma rede de computadores comprometidos (*bots*), conectados à Internet e controlados por um atacante remoto (*botmaster*) [38]. As *botnets* utilizadas para a disseminação de *spam* geralmente são compostas por computadores comprometidos de usuários, onde foi instalado algum tipo de *malware* (código malicioso). Esses computadores são controlados remotamente, sem o consentimento do usuário, para a disseminação de *spam* (Figura 4.5).

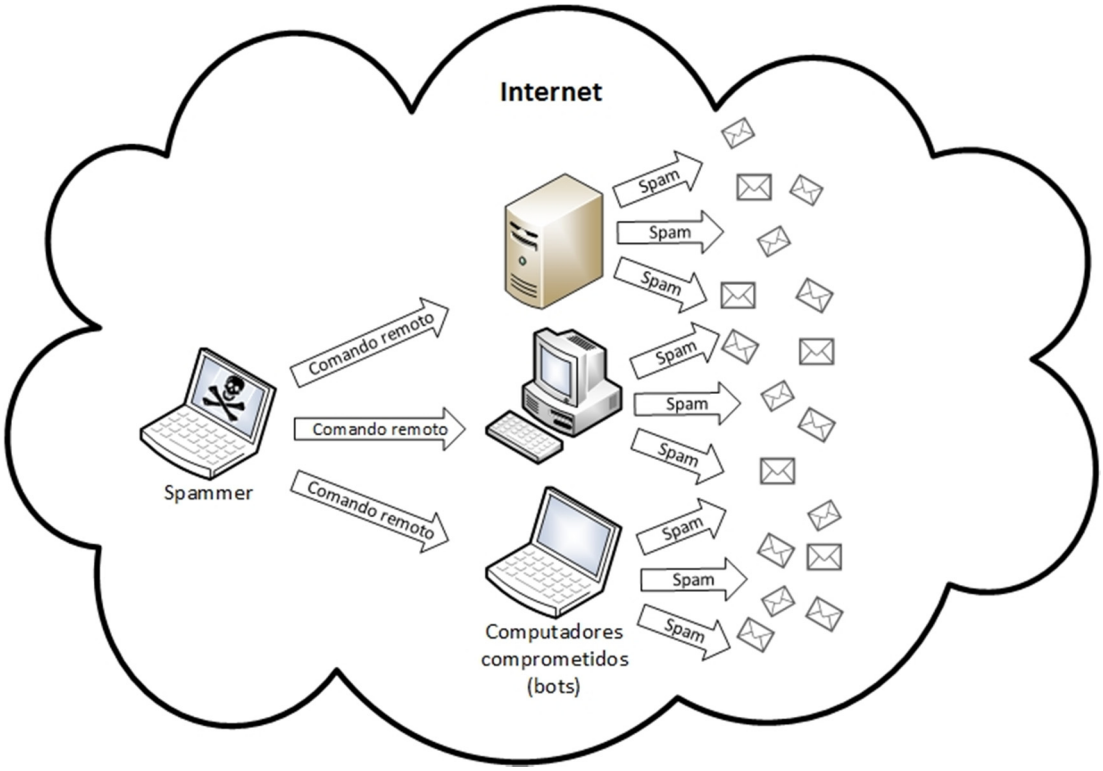



**Figura 4.5. Envio através de botnets.**

Como uma *botnet* tem grande escalabilidade, a dificuldade em detectar o *spam* a partir do endereço de origem é significativamente aumentada. Esta prática também facilita o anonimato do *spammer*, pois utiliza computadores de pessoas comuns, que podem estar localizados em pontos geográficos distantes, espalhados por diferentes países ou continentes, sem precisar configurar temporariamente servidores SMTP que denunciariam a sua localização. Assim, a quantidade de endereços IP de origem (computadores comprometidos) e a variedade de faixas de endereços IP (devido às diferentes localizações geográficas) são tão grandes que não há como realizar um bloqueio eficaz através dos endereços de origem.

O uso de *botnets* facilita anonimato e, portanto, é uma das técnicas mais utilizadas para a disseminação de *phishing* de e-mail, além de outras ameaças existentes na Internet. Como o *malware* presente nos computadores dos usuários costuma possuir poucas linhas de código, o *spam* enviado através de *botnets* geralmente se enquadra também na categoria de mecanismo simples de envio (seção 4.3.1.a).

**e) Envio através de *open relays***

*Open relays* ou relays abertas são servidores SMTP onde qualquer pessoa ou sistema pode se conectar e enviar e-mails livremente através deste, sem precisar de qualquer tipo de autenticação. A conexão é feita, quase na totalidade dos casos, sem o consentimento ou conhecimento da organização responsável pelo servidor.

Durante o final da década de 90 até o início dos anos 2000, o envio de mensagens pelos *spammers* através de servidores *open relay* era uma prática muito comum. Na época, os desenvolvedores de MTAs realizaram mudanças nos códigos e na configuração padrão dos sistemas para assegurar que as instalações padrão fossem *closed relays* (relays fechadas) e tornar a criação de uma *open relay* mais difícil, de modo a permitir que os e-mails fossem enviados através do servidor somente por usuários autorizados [39]. Entre 2012 e 2013, o projeto Spamhaus registrou cerca de 4 mil registros de *open relays*, sendo que diariamente os *spammers* descobrem e exploram de 10 a 20 novas *relays abertas* [39]. A Figura 4.6 ilustra o envio de *spam* através de *open relays*.

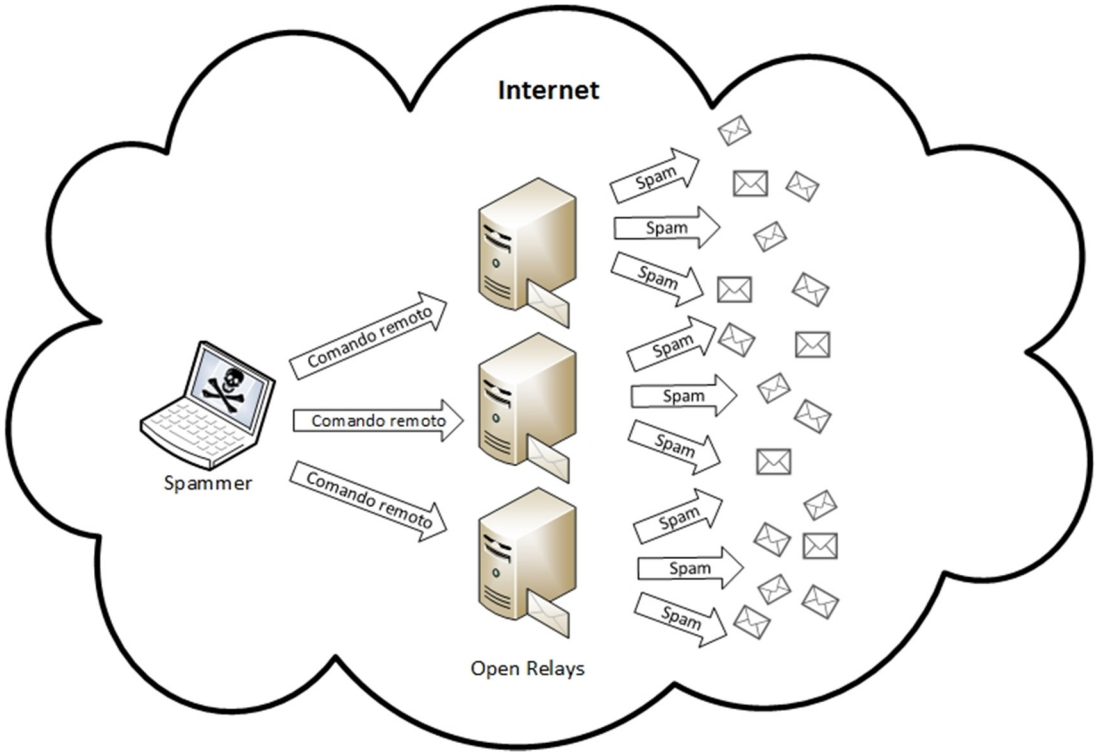



**Figura 4.6. Envio de spam através de open relays.**

O bloqueio de servidores SMTP *open relay* é complicado devido à grande quantidade desses servidores vulneráveis espalhados pela Internet e pelo surgimento constante de novas *relays* abertas a cada dia. Além disso, o bloqueio do endereço IP de origem, neste caso, também pode acarretar o bloqueio indevido dos e-mails de uma organização legítima.

### 4.3.2. Técnicas Baseadas no Conteúdo E-mail

Diferentemente das técnicas apresentadas na seção 4.3.1, as técnicas de envio de *spam* baseadas no conteúdo do e-mail estão, na maioria dos casos, associadas a violação de mecanismos *antispam* que se baseiam nas informações coletadas a partir do corpo da mensagem. Essas técnicas vão desde a inserção proposital de palavras, que confundem o

classificador de e-mails, até o uso de subterfúgios técnicos, como recursos da linguagem de hipertexto (HTML) incorporada ao e-mail.

A identificação dessas técnicas dará sequência às que foram apresentadas na seção 4.3.1, porém com foco exclusivo no corpo da mensagem. As principais técnicas são apresentadas nos itens entre a e f.

**a) Inserção proposital de palavras**

A *inserção proposital de palavras* é utilizada pelos *spammers* para confundir os classificadores de *spam*, que usam a ocorrência de palavras mais comuns para a detecção do ataque. É o caso dos classificadores bayesianos [40], que classificam as mensagens com base nas palavras que ocorrem com mais frequência, tanto para *spam* como para *não-spam*. Dependendo das palavras existentes no corpo da mensagem, o classificador gera uma probabilidade do e-mail ser *spam* ou não.

Para confundir o classificador, o *spammer* insere de forma proposital, palavras (texto comum) em e-mails legítimos no conteúdo do *spam*, fazendo com que a técnica de classificação utilizada caracterize a mensagem como *não-spam*, porque há predominância de texto legítimo no e-mail.

**b) Troca ou inserção proposital de caracteres**

A *troca ou inserção proposital de caracteres* (também conhecida por *ofuscação textual*) é outra técnica utilizada pelos *spammers* para violar os mecanismos de detecção baseados no texto da mensagem. Diferentemente da *inserção proposital de palavras*, em vez de inserir palavras que possam confundir o classificador, é realizada a troca, exclusão ou inserção proposital de caracteres nas palavras mais comuns. Este tipo de técnica prejudica a detecção tanto nos mecanismos baseados na probabilidade de mensagens como em filtros baseados em regras de detecção com palavra-chave (seção 4.3).

Por exemplo, a palavra 'viagra', muito comum em mensagens de *spam*, com o uso desta técnica poderia se apresentar de diversas formas (e.g. V1AGR4, v.i.a.g.r.a etc.), chegando a mais de *seiscentos quintilhões de variações* (*600.426.974.379.824.381.952*) [12]. Este número torna a detecção dessa e demais palavras e suas variações praticamente impossível porque esta quantidade fantástica de combinações deveria ser gerada em tempo real para todas as palavras ou armazenada em alguma base de dados. Em ambos os casos, os mecanismos de detecção baseados na ocorrência de palavras precisam comparar, no e-mail recebido, cada palavra do corpo da mensagem com todas as combinações possíveis (*strings* ofuscadas) de todas as palavras possíveis. O número de combinações é possível devido ao grande número de caracteres existentes no padrão *Unicode* que, em sua versão 8.0, reúne 120.672 códigos que representam caracteres de vários idiomas, ideogramas e coleções de símbolos [41].

No caso da substituição de caracteres, o problema pode aumentar mais ainda caso haja substituições de um caractere da palavra original por dois ou mais caracteres. A Tabela 4.1 mostra alguns exemplos.

A Tabela 4.1 apresenta alguns exemplos de ofuscação textual que podem ocorrer. A forma mais simples é a inserção de caracteres no meio da palavra como espaços, pontos, hifens etc. Os caracteres inseridos podem variar na mesma palavra (e.g. 'V I.A-G.R A'). Além da simples inserção de caracteres no meio de uma palavra específica, também pode ocorrer uma substituição $N \rightarrow 1$, ou seja, onde um ou mais caracteres ($N$) são utilizados para representar um caractere específico. É possível ainda que uma única palavra possua uma combinação de mais de um dos tipos de ofuscação apresentados na Tabela 4.1 (e.g.

'P|-|AR.M/\CE.UTI.C@L'). Com a grande possibilidade de inserções, substituições e combinações de caracteres para realizar a ofuscação textual, fica claro o tamanho da complexidade do problema.

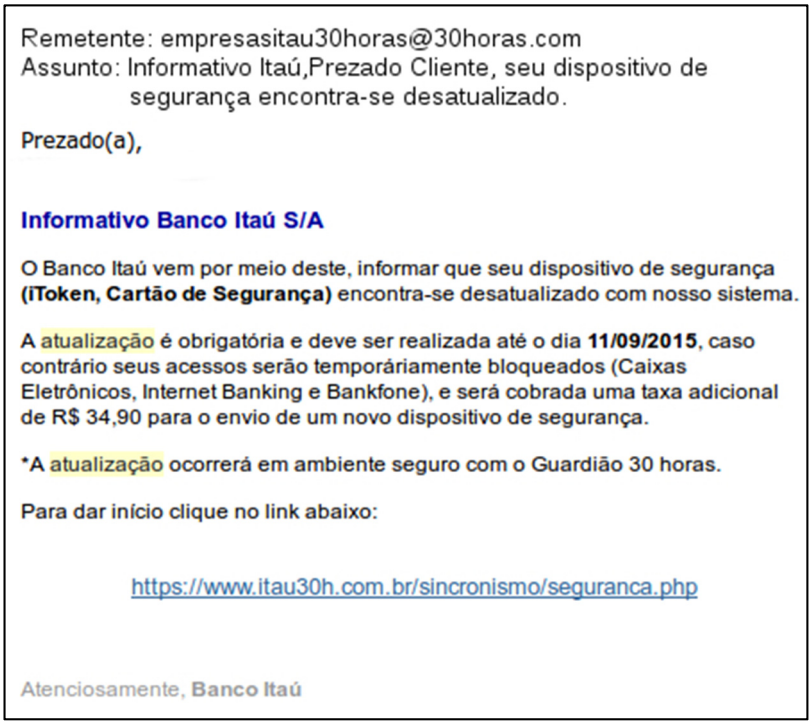
**Tabela 4.1. Exemplos de ofuscação textual.**

| Tipo de Ofuscação | Exemplos |
|---|---|
| Inserção de caracteres | 'P h a r m a c e u t i c a l', 'V.I.A.G.R.A', 'T-i-c-k-e-t', 'V I.A-G.R A' |
| Substituição 1⟶1 | 'P**h**@rmaceutic**a**1', 'V1**A**GR**4**', 'T1C**K**ET' |
| Substituição 2⟶1 | 'P**H**/\RMACEUTIC/\L', 'VI/\GR/\', 'TIC**I<**ET' |
| Substituição 3⟶1 | 'P\|-\|ARMACEUTIC/-\L', 'VI/-\GR/-\' |

Liu e Stamm [42] realizaram experimentos para comprovar o quanto a ofuscação de palavras pode prejudicar o resultado de um classificador. Para a realização dos testes, foram substituídos termos originais (sem ofuscação) em mensagens da base de *spam* por caracteres *Unicode* que possuem semelhança visual com o caractere original. Depois de treinar a ferramenta SpamAssassin [43], foram realizados testes de classificação nas bases originais (sem ofuscamento), com ofuscamento e na base desofuscada. Os resultados obtidos demonstram que termos ofuscados têm grande impacto no resultado do classificador. O Spam Assassin atribui uma nota (*score*) para o e-mail que, quanto mais alta, maior é a probabilidade de ser *spam*. Em um dos experimentos, as mensagens de *spam* originais receberam notas de 7,9 a 21,7. Com os termos ofuscados, as notas foram de 1,9 a 5,94, ou seja, muito abaixo do que um *spam* normalmente receberia.

### c) Conteúdo falso

Uma prática comum utilizada pelos *phishers* (*spammers* que disseminam *phishing*) é utilizar um texto bem convincente e muito parecido com mensagens legítimas. Mas, que na verdade ludibriam o usuário do sistema de e-mail, convencendo-o a realizar alguma ação que poderá torná-lo vítima de algum tipo de fraude.



Remetente: empresasitau30horas@30horas.com
Assunto: Informativo Itaú,Prezado Cliente, seu dispositivo de segurança encontra-se desatualizado.

Prezado(a),

**Informativo Banco Itaú S/A**

O Banco Itaú vem por meio deste, informar que seu dispositivo de segurança **(iToken, Cartão de Segurança)** encontra-se desatualizado com nosso sistema.

A atualização é obrigatória e deve ser realizada até o dia **11/09/2015**, caso contrário seus acessos serão temporáriamente bloqueados (Caixas Eletrônicos, Internet Banking e Bankfone), e será cobrada uma taxa adicional de R$ 34,90 para o envio de um novo dispositivo de segurança.

*A atualização ocorrerá em ambiente seguro com o Guardião 30 horas.

Para dar início clique no link abaixo:

https://www.itau30h.com.br/sincronismo/seguranca.php

Atenciosamente, **Banco Itaú**

**Figura 4.7. Mensagem de um *phishing* de e-mail.**

A Figura 4.7 é um exemplo de conteúdo de uma mensagem de um phishing. O texto tenta convencer a vítima que o seu dispositivo de segurança utilizado para acessar serviços bancários está desatualizado (o texto possui, inclusive, a palavra 'atualização' destacada em amarelo para chamar a atenção do usuário). Para atualizá-lo o usuário deve acessar o link informado, que o levará a baixar um malware que depois poderá lhe causar prejuízos financeiros, por exemplo, devido ao roubo de senhas bancárias.

Este tipo de e-mail acaba fugindo à regra pois, além de forjar o remetente (seção 4.3.1.c), possui características textuais muito parecidas com e-mails legítimos, tornando a mensagem ainda mais convincente aos olhos do usuário (vítima). Esses e-mails também podem conter outras artimanhas utilizadas pelos phishers, o uso de recursos da linguagem de hipertexto HTML para enganar as vítimas (assunto que será apresentado na seção 4.3.2.e). Por se parecerem com e-mails que estão presentes em bases de não-spam, os classificadores baseados no texto da mensagem muitas vezes não são eficientes contra este tipo de spam.

**d) Uso de imagens**

À medida que os mecanismos de detecção de *spam* textuais foram se tornando mais eficientes, os *spammers* passaram a inserir o texto da mensagem dentro de imagens, tornando os classificadores tradicionais ineficazes para este tipo de *spam* [10]. Uma das soluções para isto foi o uso de OCR (*Optical Character Recognition*) – Reconhecimento Ótico de Caracteres, que realiza a extração do texto contido nessas imagens (pré-processamento) e submetendo-o em seguida para o processamento textual. Após a adoção de técnicas de OCR como solução do problema, os *spammers* começaram a utilizar técnicas para dificultar o processamento das imagens.


**Figura 4.8. Imagem com técnica de ofuscação. Adaptado de [20].**

As técnicas utilizadas pelos spammers para dificultar o tratamento dessas imagens (e.g. ofuscação de imagens, Figura 4.8) aumentam consideravelmente a complexidade da classificação dos e-mails. Pela sua eficiência, esta técnica se tornou o tipo de spam mais explorado recentemente. O ofuscamento de imagens consiste em utilizar uma imagem de fundo que dificulte, de alguma maneira, o reconhecimento e a extração do texto durante o pré-processamento, mas sem prejudicar a mensagem do texto para a visão humana.

Esta prática passou a ser comum a partir de 2006 [10], sendo que no mesmo ano a quantidade de e-mail com este tipo de spam quadruplicou, representando de 25 a 45% do total, dependendo do dia [11]. Outro agravante é o tamanho médio do spam de imagem,

que é em média cerca de 10 vezes maior do que o spam textual, consumindo maior banda para trafegar na rede e mais recursos de armazenamento [16], além de necessitar de tempo e capacidade adicional de processamento no MTA de destino.

A Figura 4.8 é um exemplo de spam de imagem com técnica de ofuscação. O fundo "poluído" e a variação de cores tanto no fundo como nas letras da imagem dificultam consideravelmente a aplicação do OCR para a extração textual. Para visão humana, entretanto, é possível compreender a mensagem com um mínimo de dificuldade.

**e) Uso de recursos da linguagem HTML**

O uso de recursos de hipertexto (HTML) de forma mal-intencionada é muito comum em casos de estelionato, como no *phishing* de e-mail. Nesse caso, os recursos da linguagem podem ser utilizados para ocultar informações da vítima ou até mesmo confundi-la.

Um dos exemplos mais comuns do uso de recursos HTML é quando o texto âncora, ou seja, o texto visível para o usuário é uma URL (*Uniform Resource Locator*) de algum site legítimo, mas que aponta para um domínio fraudulento diferente do endereço visível para o usuário, por exemplo:

*<a href="http://playpal.com"> http://www.paypal.com/login.php </a>*

Neste exemplo, o usuário verá a URL "http://www.**pay**pal.com/login.php", porém será redirecionado para o endereço "http://**play**pal.com". Um usuário mais experiente saberá que um simples passar do ponteiro do mouse sobre o link provavelmente mostrará a verdadeira URL por trás do texto âncora, porém, esta técnica costuma funcionar com os usuários mais desatentos ou inexperientes.

Um outro exemplo poderia ser:

*<img src=http://www.dpf.gov.br/logo.png> Você está intimado a comparecer em nossa delegacia! <a href="http://*badsite.com/malware.exe*"> Clique aqui para saber o motivo </a>.*

Neste caso, a vítima enxergaria a mensagem "*Você está intimado a comparecer em nossa delegacia! Clique aqui para saber o motivo"*, com uma imagem do logo da organização, retirado diretamente do Portal da Polícia Federal e contendo um link para o endereço "http://badsite.com/malware.exe", que aponta para um arquivo executável que provavelmente contém algum tipo de *malware*.

No trabalho de Olivo, C. K., Santin, A. O. e Oliveira, L. S. [2] há exemplos de vários outros casos de uso do HTML que são comuns em *phishing*.

**f) Campanha Publicitária (*E-mail marketing*)**

Campanhas publicitárias (*e-mail Marketing* ou marketing por e-mail) são mensagens com fins publicitários que geralmente são enviadas por um MTA com considerável poder de processamento, que possui um domínio autêntico para envio de e-mails, tratamento de erros etc. É importante ressaltar que o *e-mail marketing* não é exatamente uma técnica de disseminação de *spam* baseada no conteúdo da mensagem, pois nem sempre o objetivo primário destes e-mails é causar problemas ao usuário ou administradores de servidores de e-mail. O problema dessas mensagens é sua classificação pelos usuários (que podem considerá-las mensagens não solicitadas – *spam*), esta é a razão pela qual esta categoria de e-mails está incluída nesta seção.

Diferentemente de e-mails que promovem a venda de medicamentos (e.g. Viagra, Cialis etc.) e outras categorias de *spam* que são indesejados por quase todos os usuários, este tipo de e-mail pode ser do interesse de alguns por conter assuntos de interesse específico, tais como: promoções de produtos, lançamentos etc. A maioria destes e-mails de campanhas publicitárias também oferece ao usuário a opção de remoção do seu endereço da *mailing list*, para que não receba mais o que considera *spam*.

### 4.3.3. Considerações sobre as técnicas de disseminação de spam

Conforme apresentado nas seções anteriores, há uma grande variedade de técnicas e tipos de *spam*. A Tabela 4.2 sumariza as principais características, objetivos e principais dificuldades de detecção para cada tipo de *spam*.

A grande diversidade de técnicas de disseminação ou tipos de *spam* aumenta consideravelmente a complexidade do problema. Sem entender a complexidade do problema, ao analisar uma técnica específica de detecção, é impossível ter a compreensão necessária da abordagem apresentada, identificando em que momento o mecanismo de detecção poderá falhar.

**Tabela 4.2. Exemplos de ofuscação textual**

| Tipo de Spam | Características | Objetivos |
|---|---|---|
| **Mecanismo simples de envio** | • Normalmente ocorre o envio de milhares, ou mesmo dezenas de centenas de milhares de e-mails.<br>• Envio sem tratamento de erros.<br>• Transmissão da mensagem sem tentativa de reenvio em caso de erro.<br>• Comum em casos de computadores comprometidos por *malwares* que disseminam *spam*.<br>• Sistemas de envio de e-mail sem robustez.<br>• Poucas linhas de código são necessárias para o mecanismo de envio.<br>• Baixo custo computacional para envio da mensagem. | • Atingir o maior número possível de destinatários no menor tempo possível, sem se preocupar com erros de transmissão. |
| **Envio com tratamento de erros** | • Mecanismos de envio mais robustos.<br>• Ocorre o tratamento de erros ou retransmissão da mensagem.<br>• Muito comum em casos de *e-mail marketing*. | • Dificultar a detecção do *spam* contra técnicas que são eficazes contra *mecanismos simples de envio*.<br>• Possibilitar a validação de endereços de e-mail, excluindo aqueles que retornam com erro, facilitando o controle do processo de *spamming*. |
| **Envio com substituição de remetente ou transmissor** | • Ocorre a falsificação do endereço IP ou do endereço de e-mail do remetente da mensagem.<br>• No caso da falsificação do e-mail, a técnica é possível devido às limitações do protocolo SMTP.<br>• Muito comum em casos de *phishing* de e-mail. | • Violar os mecanismos baseados em detecção de *spam* através do endereço de e-mail ou IP do remetente.<br>• Enganar a vítima, fazendo-a acreditar que o e-mail foi encaminhado de uma fonte confiável. |
| **Envio através de *botnets*** | • Envio de *spam* através de computadores comprometidos por *malwares*, sem o consentimento dos seus usuários. | • Ocultar a localização do *spammer*, já que o envio é feito através de computadores comprometidos de terceiros. |

| Tipo de Spam | Características | Objetivos |
|---|---|---|
| | • Grande escalabilidade da *botnet*, o que dificulta a detecção do *spam* a partir do endereço de origem.<br>• Também se enquadra na categoria *mecanismo simples de envio*. | • Atingir o maior número possível de destinatários no menor tempo possível, através do envio massivo de e-mails, devido à alta escalabilidade das *botnets*. |
| **Envio através de *open relays*** | • Servidores SMTP que permitem que qualquer pessoa ou sistema se conecte livremente, utilizando-os para a disseminação de *spam*.<br>• Envio realizado sem o consentimento da organização responsável pelo servidor SMTP configurado como *open relay*. | • Dificultar a localização do *spammer*, já que o envio é feito através de servidores SMTP que pertencem a terceiros.<br>• Dificultar o bloqueio a partir do endereço IP de origem, devido à grande quantidade de servidores com esse tipo de vulnerabilidade na Internet. |
| **Inserção proposital de palavras** | • Inserção proposital de palavras comuns em e-mails que não são *spam*. | • Enganar os classificadores baseados na ocorrência de palavras mais comuns em mensagens de *spam*. |
| **Troca ou inserção proposital de caracteres** | • Técnica também conhecida como ofuscação textual.<br>• Ocorre a troca, exclusão ou inserção proposital de caracteres em palavras comuns em *spam*.<br>• Mesmo com a troca dos caracteres em determinadas palavras, a compreensão humana não é prejudicada.<br>• Possibilidade de geração de inúmeras variações de uma única palavra.<br>• Considerável complexidade de detecção. | • Enganar os classificadores baseados na ocorrência de palavras mais comuns em mensagens de *spam*. |
| **Conteúdo falso** | • Texto bem convincente, com características textuais semelhantes a mensagens legítimas.<br>• Normalmente também é utilizado com a *forja de remetente*.<br>• Normalmente utiliza recursos hipertexto da linguagem HTML para enganar as vítimas.<br>• Classificadores baseados em características textuais da mensagem muitas vezes não são eficientes contra este tipo de *spam*. | • Ludibriar o usuário do sistema de e-mail, convencendo-o a realizar alguma ação que poderá torná-lo vítima de algum tipo de fraude. |
| **Uso de imagens** | • Texto da mensagem passa a ser "embutido" em imagens.<br>• Um dos tipos de *spam* mais explorados na literatura.<br>• Pode ocorrer o uso de técnicas de ofuscação de imagens.<br>• Tamanho médio do *spam* de imagem é dez vezes maior do que o spam textual.<br>• Consome mais recursos de armazenamento e gera maior tráfego na rede.<br>• Necessita de técnicas adicionais de OCR (pré-processamento) para extrair o texto da imagem para posterior processamento textual. | • Burlar os mecanismos *antispam* baseados em características textuais.<br>• Nos casos de ofuscamento de imagens, o objetivo é dificultar o tratamento desse tipo de *spam*. |
| **Uso de recursos da** | • Seu uso mal-intencionado é muito comum em casos de estelionato, como no *phishing* de e-mail. | • No caso de *phishing*, o objetivo é ocultar informações da vítima ou confundi-la. |

| Tipo de Spam | Características | Objetivos |
|---|---|---|
| **linguagem HTML** | | |
| **E-mail marketing** | • Normalmente é encaminhado através de MTAs robustos.<br>• Geralmente o envio de e-mails é realizado com o tratamento de erros.<br>• Não há unanimidade na classificação desses e-mails por parte dos usuários (se é *spam* ou não-*spam*).<br>• A maioria desses e-mails permite que o usuário remova seu endereço da lista de envio. | • Realizar campanhas publicitárias por e-mail, usando assuntos que podem ser do interesse de apenas alguns usuários.<br>• Nem sempre o objetivo primário desses e-mails é causar problemas ao usuário e administradores de servidores de *e-mail*. |

Com o objetivo de aumentar a compreensão sobre o problema e complexidade do *spam*, a seção 4.3 apresentou as principais técnicas de disseminação utilizadas pelos *spammers*, com o objetivo de facilitar a análise das técnicas de classificação de e-mails que serão apresentadas nas próximas seções.

## 4.4. Principais técnicas utilizadas para detecção de spam

Diante das técnicas mais utilizadas para a disseminação de *spam*, várias propostas foram criadas prometendo solucionar ou mitigar o problema. Esta seção apresentará as principais técnicas encontradas na literatura.

### 4.4.1. Técnicas de detecção de spam baseadas em regras, políticas ou protocolos

#### a) Whitelists e Blacklists

O exemplo mais comum de regra para bloqueio de *spam* é o uso de listas de bons (*whitelists)* e maus (*blacklists)* remetentes. Basicamente, a regra consiste em aceitar ou rejeitar todo e-mail, domínio ou IP contido nessas listas. No caso das *whitelists*, há a possibilidade de liberar o e-mail destinado a caixa de entrada do usuário sem precisar passar pelos demais mecanismos e classificadores, agilizando o processo de entrega e diminuindo a carga de processamento das mensagens no MTA de destino. Este tipo de configuração, entretanto, pode ser um risco para a segurança por aceitar qualquer mensagem de endereços que estejam na *whitelist*. Da mesma forma são utilizadas as *blacklists* para o bloqueio das mensagens. A Figura 4.9 ilustra o funcionamento dessas listas.

Na Figura 4.9 o e-mail que não é *spam* é enviado por um remetente que consta na *whitelist* do MTA de destino, sendo entregue diretamente ao MUA do usuário sem qualquer tipo de processamento da mensagem. Uma das vantagens do uso de *whitelists* é evitar o bloqueio indevido e reduzir o custo computacional para o processamento de e-mails de fontes confiáveis. Entretanto, se esses remetentes "confiáveis" estiverem comprometidos por algum tipo de *malware*, poderá representar um sério risco à segurança do servidor de seus usuários e redes as quais estiver conectado.

Na Figura 4.9 no exemplo em que um *spam* é encaminhado, o endereço do *spammer* está presente na *blacklist* e, portanto, a mensagem é bloqueada. O uso de *blacklists* é útil para o bloqueio de fontes de *spam* conhecidas.

O uso destas listas, apesar de parecer eficiente, possui várias limitações e o seu uso é recomendado somente em último caso. O bloqueio do endereço IP ou domínio pode

causar problemas quando o remetente utiliza o servidor de SMTP de algum provedor (e.g. Yahoo, Gmail etc.), pois acaba por bloquear todos os remetentes que o utilizam. Já o bloqueio do e-mail do remetente pode ser ineficiente, visto que o mesmo pode ter sido forjado (veja a seção 4.3.1.c).

No caso do *phishing* (seção 4.3.2.c) a origem da mensagem (e.g. endereço IP, URL alvo do *phishing*, e-mail forjado etc.) costuma mudar constantemente para evitar seu rastreamento. Além disso, a dificuldade de administração dessas listas pode se tornar muito complexa, pois o fluxo de mensagens pode ser muito intenso no servidor SMTP onde a filtragem é realizada. Assim, esta abordagem geralmente é ineficiente [2].

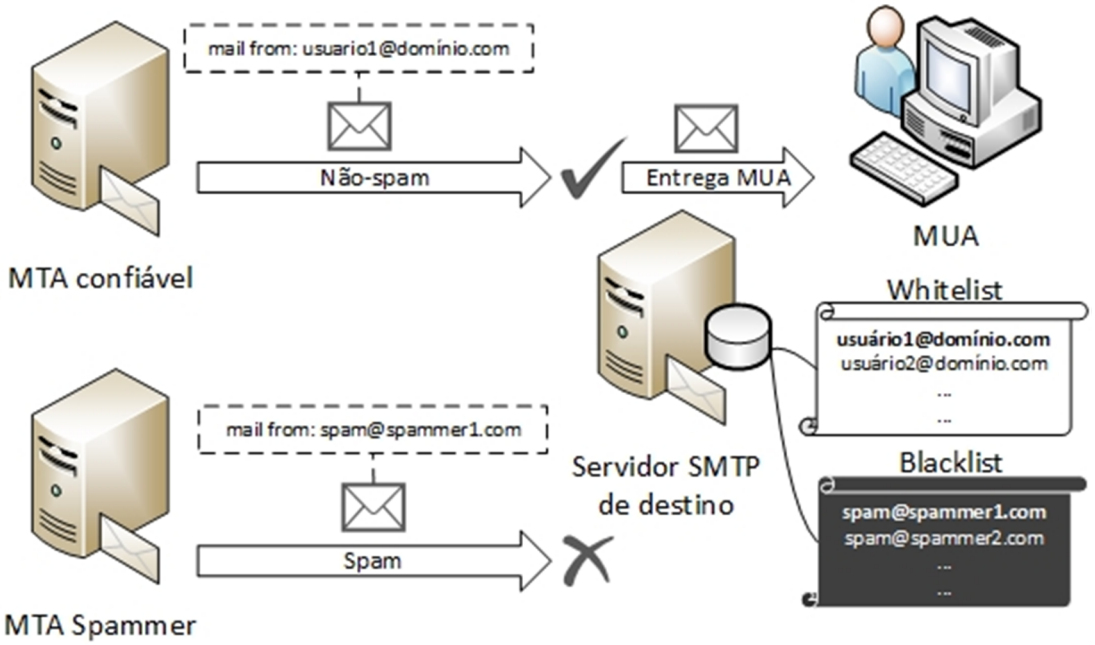



**Figura 4.9. Funcionamento de whitelists e blacklists.**

### b) Mecanismos *antispam* baseados em palavras-chave

De modo similar as *blacklists*, também é possível bloquear e-mails que contenham determinadas palavras no corpo da mensagem. As palavras inseridas na lista de palavras-chave podem fazer uso de expressões regulares para identificar algumas variações de *strings* de caracteres.

Na figura 4.10 há um exemplo de palavras-chave utilizadas no bloqueio de *spam*. De acordo com Jargas, A. M. [44], uma expressão regular é um método formal de se especificar um padrão de texto. A expressão, quando aplicada em um texto qualquer, retorna sucesso caso este texto obedeça a todas as suas condições. Na Figura 4.10, entre colchetes estão os caracteres (expressão regular) que podem ocorrer em determinada posição da *string*. Apesar de ser um exemplo simples, a complexidade das expressões regulares pode ser muito maior, sendo muito útil no momento de compor essas listas de palavras, abrangendo um número considerável de variações de uma mesma *string* de caracteres. Entretanto, sem a perícia adequada, o uso de expressões regulares pode causar o bloqueio indevido dos e-mails.

O bloqueio a partir de palavras-chave deve ser utilizado somente em último caso ou de maneira paliativa, por exemplo, bloqueando um *spam* que não está sendo detectado pela técnica de classificação textual, de forma temporária, usando assunto da mensagem ou alguma palavra ou frase no corpo do e-mail para selecionar o alvo do bloqueio. Como esse tipo de bloqueio não considera o contexto da mensagem como um todo e não calcula

nenhum tipo de probabilidade da mensagem ser ou não *spam*, a inserção de palavras na lista de bloqueio deve ser feita com bastante cautela, a fim de evitar o bloqueio indevido de mensagens.

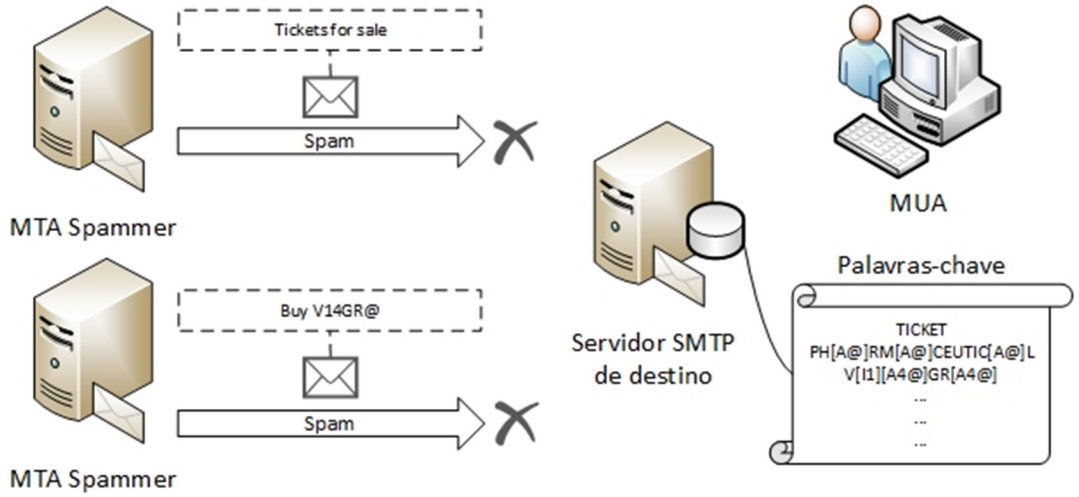



**Figura 4.10. Bloqueio por palavras-chave.**

**c) *Greylisting***

Em configurações mais rudimentares de *spam* a maioria das mensagens é enviada através de programas pouco robustos, tentando atingir o máximo de destinatários possíveis em um determinado tempo (seção 4.3.1.a). Não dispondo de um MTA completo, as mensagens de *spam* que não fossem aceitas pelo servidor de destino não eram colocadas em fila para reenvio posterior. Isso é muito comum atualmente, em computadores comprometidos em uma *botnet* para envio de *spam* (seção 4.3.1.d). Além disso, para evitar seu rastreamento, é comum o *spammer* não possuir nenhum domínio registrado em seu nome. Esse fato levou à criação de diversas outras técnicas baseadas em listas, como o *greylisting* [45] e o SPF (*Sender Policy Framework*) [46].

A *greylisting* (Figura 4.11) consiste numa recusa inicial da mensagem recebida pelo MTA de destino (evento 1), supondo que o *spammer* não dispõe de um MTA completo que reenviará a mensagem em caso de falha na entrega. O endereço do remetente é colocado temporariamente na *greylist*, onde permanecerá por um tempo determinado aguardando o reenvio da mensagem (evento 2). O servidor SMTP de destino enviará uma mensagem de erro (evento 3) informando que a mensagem foi colocada na *greylist*. Na origem, se o MTA estiver configurado devidamente para reprocessar as mensagens, o e-mail é colocado numa fila para reenvio (evento 4). Na nova tentativa de envio da mensagem, o servidor SMTP de destino fará a checagem do endereço do remetente, que desta vez estará na *greylist*, aceitando a mensagem (evento 5). O tempo que os endereços permanecem na *greylist* e o tempo que em que ocorre uma nova tentativa de envio da mensagem, correspondendo a um detalhe de configuração dos servidores.

Levine, J. R. [47] cita várias situações em o *greylisting* pode falhar, por exemplo:

(1) Perda de e-mails legítimos, caso o MTA de origem não esteja configurado para reenviar mensagens com falha na entrega;

(2) Atraso na entrega de e-mails (tempo entre o primeiro e segundo envio), o que se agrava dependendo da configuração do MTA de origem;

(3) Máquinas de envio de *spam* mais robustas podem suportar o reenvio de mensagens, o que inviabiliza o uso desta técnica (seção 4.3.1.b).

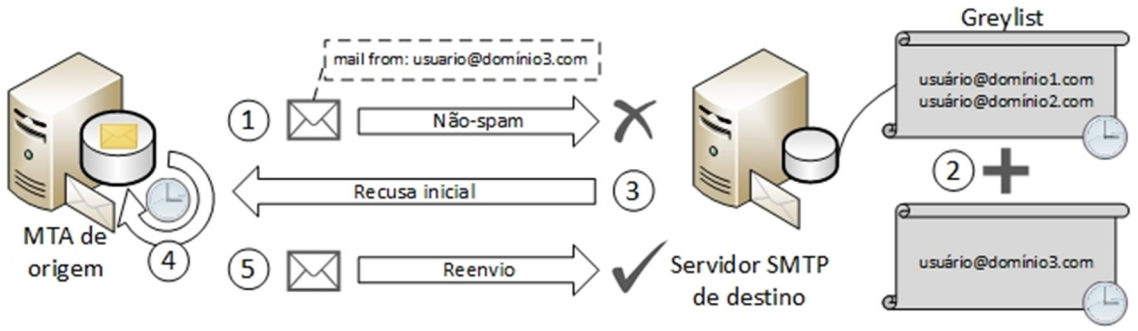



**Figura 4.11. Cenário de uso de Greylisting.**

**d) Framework com Política de Remetente (SPF - *Sender Policy Framework*)**

O SPF é um padrão aberto que especifica um procedimento para prevenir a substituição do remetente do e-mail (seção 4.3.1.c – *envio com substituição de remetente ou transmissor*). Mais especificamente, a versão atual do SPF (SPFv1 ou SPF Clássico) protege o campo *"envelope sender address"* (também conhecido como *return-path*), que é o campo utilizado pelos MTAs de origem e destino na entrega das mensagens, inclusive nos casos de erro na entrega do e-mail e retorno ao remetente [46].

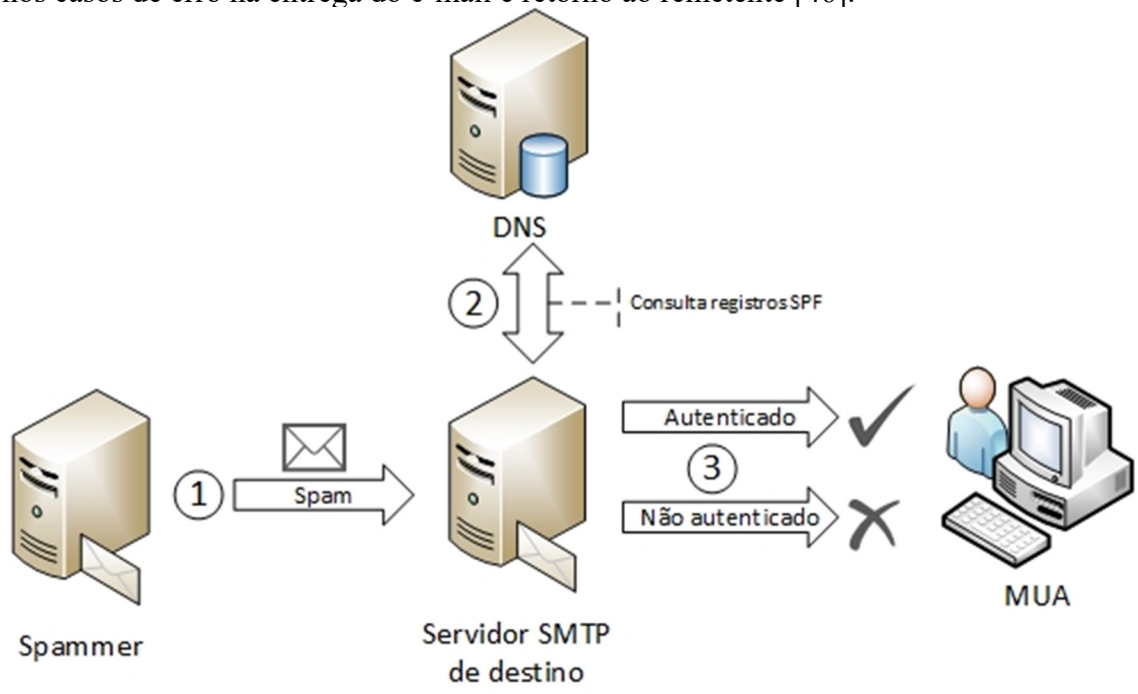



**Figura 4.12. Cenário de uso de SPF**.

O SPF funciona de modo que o proprietário de um domínio possa especificar quais servidores são utilizados para enviar e-mails. Para que o procedimento funcione, primeiramente o proprietário do domínio remetente deve publicar um registro SPF no DNS (*Domain Name System*), o qual contém o endereço de um servidor autorizado a enviar e-mail em nome do remetente. Além disso, o servidor de destino deve fazer a checagem desse registro, rejeitando a mensagem caso essa não seja originada em um endereço especificado na política SPF. Uma vez que a autenticidade do domínio do servidor do remetente é confirmada, este poderá ser adicionado a alguma lista ou sistema de reputação [46]. A Figura 4.12 ilustra o funcionamento desta técnica.

Atualmente, é sabido que boa parte dos *spammers* possui os recursos de um MTA completo (seção 4.3.1b), com a possibilidade de reenvio da mensagem em caso de falha

(ou na recusa da primeira mensagem pela *greylisting*), ou mesmo publicando o registro SPF do seu servidor de e-mail (de onde o *spam* é enviado), tornando esses mecanismos ineficazes na maioria dos casos [9, 47]. Um bom exemplo da limitação da proposta do SPF é o *"spam* comercial" enviado por empresas de *e-commerce*, onde as mensagens publicitárias, na maioria das vezes, são encaminhadas através de servidores de e-mail completos (que podem, inclusive, possuir registros SPF publicados). Além disso, nem todos os servidores de domínios confiáveis (e.g. bancos, órgãos governamentais, empresas, etc.) possuem seus registros SPF publicados, não impedindo que terceiros mal-intencionados enviem e-mails em nome desses domínios.

**e) Técnicas Baseadas em Assinatura**

Várias técnicas baseadas na assinatura das mensagens foram desenvolvidas para mitigar o problema do *spam*. Uma das mais conhecidas, o DKIM (*Domain Keys Identified Mail*), define um mecanismo pelo qual as mensagens de e-mail podem ser assinadas digitalmente, permitindo que um domínio assinante reivindique a responsabilidade pela introdução da mensagem no fluxo de e-mails. Os destinatários da mensagem podem verificar a assinatura solicitando a chave pública diretamente ao domínio assinante e assim confirmar que a mensagem foi verificada por alguém em posse da chave privada para o domínio remetente [48].

Em outras palavras, o DKIM é uma especificação do IETF (*Internet Engineering Task Force*) que define um mecanismo para autenticação de e-mail baseado em criptografia de chaves públicas. Através do uso do DKIM (Figura 4.13) uma organização assina digitalmente as mensagens que envia, permitindo ao receptor confirmar a autenticidade das mensagens que recebe. Para verificar a assinatura digital, a chave pública é obtida por meio de consulta ao DNS do domínio do remetente. Ao contrário do SPF, que verifica somente o envelope, o DKIM verifica o cabeçalho da mensagem. Assim, o DKIM impõe um custo computacional adicional para processar cada mensagem, tanto para o MTA remetente quanto para o receptor [49].

O DKIM, além de verificar a assinatura da mensagem, também possibilita a verificação da integridade do conteúdo do e-mail. A autenticação da mensagem auxilia no controle de *spam* e *phishing* de e-mail [50].

Antes do DKIM, houve outros quatro padrões IETF para assinatura de e-mails [50], como listado a seguir:

- PEM - *Privacy Enhaced Mail* (1987) [51];
- PGP – *Pretty Good Privacy* (1991), padronizada mais tarde como OpenPGP [52];
- MOSS - *MIME Object Security Services* (1995), originado do PEM [53];
- S/MIME – *Secure MIME*, desenvolvido de forma independente pela RSA Security e mais tarde padronizado pela IETF [54].

De modo geral, as técnicas de autenticação de e-mails baseadas em assinatura poderiam ser a solução definitiva para problemas como o *phishing* de e-mail. Entretanto, a falta de padronização da técnica de autenticação utilizada no serviço de e-mail e a não-obrigatoriedade do uso da mesma impede a solução do problema. Na verdade, o próprio protocolo SMTP deixa a desejar quando tolera a ausência de uma técnica de autenticação. Essas técnicas consumem recursos computacional e configurações adicionais no sistema de e-mail e, portanto, nem sempre são adotadas. Além disso, um remetente (MTA de origem) que não assina suas mensagens não necessariamente é uma fonte não-confiável, mas a recusa de mensagens desta fonte nem sempre é possível. O mesmo vale para as

mensagens autenticadas, que mesmo que assinadas digitalmente, não são necessariamente de fontes confiáveis. A impossibilidade de verificação da autenticidade de todas as mensagens, somada ao fato que a assinatura em uma mensagem não significa que a mesma não seja *spam*, faz com que as técnicas de autenticação não representem a solução definitiva para o problema.

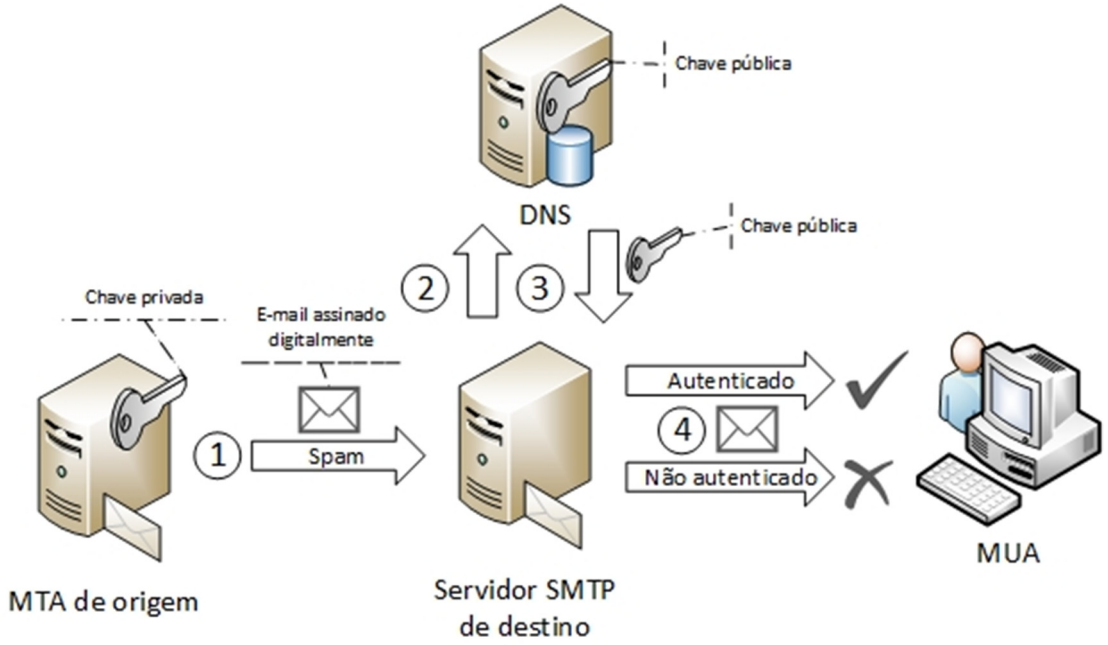



**Figura 4.13. Cenário de uso de DKIM.**

### 4.4.2. Técnicas de detecção de spam baseadas reconhecimento de padrões

Reconhecimento de padrões consiste na descoberta automática de padrões nos dados, usando algoritmos computacionais, que permitem classificar os dados em diferentes categorias [55]. Técnicas de reconhecimento de padrões são muito utilizadas no combate ao *spam*. Em se tratando de detecção de *spam*, algumas das técnicas ou classificadores mais utilizados são as abordagens Bayesianas, Redes Neurais, Árvores de Decisão e SVM (*Support Vector Machines*).

De um modo geral, as abordagens propostas para detecção de *spam* podem ser divididas em dois tipos:

(1) As que utilizam essencialmente a frequência de ocorrência de palavras como característica;

(2) As que se utilizam de várias características presentes no e-mail como informações de cabeçalho, uso de recursos específicos da linguagem HTML etc.

Também é possível encontrar propostas que combinam ambas [56]. Dentre as propostas que utilizam a frequência de palavras para a classificação de e-mails, os filtros bayesianos estão entre os mais comuns. A teoria da decisão bayesiana é uma abordagem estatística fundamental para o problema de classificação de padrões [40], sendo o classificador *Naïve Bayes* (NB) um dos mais utilizados para a classificação textual [57] (inclusive para a classificação de e-mails). Uma das razões é a sua simplicidade, o que o torna fácil de ser implementado e com boa taxa de acerto, comparando-se a outros algoritmos mais elaborados de aprendizagem de máquina [58]. O algoritmo NB tem sido um dos mais utilizados em propostas comerciais de filtros *antispam* [59].

Chen, C. e seus colegas [59] fizeram uma comparação de três propostas que utilizam NB [60, 61, 62] e mais quatro abordagens tradicionais (SVM, C4.5, NB e KNN). De um modo geral, a proposta apresenta alguns aprimoramentos no algoritmo NB, tentando comprovar que é melhor utilizar o algoritmo NB aprimorado do que métodos tradicionais de classificação (incluindo o próprio NB). Porém, ou autores não apresentam algumas informações essenciais, tais como taxa de falsos positivos.

Drucker, H., Wu, S., e Vapnik, V. N [63] apresentaram um trabalho que, embora seja antigo para área, apresenta vários conceitos muito utilizados em técnicas baseadas na detecção por frequência de ocorrências de caracteres, tais como TF (*Term Frequency*) – quantidade de vezes que uma palavra aparece em um documento, TF-IDF (*Term Frequency – Inverse Document Frequency*) – que define a importância de um termo dentro de uma coleção de documentos, e *Stop List* – lista de termos que não devem aparecer no vetor de características. Adicionalmente, também apresentam várias técnicas muito utilizadas para avaliar a performance do classificador e validação dos resultados, além de fazer um estudo do uso do classificador SVM em comparação com outros algoritmos de classificação (*Ripper*, *Rocchio* e *Boosting Decision Trees*).

Além das abordagens mais tradicionais de reconhecimento de padrões utilizadas para a detecção de spam, algumas propostas buscam a solução dos problemas sob uma outra perspectiva. Ma, W. e seus colegas [64], por exemplo, exploram a detecção de *spam* através de "seleção negativa", ou seja, uma linha de reconhecimento de padrões para casos em que não houve uma etapa de aprendizado (AMO – *Artificial Immune Systems*). Uma das vantagens seria a pro-atividade na detecção, sem necessitar uma etapa de treinamento, o que quase sempre é necessário nas abordagens *antispam* tradicionais.

Outra técnica que visa facilitar a etapa de aprendizado é o *co-training*, que permite a construção de um classificador com um número relativamente pequeno de amostras rotuladas [65]. Após a construção desse primeiro classificador de taxa de acerto "fraca", esse mesmo classificador vai se tornando mais robusto com e-mails não rotulados. O princípio base desta técnica é considerar que um classificador "fraco", feito com amostras rotuladas, pode encontrar amostras muito similares em uma base não rotulada. Então, vão sendo adquiridas novas amostras rotuladas para alimentar a base e refazer seu treinamento. Kiritchenko e Matwin [66] utilizaram *co-training* para a classificação de e-mails, realizando 50 treinamentos na base, à medida que mais e-mails eram classificados a taxa de acerto aumentava. No primeiro treinamento foi possível classificar corretamente 90% das mensagens com o classificador SVM, mas após 50 iterações de treinamento o resultado aumentou para 94%.

Dentro desta mesma área, alguns trabalhos propuseram soluções para combater as técnicas de ofuscação textual utilizadas pelos *spammers*. Braga e Ladeira (2007) propuseram uma abordagem que considera a dinamicidade do spam, ou seja, a quantidade de variações que costumam surgir visando burlar os mecanismos de detecção [67]. A abordagem é composta por três módulos, com função de pré-processamento, classificação e adaptação. A etapa de pré-processamento transforma a mensagem em valores numéricos, fazendo uso de árvores de Huffman adaptativas (árvores FGK) e um algoritmo de ordenação das mensagens com base na representação vetorial das palavras que a compõe. A vantagem de utilizar uma árvore adaptativa é a sua capacidade de acrescentar novas folhas sem precisar criar uma nova árvore desde o início. Na etapa de classificação foi utilizado o classificador SVM. A adaptação das mensagens é feita através de uma técnica chamada envelhecimento exponencial. Para retreinar o classificador no tempo $i$+1, são utilizadas novas mensagens que chegaram até o tempo $i$+1, porém nem todas as

mensagens do tempo *i* são escolhidas, dando maior importância às mensagens mais recentes.

O número de mensagens escolhidas em cada conjunto treinado decresce exponencialmente, e a periodicidade com que ocorre cada treinamento pode ser ajustada. O interessante dessa abordagem é que a mesma considera que as mensagens de *spam* mudam ao longo do tempo, possibilitando que, a cada novo treinamento, seja dado uma maior importância às mensagens mais recentes. Além disso, a abordagem também faz uso de um modelo numérico (Árvore de Huffman Adaptativa - FGK), o que acelera todo o processo se comparado ao uso de texto na aplicação das técnicas propostas.

Uma outra proposta parecida foi apresentada por Zhou et. Al [69], também fazendo uso de árvores FGK [68]. Os resultados mostram que o modelo com Árvore de Huffman Adaptativa levou vantagens sobre outras sete técnicas de classificação em oito dentre dez testes realizados. Seguindo a mesma linha de utilizar métodos de compressão para a detecção de *spam* em geral, alguns trabalhos propuseram o uso do Modelo de Markov [69, 70, 71, 72]. Modelos de compressão estatísticos podem ser utilizados para estimar a probabilidade de uma certa sequência de caracteres, podendo ser aplicados como classificadores Bayesianos.

Lee e Ng [74] utilizaram um método que faz uso do Modelo de Markov para o desofuscamento de palavras em *spam,* o qual eles chamam de Árvore Léxica do Modelo Escondido de Markov (LT-HMM – Lexicon Tree Hidden Markov Model) [73]. A abordagem integra ao Modelo de Markov um dicionário (léxico) e informações de contexto. Por exemplo, para que a palavra ofuscada *mortg3ge* seja convertida para *mortgage* (hipoteca), é necessário saber que *mortgage* é uma palavra do idioma inglês. O dicionário é construído através do modelo proposto, ou seja, da árvore léxica que na verdade é uma árvore de prefixos. Os resultados apresentados comprovam que a técnica é eficiente para desofuscar palavras, porém a abordagem possui um ponto fraco que é utilizar apenas os caracteres da tabela ASCII, deixando de lado os milhares de caracteres que estão na tabela Unicode. Além disso, a LT-HMM possui um número muito grande de estados (110.919 estados), o que não é muito conveniente para aplicações práticas.

Sculley et al. [75] buscaram solucionar o problema da ofuscação textual com o uso de algoritmos de busca aproximada. Uma das motivações para o uso desse tipo de técnica é que elas têm sido aplicadas com sucesso em áreas de biologia computacional com dados genômicos, já que as *strings* formadas por esse tipo de dados possuem inserções, substituições e exclusões de caracteres causadas por motivos evolucionais, fazendo semelhança com o que ocorre em termos comuns em *spam.* A técnica utilizada para fazer a busca aproximada de s*trings* é conhecida como *k-mers*. O problema do uso desse tipo de técnica é o custo computacional necessário para realizar a busca. Para resolver esse problema, foi proposto o uso de variantes desse método que utilizam os *kernels gappy* e *wildcard*, juntamente com o classificador *perceptron with margins*. Os resultados apresentados, apesar de interessantes, foram obtidos em uma base de *spam* que não possuía muitas variações de caracteres, se comparado ao grande número de possibilidades dentro do padrão Unicode (segundo o artigo, havia somente 25 variações da palavra 'viagra' com uma variedade mínima de caracteres visualmente semelhantes). Além disso, apesar de ser mencionado a importância de um bom desempenho para esse tipo de aplicação, não é apresentado nenhum resultado que mostre a sua aplicabilidade em um ambiente real de produção.

As técnicas de reconhecimento de padrões estão entre as mais utilizadas e mais pesquisadas na literatura relacionada ao combate ao *spam.* Devido à variedade de técnicas existentes, há vários trabalhos que analisam ou testam várias dessas técnicas na

classificação de e-mails [76, 77]. Embora bastante promissores, os classificadores mais utilizados no combate ao *spam* passam a ser do conhecimento dos *spammers*, que podem explorar determinadas limitações em etapas importantes da classificação, tal como a aprendizagem. Por exemplo, um *spammer* pode enviar mensagens que contenham propositalmente palavras frequentes em e-mails que não são *spam,* deste modo proporcionalmente estas palavras gerariam falsos positivos, se essas mensagens forem incluídas na base de treinamento. Esta técnica conhecida como *evasion* [78], costuma ser utilizada em outras aplicações relacionadas à segurança, como sistemas IDS (*Intrusion Detection System*) [79].

Embora não haja um consenso sobre qual é a melhor técnica de reconhecimento de padrões para classificação de e-mails, porque mensagens de *spam* não se enquadram em alguns casos muito específicos (e.g. técnicas da seção 4.2), várias abordagens desta área conseguem classificar corretamente a maioria das mensagens, justificando sua aplicação nas ferramentas disponíveis atualmente.

As técnicas baseadas em reconhecimento de padrões geralmente são indiferentes em relação às técnicas de *spam* apresentadas na seção 4.3.1, porém estão fortemente relacionadas às técnicas da seção 4.3.2. Como a aprendizagem de máquina se tornou algo popular entre as ferramentas *antispam*, várias das técnicas da seção 4.3.2 foram criadas especialmente para prejudicar a performance dos classificadores, aumentando o número de falsos positivos e falsos negativos.

### 4.4.3. Técnicas de detecção de spam baseadas em redes sociais

Algumas abordagens mais recentes de detecção de spam vêm explorando o sucesso crescente das redes sociais para obter informações que possam ajudar na classificação de mensagens. Li e Shen [80] sugerem que as informações obtidas em redes sociais (e.g. assuntos de interesse, esportes, religião, política etc) podem servir de guia para decidir se as informações contidas em um determinado e-mail devem ser consideradas *spam* para um usuário ou não. Um e-mail com propaganda usa as palavras-chave "perca peso", este tema pode ser considerado *spam* para um grupo e muito interesse para outro grupo com obesos, por exemplo. Assim, o interesse manifestado nas redes sociais ajudaria a separar os dois grupos e classificar corretamente os e-mails para cada caso.

A abordagem aplicada em MailRank [81] não utiliza uma rede social específica para a coleta de informações, mas propõe a criação de uma espécie de "rede de usuários de e-mail", que seria uma forma de identificar quem se comunica com quem, formando uma espécie de rede social de comunicação baseada e-mails. Através de um algoritmo de reputação, um servidor central identifica quais usuários são ou não são *spammers*. Por exemplo, se um usuário possui uma reputação alta (não é *spammer*) e encaminha um e-mail para um outro usuário, o destinatário pode ganhar uma pontuação que o faz ser rotulado como não-*spammer*. Embora interessante, o ponto fraco da proposta é que os resultados apresentados são baseados em um cenário simulado. Como esse ambiente não simula alguns problemas que podem existir em um cenário real (e.g. propagação de *malwares*, *bots* etc.) os resultados podem não ser muito conclusivos.

Do mesmo modo que as informações obtidas através de redes sociais podem ser úteis no combate ao *spam*, *spammers* ou *phishers*, pode-se fazer uso dessas informações para realizar ataques direcionados. Informações simples como orientação sexual e até mesmo data de aniversário do usuário podem disparar um e-mail com felicitações e propagar spam ou *malwares*. Um estudo realizado na Universidade de Michigan utilizou informações de 7 mil usuários do Facebook, identificou que muitos usuários que

possuíam um perfil com informações restritas, possuem perfis em outras redes sociais que disponibilizam publicamente algumas informações. Os resultados mostraram que um *spammer* poderia atingir 85% dos usuários com um ataque direcionado, iniciando pelo Facebook e utilizando outras redes sociais, conforme for o seu interesse [82].

Embora as redes sociais tenham informações relevantes para o combate ao *spam*, ainda são muito recentes se comparado ao serviço de e-mail, que se tornou essencial à maioria das pessoas há muito tempo. Nem todos os usuários da Internet possuem cadastro em redes sociais, assim em muitos casos não é possível a coleta dessas informações. Além disso, a grande variedade de redes sociais, com diferentes finalidades, diferentes formas de funcionamento e identificação na mesma, isto tudo pode se tornar um complicador quando for se buscar essas informações.

### 4.4.4. Técnicas de detecção de spam de imagem

Conforme apresentado na seção 4.3.2, o *spam* de imagem é uma técnica utilizada pelos *spammers* para disfarçar mensagens de texto usando imagens e, portanto, tornando impossível a sua interpretação por classificadores textuais. Uma das soluções para esse tipo de problema é o uso de técnicas de OCR (*Optical Character Recognition*) – Reconhecimento Ótico de Caracteres, que faz a conversão de imagens em texto (com uma fase pré-processamento) e depois as analisa usando classificador textual.

As técnicas utilizadas para tratar *spam* de imagem também estão relacionadas à área de reconhecimento de padrões (seção 4.4.2), porém utilizam técnicas mais específicas. Estas técnicas estão relacionadas principalmente ao pré-processamento das imagens, que deve ser eficiente, no intuito de facilitar o reconhecimento dos caracteres que serão entrada de um classificador de *spam* em formato textual.

Uma das técnicas mais utilizadas para a detecção de spam de imagem é o histograma [10, 17, 18, 19], representação gráfica da distribuição dos dados no espectro das tonalidades da imagem. Em imagens de *spam* poderia ser utilizado um histograma de cores, por exemplo, identificando padrões de histogramas para imagens de *spam* e para imagens de e-mails legítimos. Há ainda abordagens que exploram outros atributos bem específicos das imagens, como será apresentado a seguir.

Li et al. [10] apontam como principal problema a detecção de imagens com um fundo mais complexo (e.g. texto sobreposto a uma fotografia), pois nesses casos as características globais da imagem, tais como cor, textura e formato, são similares a imagens de e-mails legítimos e, portanto, diminuindo a taxa de detecção. De acordo com os autores, embora as características globais possam mudar depois do uso das técnicas de ofuscamento utilizadas pelos *spammers,* as características locais (SIFT - *Scale-Invariant Feature Transform,* MSER - *Maximally Stable Extremal Regions, Gabor wavelet, etc)*, permanecem inalteradas. Logo, as características locais que focam em detalhes da extração de características da imagem são essenciais para a detecção de *spam* de imagem.

Biggio, B. e seus colegas [20] consideram que o texto contido em uma imagem de *spam* possui a mesma essência de um texto contido em um *spam* textual. Logo, se for resolvido o problema de extração do texto da imagem, na sequência é possível aplicar técnicas de detecção de *spam* textual (e.g. filtro bayesiano) que a taxa de acerto será boa. A abordagem proposta possui uma etapa de pré-processamento onde o texto inserido nas imagens é convertido em texto puro. A etapa de treinamento é realizada após o pré-processamento – imagens convertidas em texto. Os resultados mostram que esse tipo de estratégia é eficiente somente nos casos em que as imagens não tiveram algum tipo de ofuscamento. Para as mensagens com ofuscamento, a detecção do *spam* é feita

considerando que a imagem que oferece algum tipo de dificuldade para o OCR é *spam*. O parâmetro utilizado para detectar a dificuldade imposta ao OCR foi medido através da "complexidade perimétrica" da imagem.

Os autores Biggio, B. e seus colegas [21], em outro artigo, consideram que as abordagens da literatura para *spam* de imagem estão focadas somente no tratamento das imagens do *spam,* não se preocupando na identificação das imagens ofuscadas. A proposta apresentada é uma continuação do trabalho anterior [20], onde o foco maior está na identificação dessas imagens. A identificação é feita com base num valor de "ruído", linearizando numa escala de 0 (sem ruído) a 1 (com muito ruído). A novidade em relação à abordagem anterior é que a complexidade perimétrica, que geralmente é utilizada para uma imagem inteira, passa a ser utilizada em diversos pontos da imagem, sendo possível identificar caracteres quebrados ou agregados ao ruído. Foram identificados padrões para (i) caracteres sem ruído; (ii) caracteres quebrados com pouco ruído no fundo; (iii) dois ou mais caracteres conectados pelo ruído e (iv) imagens nas quais o texto é colocado em cima de um fundo irregular.

Por ser um problema complexo, existe uma grande variedade de propostas para solucionar o *spam* de imagem, cada uma tentando e explorando algum aspecto específico do problema. Existem ainda, abordagens que propõem a exploração de ambas as características, textuais e das imagens [17]. De um modo geral, pode-se dizer o *spam* de imagem (seção 4.3.2) é um dos tipos mais explorados por *spammers* e estudados atualmente. As justificativas para isso podem ser o grande percentual de ocorrência desses e-mails em relação ao total global de *spam* (alta relevância do problema) e a grande variedade de técnicas de reconhecimento de padrões utilizadas para o processamento de imagens. Além disso, diferente de outras técnicas utilizadas na área, as quais visam a detecção de *spam* em geral, as abordagens existentes para a detecção deste tipo de *spam* geralmente buscam combater uma técnica bem específica de disseminação de *spam* (seção. 4.3.2.d).

### 4.4.5. Outras técnicas de detecção de spam

Há ainda algumas abordagens onde a principal contribuição não está totalmente relacionada a uma das categorias apresentadas anteriormente. Por exemplo, abordagens onde o principal fator para a detecção está relacionado a uma questão de infraestrutura ou baseado na colaboração de usuários ou outros servidores, em alguns casos utilizando uma das técnicas apresentadas nas seções anteriores.

Liu e seus colegas [83] propuseram uma infraestrutura *antispam* baseada em *grid* computacional. A abordagem considera que um e-mail é *spam* somente se for enviado para um número grande de destinatários, contabilizando o número de pessoas que receberam o e-mail em uma base denominada *CopyRank.* Um grupo de servidores é configurado para atrair mensagens de *spam*, que passam por um filtro bayesiano distribuído que encaminha as informações aos computadores clientes. Supõe-se que a taxa de falsos positivos será baixa uma vez que somente e-mails com um *CopyRank* alto poderão ser classificados como *spam.* A *grid* funciona com um *plugin* instalado no MUA (*Mail User Agent*), que se conecta com o servidor mais próximo da *grid* toda vez que recebe um e-mail. O cliente encaminha um *checksum* do e-mail recebido, e com base nele o servidor atribui um *CopyRank.* Além do *CopyRank*, o filtro bayesiano faz uma análise que indica se o e-mail é ou não *spam.* Os servidores trabalham de forma cooperada para manter uma tabela de *CopyRanks.*

Além de propostas como a que foi apresentada por Liu et al., utilizando reconhecimento de padrões, uma base de reputação e *grid computacional*, há outras

abordagens que são essencialmente baseadas em redes P2P e na colaboração dos usuários para reportar *spam* [14, 15].

Apesar destas técnicas serem inovadoras, assim como as apresentadas anteriormente, tem bom desempenho enquanto não forem conhecidas pelos *spammers*. Depois que se tornam publicas as técnicas perdem eficácia porque os *spammers* inventam maneiras de burlar seus mecanismos de identificação do *spam*. No caso de técnicas baseadas em reconhecimento de padrões, por exemplo, os modelos podem ser regerados, porém na prática a situação parece ser uma batalha interminável entre os desenvolvedores de técnicas de detecção de *spam* e os *spammers*.

### 4.4.6. Considerações acerca das técnicas de detecção de spam

Conforme apresentado nas seções 4.4.1 a 4.4.5, existe uma grande variedade de técnicas que buscam o combate ao *spam*. Essas técnicas, com exceção daquelas que são um objeto de pesquisas científicas mais recentes, normalmente estão ou podem estar incorporadas a ferramentas *antispam* utilizadas pelos administradores de sistema de e-mail. O *SpamAssassin*, uma das ferramentas *antispam* mais conhecidas, utiliza redes neurais e métodos estatísticos de classificação bayesiana para atribuir um *score* que representa a probabilidade de um e-mail ser *spam* ou não [43]. A decisão é baseada num limiar padrão (*threshold*) que também pode ser ajustado pelo administrador do serviço de e-mail. Adicionalmente, também são utilizadas regras para auxiliar a classificação dos e-mails.

A Tabela 4.3 resume as principais vantagens e desvantagens das técnicas de detecção de *spam* apresentadas nesta seção. As vantagens e desvantagens apresentadas na tabela não estão relacionadas somente à eficiência dessas técnicas em relação as técnicas de disseminação de *spam* apresentadas na seção 4.3. Mas, podem apresentar vantagens e desvantagens relacionadas a aspectos funcionais, como dificuldade de gerenciamento ou custo computacional.

Após já terem sido abordadas as principais técnicas utilizadas pelos *spammers* e as principais técnicas disponíveis para detecção do *spam*, já é possível perceber que nenhuma técnica de detecção é capaz de funcionar bem isoladamente. Uma análise mais detalhada sobre a relação de técnicas de disseminação de spam e as técnicas de detecção será apresentada na seção 4.5.

**Tabela 4.3. Resumo das técnicas de detecção de spam**

| Técnica | Vantagens | Desvantagens |
|---|---|---|
| *Whitelists e Blacklists* | • Pode agilizar o processo de entrega da mensagem em casos de não-*spam*.<br>• Pode agilizar o processo de recusa da mensagem em casos de *spam*.<br>• Reduz o custo computacional necessário para analisar a mensagem.<br>• É eficiente contra fontes conhecidas de *spam*. | • As *whitelists* podem significar um risco para a segurança do servidor.<br>• Seu uso possui muitas limitações.<br>• É ineficiente quando o remetente ou IP de origem da mensagem é substituído.<br>• Impõe dificuldade de gerenciamento em servidores com grande fluxo de e-mails. |
| Filtros baseados em palavras-chave | • Útil em casos em que o classificador falha em classificar um determinado *spam* de e-mail.<br>• Sua eficiência pode ser potencializada com o uso de expressões regulares. | • Uso deve ser feito com cautela, a fim de evitar o bloqueio indevido de mensagens.<br>• Uso da técnica deve ser considerado somente em último caso ou de maneira paliativa. |

| Técnica | Vantagens | Desvantagens |
|---|---|---|
| *Greylisting* | • Técnica eficiente quando o *spammer* não realiza o reenvio dos e-mails com falha na entrega. | • Perda de e-mails legítimos, quando o MTA de origem não faz o reenvio de mensagens.<br>• Possibilita atrasos na entrega de e-mails (entre o primeiro e segundo envio).<br>• É ineficiente contra máquinas de envio de *spam* mais robustas. |
| SPF | • Útil contra alguns casos de *phishing* de e-mail. | • Ineficiente quando o *spammer* possui um registro SPF.<br>• Frequente ausência de registros SPF publicados no DNS. |
| Técnicas baseadas em assinatura | • Permitem confirmar a autenticidade da mensagem.<br>• Útil contra alguns casos de *phishing* de e-mail.<br>• Algumas técnicas permitem atestar a integridade do conteúdo do e-mail. | • Falta de padronização da assinatura no serviço de e-mail.<br>• Indisponível em vários servidores de e-mail, dado que o protocolo SMTP, por padrão, não exige o uso de uma técnica baseada em assinatura.<br>• Aumenta o custo computacional para envio do e-mail tanto na origem quanto no destino da mensagem.<br>• Não garante que uma mensagem autenticada seja de fonte confiável. |
| Técnicas baseadas em reconhecimento de padrões | • Estão entre as técnicas mais utilizadas na detecção de *spam* e mais pesquisadas na literatura.<br>• Apresenta variedade de técnicas que podem ser utilizadas.<br>• Conseguem bons resultados na classificação, se comparado à outras técnicas. | • Os *spammers* costumam explorar as carências de algumas das etapas da classificação para não serem detectados por esse tipo de técnica.<br>• A maioria das técnicas necessita de uma etapa de aprendizagem. |
| Técnicas baseadas em redes sociais | • Podem explorar o sucesso crescente das redes sociais para auxiliar a detecção de *spam*.<br>• Algumas abordagens possibilitam a classificação das mensagens com base na percepção e interesse do usuário sobre determinados assuntos. | • Nem todos os usuários possuem cadastros em redes sociais, impossibilitando a coleta de informações em alguns casos.<br>• Variedade de redes sociais aumenta a complexidade da busca por informações.<br>• Tipo de técnica pouco explorada se comparada às demais. |
| Técnicas de detecção de spam de imagem | • Também estão relacionadas à área de reconhecimento de padrões (área muito explorada na Ciência da Computação).<br>• Estão entre as técnicas mais exploradas na literatura. | • Focam somente em um tipo específico de *spam*.<br>• Estão sujeitas a técnicas de ofuscamento de imagens. |
| Outras técnicas de detecção de spam | • Flexibilidade na criação de novas técnicas.<br>• Podem utilizar aspectos de infraestrutura, tais como redes P2P, *grid* computacional, colaboração de usuários na classificação etc. | • Podem ser facilmente burladas pelos *spammers*, inviabilizando sua proposta, uma vez que normalmente não podem ser facilmente adaptadas quando há mudança nas técnicas de *spam*. |

## 4.5. A eficiência das técnicas de detecção

Esta seção apresenta uma análise da relação entre as técnicas de disseminação de *spam* e as estratégias que podem ser utilizadas para mitigar o problema. Conforme mencionado anteriormente, o objetivo é fazer essa análise sempre a partir da técnica de disseminação em vez de utilizar as técnicas de detecção como ponto de partida, em contraposição ao que é convencionalmente feito na literatura [4, 22, 23, 24, 25, 26].

Esta análise, a partir do ponto de vista da técnica de disseminação, visa oferece vantagens em relação a abordagem tradicional. Pois, quando a análise é realizada a partir da técnica de detecção, normalmente não é apresentado previamente a grande variedade de técnicas que os *spammers* costumam utilizar no envio dos seus e-mails, causando a falsa impressão de que a técnica de detecção conseguirá mitigar o problema como um todo. Porém, quando a análise é realizada da ótica da técnica utilizada pelo *spammer*, é possível identificar quais técnicas *antispam* existentes podem ser eficientes contra aquele *spam* e quais são ineficazes ou podem falhar.

### 4.5.1 Análise das técnicas

As técnicas de disseminação (TD) mais importantes serão analisadas a seguir, considerando-se as técnicas *antispam* (TA) que podem ser utilizadas em cada caso. A relação entre as técnicas é avaliada com um *score* que vai de 1 (☆) até 5 (☆☆☆☆☆), sendo o seguinte o significado dos *scores.*

**Tabela 4.4. Scores utilizados na avaliação das técnicas *antispam*.**

| | |
|---|---|
| ☆ | A TA é totalmente ineficiente em relação a TD. |
| ☆☆ | A TA é pouco eficiente em relação a TD. |
| ★★★ | A TA é neutra em relação a TD. |
| ☆☆☆☆ | A TA é muito eficiente em relação a TD. |
| ☆☆☆☆☆ | A TA é totalmente eficiente em relação a TD. |

A Tabela 4.5 sumariza a relação entre as técnicas de disseminação (TD) apresentadas na seção 4.3 às técnicas *antispam* (TA) apresentadas na seção 4.4.

Em alguns casos, a TA não tem influência positiva nem negativa em relação à técnica de disseminação do *spam*. Esses casos (destacados com as estrelas sólidas) receberam três estrelas. Por exemplo, as técnicas baseadas em reconhecimento de padrões (e neste caso inclui-se a detecção de *spam* de imagem) não tem foco na forma de envio da mensagem, mas receberam três estrelas (avaliação neutra) devido a sua eficiência na detecção do conteúdo da mensagem, independente da forma de envio. Ou seja, a avaliação com uma ou duas estrelas (TA totalmente ou pouco ineficiente) seria imprecisa, pois haverá casos em que esta será bem-sucedida, independente da técnica de envio. Já a técnica *greylisting*, por exemplo, não tem foco no conteúdo do e-mail mas poderá ser eficiente dependendo da técnica de envio.

Há outros casos em que a técnica baseada no conteúdo normalmente está associada à forma de envio. Por exemplo, e-mails com o conteúdo falso normalmente estão associados às formas de envio que facilitam o anonimato (e.g. através de *botnets*). Nessas situações, as técnicas são avaliadas caso a caso ao invés de simplesmente receberem a avaliação neutra.

a) **Mecanismo simples de envio**

Os sistemas utilizados para disseminação de *spam* que se enquadram nesta categoria se caracterizam pelo envio de e-mails através de mecanismos pouco robustos, sem tratamento de erros ou reenvio de mensagens em caso de falha na entrega. Estas características fazem com que o *greylisting* seja muito eficiente contra este tipo de *spam*. Contudo, essa técnica não recebeu cinco estrelas pois está sujeita a falhas em algumas situações. O *greylisting* pode falhar por haver MTAs de domínios de organizações confiáveis que não estão configurados devidamente para tratar o reenvio no caso de uma recusa inicial de um e-mail. Isto causaria o não recebimento de um e-mail que não é *spam*.

O SPF não poderá ser utilizado de maneira isolada para bloquear a mensagem, que pode ser de uma origem confiável, provavelmente por não ter seus registros publicados no DNS. Da mesma forma, o uso de assinaturas digitais também não é totalmente eficiente, pois depende que o remetente utilize tal recurso. Além disso, o SPF e a assinatura digital possuem foco na autenticação da mensagem, e não em características específicas dos mecanismos simples de envio.

O bloqueio através de *blacklists* funcionará somente após a identificação da origem do *spam*, e o bloqueio por palavras-chave só será útil após o conhecimento do conteúdo da mensagem, que poderá ser alterado propositalmente pelo *spammer* em pouco tempo.

b) **Envio com tratamento de erros**

Quando o sistema de envio de *spam* é capaz de reenviar mensagens em caso de falha na entrega, técnicas como o *greylisting* se tornam totalmente ineficientes. No caso do SPF, alguns servidores podem possuir seus registros publicados no DNS, inviabilizando o uso da técnica. *Blacklists* poderiam ser muito eficientes contra as fontes conhecidas de *spam*, uma vez que nesta categoria os servidores de origem do *spam* não costumam mudar com tanta frequência, porém, pode haver exceções. O bloqueio por palavras-chave só terá utilidade após o conhecimento do conteúdo da mensagem, o que ocorre somente depois que boa parte do *spam* é recebido. No caso das assinaturas digitais, como alguns servidores podem pertencer a organizações confiáveis, nada impede que o e-mail seja assinado digitalmente pela empresa responsável pelo envio, o que não muda a possibilidade da mensagem ser *spam*.

c) **Substituição de remetente**

Quando o *spammer* utiliza artifícios como a substituição (forja) do remetente, técnicas como *blacklists* que utilizam o nome do domínio do remetente ou o endereço completo do e-mail do remetente, são totalmente ineficientes, uma vez que o *spammer* pode trocar o endereço do remetente quantas vezes quiser e utilizar um domínio de uma entidade confiável (que não deve ser bloqueado) para parecer um e-mail legítimo. O *greylisting* também é indiferente em relação a esta técnica, podendo falhar em muitos casos.

Por outro lado, técnicas como assinatura digital e SPF podem ser muito eficientes contra este tipo de *spam*, falhando apenas em alguns casos. Por exemplo, se o MTA de destino recebe um *phishing* que se apresenta como sendo de uma organização confiável (e.g. Bancos, órgãos governamentais etc.), a TA pode atestar a validade do remetente da mensagem através da chave pública ou do registro SPF que foi publicado no DNS. Contudo, estas técnicas não receberam cinco estrelas, pois dependem que o remetente (ou o domínio de uma organização confiável, no caso do *phishing*) assine as mensagens, ou tenha seus registros SPF publicados. Além disso, as vítimas de *phishing* também podem

ser enganadas quando é utilizado um domínio muito parecido com o legítimo (e.g. playpal.com em vez de paypal.com).

O bloqueio por palavras-chave, assim como no *envio com tratamento de erros*, só terá utilidade após o conhecimento do conteúdo da mensagem, normalmente depois que o MTA de destino já recebeu boa quantidade do *spam*. Neste caso, entretanto, a gravidade é bem maior pois pode se tratar de uma ameaça como o *phishing.*

d) **Envio através de *botnets***

O envio através de *botnets* tem muitas semelhanças com o mecanismo simples de envio, porém, com um agravante "o bloqueio por *blacklists* se torna ainda menos eficiente visto que a origem das mensagens pode mudar mais ainda". Isto acontece porque uma *botnet* é uma rede de grande escalabilidade, formada por computadores comprometidos de usuários espalhados pela Internet. O bloqueio por palavras-chave só terá utilidade após o conhecimento do conteúdo da mensagem, após muito conteúdo *spam* ter sido recebido. Por outro lado, a técnica *greylisting* é muito eficiente contra esse tipo de *spam*, visto que os mecanismos de envio utilizados pelos *spammers* possuem as mesmas características dos mecanismos simples de envio, ou seja, não tratam o reenvio da mensagem.

O uso de assinaturas digitais e SPF pode possuir boa eficiência contra este tipo de envio, que é característico de mensagens como o *phishing*, onde o *spammer* encaminha seus e-mails fraudulentos através de computadores comprometidos que estão espalhados pela Internet, com o objetivo de dificultar a sua localização.

e) **Envio através de *open relays***

Quando se trata de *spam* enviado através de *open relays*, o uso de *blacklists* se torna complicado pois o bloqueio de um IP do MTA de origem pode ocasionar também o bloqueio de todas as mensagens de uma organização confiável. O bloqueio de palavras-chave, como de costume, possui pouca eficiência se utilizada isoladamente e só deverá ser considerado quando todas as outras técnicas falharem. O *greylisting* terá pouca eficiência visto que a *open relay*, que é um servidor que normalmente pertence a uma organização confiável, normalmente é capaz de reenviar a mensagem após uma recusa inicial. O SPF e a assinatura digital só terão utilidade nos casos já previstos anteriormente, quando há a substituição de remetente e todos os pares da comunicação utilizam tais recursos.

f) **Inserção proposital de palavras em mensagens de e-mail**

Quando o *spammer* utiliza recursos textuais para enganar os classificadores de e-mails, normalmente o envio das mensagens é mal-intencionado. Ou seja, não se trata de um *spam* que permite que o usuário opte por não receber ou de uma simples divulgação com fins publicitários, feita por uma organização confiável. Nesse caso, a técnica de *spam* baseada no conteúdo dificilmente utiliza somente recursos textuais ou visuais, podendo vir acompanhada de alguma característica que permita a variação do endereço de origem da mensagem, a fim de dificultar ainda mais a sua detecção. Assim, o uso de *blacklists* passa a ser pouco eficiente em quase todos os casos em que o *spammer* explora os recursos disponíveis no corpo da mensagem. O bloqueio por palavras-chave só terá utilidade após o recebimento (conhecimento) do conteúdo da mensagem.

As técnicas baseadas em detecção de imagem são totalmente ineficientes contra este tipo de técnica, visto que neste caso o *spammer* usa subterfúgios técnicos para burlar os mecanismos *antispam*, usando aspectos textuais da mensagem. Já as técnicas de reconhecimento de padrões com foco no texto da mensagem passam a perder sua eficiência quando este tipo de técnica é utilizado.

g) **Troca ou inserção de caracteres na mensagem de e-mail**

A troca ou inserção intencional de caracteres no e-mail provoca uma queda ainda maior na performance dos classificadores textuais baseados em reconhecimento de padrões. Pois, as palavras que são utilizadas como características para determinar se um conteúdo é *spam*, normalmente são aquelas que são modificadas pelos *spammers*.

A existência desse tipo de técnica de inserção intencional de caracteres, não faz com que a área de reconhecimento de padrões seja desconsiderada na classificação textual das mensagens. A redução na sua eficiência ocorre principalmente nos classificadores tradicionais (e.g. classificadores bayesianos e redes neurais), mas há trabalhos que buscam o aprimoramento dessas técnicas contra este tipo específico de *spam* [67, 68, 69, 73, 74, 75].

h) **Conteúdo falso**

No caso de e-mails com conteúdo falso, como o *phishing*, as técnicas tradicionais de classificação textual são pouco eficientes, pois o conteúdo da mensagem se parece muito com uma mensagem real. Alguns trabalhos buscam utilizar outros aspectos não-textuais da mensagem como característica, objetivando a classificação desse tipo de *spam* [2, 84, 85, 86, 87] através de técnicas de reconhecimento de padrões.

No caso do *phishing*, técnicas como o SPF e uso de assinatura digital podem ser muito eficientes contra este tipo de *spam* se as organizações confiáveis, às quais o *phisher* tenta personificar na sua mensagem, fazerem uso de tais recursos. Nesse caso, o receptor da mensagem poderia atestar que o e-mail não foi enviado pela organização que consta como rementem no e-mail.

i) **Uso de imagens**

O uso de texto embutido em imagens inviabiliza qualquer tipo de técnica baseada no texto da mensagem. Por esse motivo, os filtros baseados em palavras-chave passam a ser totalmente ineficientes. As técnicas de reconhecimento de padrões baseadas no processamento textual também passam a ser totalmente ineficientes, exceto se houver uma etapa de pré-processamento que extraia o texto da imagem para então submetê-lo às técnicas de classificação textual.

Já as técnicas específicas para a detecção de *spam* de imagem são muito eficientes contra este tipo de técnica. Os trabalhos presentes na literatura apresentam abordagens que vão desde o pré-processamento da mensagem (inclusive para casos de ofuscação da imagem), para posterior classificação textual através de técnicas tradicionais, até o reconhecimento do *spam* pelas características identificadas na própria imagem que contém o texto embutido.

j) **Uso de recursos HTML**

O uso de recursos da linguagem HTML pode ser mal-intencionado em muitos casos (e.g. *phishing*). Assim, a eficiência das técnicas de reconhecimento de padrões é a mesma da TD conteúdo falso. Alguns trabalhos na literatura exploram características de e-mails no formato HTML para realizar a classificação das mensagens [2, 84, 85, 86, 87].

Quando há o uso mal-intencionado de recursos HTML, normalmente há também outras características que ocorrem em e-mails fraudulentos, como a falsificação do remetente da mensagem. Técnicas como o SPF com a assinatura digital possuem muita eficiência nesses casos.

k) **E-mail marketing**

O e-mail marketing é um tipo de *spam* muito complicado de classificar. Porque normalmente é originado em servidores MTA de organizações confiáveis com fins publicitários (que normalmente não mudam de endereço com frequência ou tentam esconder a sua localização); o uso de *blacklists* pode ser um pouco mais eficiente contra esse tipo de e-mail. Porém, alguns usuários podem ter interesse em receber os e-mails de algumas organizações (e.g. sites de e-commerce) e o uso de *blacklists* deve ser considerado em último caso.

O bloqueio por palavras-chave passa a ser totalmente ineficiente, pois há o agravante de determinadas palavras presentes em algum *spam* específico também estarem presentes em um e-mail publicitário, que é do interesse dos usuários, causando o bloqueio indevido da mensagem. Técnicas baseadas na autenticação da mensagem (SPF, *greylisting* e assinaturas digitais) são pouco eficientes, já que são recursos que também podem ser utilizados pelo *spammer*.

As técnicas baseadas em reconhecimento de padrões, especificamente aquelas baseadas nas características textuais, costumam ter bons resultados na classificação dos e-mails. Há ferramentas que conseguem, inclusive, classificar boa parte dos e-mails em três categorias: *não-spam*, *spam*, e *e-mail marketing*, deixando a decisão quanto aos e-mails desta última categoria a critério do usuário final. Técnicas baseadas em redes sociais também podem ser de grande utilidade, já que essas redes podem fornecer informações sobre os assuntos de interesse do usuário.

**Tabela 4.5. Resumo da relação ente TD e TA.**

| TA / TD | | Blacklists | Palavras-chave | Greylisting | SPF | Assinatura Digital | Reconhecimento de Padrões | Redes Sociais | Spam de Imagem |
|---|---|---|---|---|---|---|---|---|---|
| **Técnicas baseadas no envio do *spam*** | **1. Mecanismo simples de envio** | ☆☆☆ | ☆☆ | ☆☆☆☆ | ☆☆☆ | ☆☆☆ | ★★★ | ★★★ | ★★★ |
| | **2. Envio com tratamento de erros** | ☆☆☆ | ☆☆ | ☆ | ☆☆☆ | ☆☆☆ | ★★★ | ★★★ | ★★★ |
| | **3. Substituição de remetente** | ☆ | ☆☆ | ☆☆☆ | ☆☆☆☆ | ☆☆☆☆ | ★★★ | ★★★ | ★★★ |
| | **4. Envio através de *botnets*** | ☆☆ | ☆☆ | ☆☆☆☆ | ☆☆☆☆ | ☆☆☆☆ | ★★★ | ★★★ | ★★★ |
| | **5. Envio através de *open relays*** | ☆☆ | ☆☆ | ☆☆ | ☆☆☆ | ☆☆☆ | ★★★ | ★★★ | ★★★ |

| TA / TD | | Blacklists | Palavras-chave | Greylisting | SPF | Assinatura Digital | Reconhecimento de Padrões | Redes Sociais | Spam de Imagem |
|---|---|---|---|---|---|---|---|---|---|
| Técnicas baseadas no conteúdo do e-mail | **6. Inserção proposital de palavras** | ☆☆ | ☆☆ | ★★★ | ★★★ | ★★★ | ☆☆☆ | ☆☆ | ☆ |
| | **7. Troca ou inserção de caracteres** | ☆☆ | ☆☆ | ★★★ | ★★★ | ★★★ | ☆☆ | ☆☆ | ☆ |
| | **8. Conteúdo falso** | ☆☆ | ☆☆ | ★★★ | ☆☆☆☆ | ☆☆☆☆ | ☆☆☆ | ☆☆ | ☆ |
| | **9. Uso de imagens** | ☆☆ | ☆ | ★★★ | ★★★ | ★★★ | ☆ | ☆☆ | ☆☆☆☆ |
| | **10. Uso de recursos HTML** | ☆☆ | ☆☆ | ★★★ | ☆☆☆ | ☆☆☆ | ☆☆☆ | ☆☆ | ☆ |
| | **11. E-mail marketing** | ☆☆☆ | ☆ | ☆☆ | ☆☆ | ☆☆ | ☆☆☆☆ | ☆☆☆☆ | ☆ |

### 4.5.1 Considerações importantes sobre as técnicas analisadas

Um fato importante em relação ao *spam* é que não existe uma técnica de detecção perfeita, mesmo que a técnica busque combater um tipo específico de *spam*, pois sempre haverá um caso em que a técnica poderá falhar. Por exemplo, o *greylisting*, técnica amplamente utilizada, é eficaz contra os mecanismos de envio de *spam* que não tratam o reenvio da mensagem. Entretanto, o *greylisting* pode ocasionar a recusa indevida da mensagem, mesmo de MTAs legítimos que, por uma questão estratégica ou má configuração, acabam não tratando o reenvio.

Na avaliação de qualquer tipo de técnica *antispam*, além de avaliar a detecção das mensagens de *spam*, também se faz necessário a avaliação do comportamento da técnica em relação a e-mails que não são *spam*. As técnicas de detecção podem falhar, tanto na classificação do *spam* como não-*spam*, quanto na classificação de um e-mail não-*spam* como *spam*. Para que uma técnica possa ser considerada eficiente, mesmo que seja apenas contra um tipo específico de *spam*, deverá possuir uma possibilidade nula de ocorrência de falsos positivos e falsos negativos, ao mesmo tempo que uma taxa de altíssima de verdadeiros positivos e verdadeiros negativos. Por esse motivo, nenhuma das técnicas foi classificada como totalmente eficiente.

Algumas técnicas não chegaram a ser avaliadas sequer como muito eficientes, como no caso das *blacklists* e palavras-chave. Isto não desqualifica essas técnicas a ponto do seu uso ser desconsiderado. Porém, essas técnicas podem ser utilizadas como ferramentas auxiliares na detecção de *spam*, o que deve ser feito com bastante cautela. Já técnicas baseadas na autenticação das mensagens, como assinaturas digitais e SPF, que em alguns casos não receberam cinco estrelas porque sua eficácia depende da sua adoção

por terceiros (e não somente pelo MTA de destino), podem ser úteis quando todos os pares da comunicação utilizam a técnica.

Para as técnicas baseadas em informações extraídas de *redes sociais*, abordagem muito recente e ainda com poucas evidências de sucesso, a avaliação foi considerada *pouco eficiente* em quase todos os casos, exceto para o item 11 da tabela (*e-mail marketing*), visto que as informações obtidas nessas redes podem dizer muito a respeito dos assuntos de interesse do usuário.

A análise das técnicas baseadas em reconhecimento de padrões é a que deve ser interpretada com mais cautela. Apesar de ter sido classificada como muito eficiente somente em um dos casos, ainda é a mais utilizada e uma das mais eficientes na detecção de *spam*, com inúmeras possibilidades de aplicação ainda não exploradas dentro do problema em questão. Ou seja, suas técnicas estão entre as mais eficientes para a detecção do *spam* em geral, apresentando uma ampla gama de possibilidades de técnicas que ainda podem ser aplicadas para mitigar o problema. Ao analisar os *scores* atribuídos em cada avaliação, deve-se considerar que a TD foi avaliada em relação a TA que especificamente visam violar técnicas baseadas em reconhecimento de padrões.

Enfim, o uso das técnicas *antispam* não deve ser de executada de maneira isolada. Ou seja, é necessário que haja uma combinação de técnicas para que o problema seja mitigado da melhor maneira possível. Essa combinação pode ser feita para atender alguma necessidade específica, otimizando os resultados da classificação dos e-mails, ou ainda para adequar o ambiente de detecção a capacidade computacional do servidor de e-mails.

## 4.6. Conclusões

O envio indiscriminado de mensagens sem o consentimento de seus destinatários, prática conhecida como *spam,* é um problema que ainda está longe de ser solucionado, apesar do e-mail estar presente na vida da maioria das pessoas atualmente. Além do aborrecimento dos usuários, o *spam* pode causar problemas como a geração de custo computacional adicional e despesas com tecnologia e infraestrutura para a detecção desse conteúdo. Após a criação de novas técnicas para a detecção de *spam*, os *spammers* costumam desenvolver novas artimanhas para burlar os mecanismos de classificação dos e-mails, gerando uma situação de competição que parece não ter fim.

Este capítulo apresentou a análise das técnicas *antispam* sob uma nova perspectiva. Primeiramente foram apresentadas as principais técnicas utilizadas na disseminação de *spam* de e-mail. Em seguida foram apresentadas as principais técnicas de detecção existentes na literatura. Com este conhecimento já foi possível avaliar a eficiência de cada técnica *antispam* que foi apresentada. Por último, foram apresentados alguns aspectos relacionados a cada técnica utilizada pelos *spammers* e a eficiência das abordagens para combatê-las. Para cada técnica de detecção de *spam* foi atribuído uma nota (*score*) que representa a sua eficiência em relação a um tipo específico de técnica de disseminação de *spam*.

A análise realizada apresentou diversos resultados interessantes. Por exemplo, ficou evidente que nenhuma técnica *antispam* é eficiente sozinha, pois sempre haverá situações que levarão à falha. As técnicas baseadas em reconhecimento de padrões, observadas nas ferramentas *antispam*, são o alvo comum de ataque dos *spammers* devido ao seu uso e eficiência.

## 4.7. Referências

[1] Klensin, J. “RFC 2821 - Simple Mail Transfer Protocol”, disponível em: <http://www.ietf.org/rfc/rfc2821.txt>. Acessado em: 29/09/2015.

[2] Olivo, C. K.; Santin, A. O.; Oliveira, L. S. “Obtaining the Threat Model for E-mail Phishing”. em: Applied Soft Computing, vol. 13, issue 12, p.4841-4848. 2013.

[3] Kleiner, K. “Happy Spamyversary! Spam Reaches 30”, disponível em: <http://www.newscientist.com/article/dn13777-happy-spamiversary-spam-reaches-30.html>. Acessado em: 29/09/2015.

[4] Hoanca, B. “How good are our weapons in the spam wars?”, em: IEEE Technology and Society Magazine, vol. 25, Issue 1, p.22-30. 2006.

[5] Whitworth, B.; Whitworth, E. “Spam and the social technical gap”, em: IEEE Computer, vol. 37, Issue 10, p.38-45. 2004.

[6] Symantec “January 2011 Intelligence Report”, disponível em: <http://www.messagelabs.com/mlireport/MLI_2011_01_January_Final_en_us.pdf>. Acessado em junho de 2012.

[7] Symantec “May 2013 Intelligence Report”, disponível em: <http://www.symantec.com/content/en/us/enterprise/other_resources/b-intelligence_report_05-2013.en-us.pdf>. Acessado em: 29/09/2015.

[8] Symantec “Symantec Internet Security Threat Report – volume 19”, disponível em: <www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf>. Acessado em 29/09/2015.

[9] Leyden, J. “Spammers embrace e-mail authentication”, disponível em <http://www.theregister.co.uk/2004/09/03/email_authentication_spam/>. Acessado em: 29/09/2015

[10] Li, P.; Yan, H.; Cui, G.; Du, Y. “Integration of Local and Global Features for Image Spam Filtering”, em: Journal of Computational Information Systems, vol. 8, p.779-789. 2012.

[11] The New York Times, “Spam Doubles, Finding New Ways to Deliver Itself”, disponível em: <http://www.nytimes.com/2006/12/06/technology/06spam.html>. Acessado em: 29/09/2015.

[12] “There are 600,426,974,379,824,381,952 ways to spell Viagra”, disponível em: <http://cockeyed.com/lessons/viagra/viagra.html>. Acessado em: 30/09/2015.

[13] Olivo, C. K.; Santin, A. O.; Oliveira, L. E. S. “Avaliação de Características para Detecção de Phishing de E-mail”, Pontifícia Universidade Católica do Paraná, Curitiba – PR, Brasil. 2010.

[14] Damiani, E.; Vimercati, S.; Paraboschi, S.; Samarati, P. “P2P-based collaborative spam detection and filtering”, em: Proceedings of the Fourth International Conference on Peer-to-Peer Computing, IEEE, p. 176-183. 2004.

[15] Dimmock, N.; Maddison, I., “Peer-to-peer collaborative spam detection”, em: ACM Crossroads Magazine, vol. 11, issue 2, p. 4-4. 2004.

[16] Liu, Q.; Qin, Z.; Cheng, H.; Wan, M., “Efficient Modeling of Spam Images”, em: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE, p. 663-666. 2010.

[17] Soranamageswari, M.; Meena, C., “A Novel Approach Towards Image Spam Detection”, em: International Journal of Computer Theory and Engineering, vol. 3, p. 84-88. 2011.

[18] Xu, C.; Chiew, K.; Chen, Y.; Liu, J. , “Fusion of Text and Image Features: A New Approach to Image Spam Filtering”, em: Practical Applications of Intelligent Systems, Advances in Intelligent and Soft Computing, Springer, vol. 124, p. 129-140. 2012.

[19] Gao, Y.; Yang, M.; Zhao, X.; Pardo, B. Wu, Y.; Pappas, T.N.; Choudhary, A., “Image Spam Hunter”, em: International Conference on Acoustics, Speech and Signal Processing, IEEE, p. 1765-1768. 2008.

[20] Biggio, B.; Fumera, G.; Pillai, I.; Roli, F., “Image Spam Filtering Using Visual Information”, em: 14th International Conference on Image Analysis and Processing, IEEE, p. 105-110. 2007.

[21] Biggio, B.; Fumera, G.; Pillai, I.; Roli, , “Image spam filtering by content obscuring detection”, em: Fourth conference on e-mail and antispam. 2007.

[22] Caruana, G.; Li, M., "A survey of emerging approaches to spam filtering", em: ACM Computing Surveys (CSUR), vol. 44, Issue 2. 2012.

[23] Gansterer, W.; Ilger, M.; Lechner, P.; Neumayer, R.; Straub, J., "Anti-Spam Methods – State-of-the-Art", disponível em: <security.taa.univie.ac.at/files/FA384018-1.pdf>. Acessado em 30/09/2015.

[24] Blanzieri, E.; Bryl, A., "A survey of learning-based techniques of email spam filtering", em: Journal Artificial Intelligence Review, vol. 29, Issue 1, p. 63-92. 2008.

[25] Goodman, J.; Cormack, G.; Heckerman, D. "Spam and the ongoing battle for the inbox", em: Communications of the ACM, vol. 50, Issue 2, p. 24-33. 2007.

[26] Wang, X.; Cloete, I., "Learning to classify email: a survey", em: Proceedings of the Fourth Conference on Machine Learning and Cybernetics, vol. 9, p.5716-5719.

[27] “The Postfix Home Page”, disponível em: <http://www.postfix.org/>. Acessado em 30/09/2015.

[28] Bernstein, D. “qmail: Second Most Popular MTA on the Internet”, disponível em: <http://qmail.linorg.usp.br/top.html>. Acessado em 30/09/2015.

[29] “Secure Enterprise Email Solutions for Business | Exchange”, disponível em: <https://products.office.com/pt-br/exchange/email>. Acessado em 30/09/2015.

[30] “Thunderbird – Software Made to Make E-mail Easier”, disponível em: <https://www.mozilla.org/pt-BR/thunderbird/>. Acessado em 30/09/2015.

[31] “Software de E-mail e Calendário Microsoft Outlook”, disponível em: <https://products.office.com/pt-br/outlook/email-and-calendar-software-microsoft-outlook>. Acessado em 30/09/2015.

[32] Crispin, M. “RFC 3501 – Internet Message Access Protocol – Version 4rev1”, disponível em: <https://tools.ietf.org/rfc/rfc3501.txt>. Acessado em 30/09/2015.

[33] Myers, J.; Rose, M. “RFC 1939 – Post Office Protocol – Version 3”, disponível em: <https://www.ietf.org/rfc/rfc1939.txt>. Acessado em 30/09/2015.

[34] Freed, N.; Borenstein, I., “RFC 2045 - Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies”, disponível em: <http://tools.ietf.org/rfc/rfc2045.txt>. Acessado em 30/09/2015.

[35] Crocker, D. “RFC 882 – Standard for the Format of ARPA Internet Text Messages”, disponível em: <http://tools.ietf.org/rfc/rfc822.txt>. Acessado em 30/09/2015.

[36] Vaudreuil, G. “RFC 3463 – Enhaced Mail System Status Codes”, disponível em: <https://tools.ietf.org/rfc/rfc3463.txt>. Acessado em 30/09/2015.

[37] "Código Penal Brasileiro", Título II, Cap. VI, Art. 171, disponível em: <http://www.planalto.gov.br/ccivil_03/Decreto-Lei/Del2848.htm>. Acessado em 30/09/2015.

[38] Wang, P.; Sparks, S.; Zou, C. "An Advanced Hybrid Peer-to-Peer Botnet", em: IEEE Transactions on Dependable And Secure Computing, vol. 7, nº 2. 2010.

[39] Bianchi, N. M., "The Return of the Open Relays", disponível em: <http://www.spamhaus.org/news/article/706/the-return-of-the-open-relays>. Acessado em: 30/09/2015.

[40] Duda, R.; Hart, P.; Stork D., "Pattern Classification", 2ª edição, Wiley-Interscience. 2000.

[41] "The Unicode Standard – Technical Introduction", disponível em: <http://www.unicode.org/standard/principles.html>. Acessado em 30/09/2015.

[42] Liu, C.; Stamm, S. "Fighting Unicode-Obfuscated Spam", em: Proceedings of the Anti-Phishing Working Group - 2nd Annual eCrime Researchers Summit, p. 45-59, ACM. 2007.

[43] "SpamAssassin – The #1 Enterprise Open-Source Spam Filter", disponível em: <http://spamassassin.apache.org/>. Acessado em 24/09/2015.

[44] Jargas, A. M. "Shell Script Profissional", 1ª edição, Editora Novatec LTDA. 2008.

[45] Harrys, E. "The Next Step in Spam Control War: Greylisting", disponível em: <http://projects.puremagic.com/greylisting/whitepaper.html>. Acessado em: 30/09/2015.

[46] "Sender Policy Framework", disponível em: <http://www.openspf.org/>. Acessado em: 30/09/2015.

[47] Levine, J. R. "Experiences with Greylisting", em: Second Conference on e-mail and Anti-Spam. 2005.

[48] Allman, E.; Callas, J.; Delany, M.; Libbey, M.; Fenton, J.; Thomas, M., "RFC 4871 - DomainKeys Identified Mail (DKIM) Signatures", disponível em: <http://www.rfc-editor.org/rfc/rfc4871.txt>. Acessado em: 30/09/2015.

[49] Antispam.br, "Domain Keys Identified Mail (DKIM)", disponível em: <http://antispam.br/admin/dkim/>. Acessado em 30/09/2015.

[50] Hansen, T.; Crocker, D.; Hallam-Baker, P. "RFC 5585 - DomainKeys Identified Mail (DKIM) Service Overview", disponível em: <http://tools.ietf.org/rfc/rfc5585.txt>. Acessado em 30/09/2015.

[51] Linn, J. "Privacy Enhancement for Internet Electronic Mail", disponível em: <https://tools.ietf.org/rfc/rfc989.txt>. Acessado em 30/09/2015.

[52] Callas, J.; Donnerhacke, L.; Finney, H.; Shaw, D.; Thayer, R. "RFC 4880 - OpenPGP Message Format", disponível em: <https://tools.ietf.org/rfc/rfc4880.txt>. Acessado em 30/09/2015.

[53] Crocker, S.; Freed, N.; Galvin, J.; Murphy, S., "RFC 1848 - MIME Object Security Services", disponível em: <https://tools.ietf.org/rfc/rfc1848.txt>. Acessado em 30/09/2015.

[54] Ramsdell, B., "RFC 3851 - Secure/Multipurpose Internet Mail Extensions (S/MIME) - Version 3.1 - Message Specification", disponível em: <https://tools.ietf.org/rfc/rfc3851.txt>. Acessado em 30/09/2015.

[55] Bishop, C. "Pattern Recognition and Machine Learning", 1ª edição, Springer. 2007.

[56] Karthika, D.; Hamsapriya, T.; Raja, M.; Lakshmi, P. "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", em: International Conference on Process Automation, Control and Computing (PACC), IEEE, p. 1-7. 2011.

[57] Schneider, K. "A comparison of event models for Naive Bayes anti-spam e-mail filtering", em: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1, ACM, p. 307-314. 2003.

[58] Androutsopoulos, I.; Paliouras, G.; Michelakis, E. "Learning to Filter Unsolicited Commercial E-Mail", em: NCSR "Demokritos" Technical Report, nº 2004/2, disponível em: <http://nlp.cs.aueb.gr/pubs/TR2004_updated.pdf>. Acessado em: 30/09/2015.

[59] Chen, C.; Tian, Y.; Zhang, C. "Spam Filtering with Several Novel Bayesian Classifiers", em: 19th International Conference on Pattern Recognition, IEEE, p. 1-4. 2008.

[60] Frank, E; Hall, M.; Pfahringer, B. "Locally Weighted Naive Bayes", em: Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, ACM, p. 249-256. 2002.

[61] Zhang, H; Jiang, L.; Su, J. "Hidden Naive Bayes", em: Proceedings of the 20th national conference on Artificial intelligence - Volume 2, ACM, p. 919-914. 2005.

[62] Webb, G.; Boughton, J.; Wang, Z. "Not so Naive Bayes: Aggregating One-Dependence Estimators", em: Machine Learning, vol. 58, issue 1, p.5-24. 2005.

[63] Drucker, H.; Wu, S.; Vapnik, V. N. "Support Vector Machines for Spam Categorization", em: IEEE Transactions on Neural Networks, vol. 10, issue 5, p. 1048-1054. 1999.

[64] Ma, W.; Tran, D.; Sharma, D. "A Novel Spam e-mail Detection System Based on Negative Selection", em: Fourth International Converence on Computer Science and Convergence Information Technology, IEEE, p. 987-992. 2009.

[65] Blum, A.; Mitchell, T. "Combining labeled and unlabeled data with co-training", em: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, ACM, p.92-100. 1998.

[66] Kiritchenko, S.; Matwin, S. "E-mail Classification with Co-training", em: Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research, IBM Press, p. 8. 2001.

[67] Braga, I.; Ladeira, M. "Um modelo adaptativo para a filtragem de spam", em: VI Encontro Nacional de Inteligência Artificial, Rio de Janeiro – RJ, Anais do XXVII Congresso da Sociedade Brasileira de Computação, p. 1381-1390. 2007.

[68] Zhou, Y.; Mulekar, M.; Nerellapalli, P. "Adaptive Spam Filtering Using Dynamic Feature Space", em: 17th IEEE International Conference on Tools with Artificial Intelligence. 2005.

[69] Bratko, A.; Filipič, B. "Spam Filtering using Character-level Markov Models: Experiments for the TREC 2005 Spam Track", em: Proceedings of the 14th Text Retrieval Conference. 2005.

[70] Chhabra, S.; Yerazunis, W.; Siefkes, C. "Spam Filtering Using a Markov Random Field Model with Variable Weighting Shcemas", em: Fourth IEEE International Conference on Data Mining, p. 347-350. 2004.

[71] Ndumiyana, D.; Sakala, L. “Hidden Markov Models and Artificial Neural Networks for Spam Detection”, em: International Journal of Engineering Research & Technology, vol. 2, issue 4. 2013.

[72] Bratko, A.; Cormack, G.; Filipič, B.; Lynam, T.; Zupan, B. “Spam Filtering Using Statistical Data Compression Models”, em: Journal of Machine Learning Research, vol. 7, p. 2673-2698. 2006.

[73] Lee, H.; Ng, A. “Spam Deobfuscation Using a Hidden Markov Model”, em: Second Conference on Email and Anti-Spam. 2005.

[74] Lee, S.; Jeong, I.; Choi, S. “Dynamically Weighted Hidden Markov Model for Spam Deobfuscation”, em: Proceedings of the 20th International Joint Conference on Artificial Intelligence, p. 2523-2529, Morgan Kaufmann Publishers Inc. 2007.

[75] Sculley, D.; Wachman, G.; Brodley, C., “Spam Filtering Using Inexact String Matching in Explicit Feature Space with On-line Linear Classifiers”, em: Proceedings of the 15th Text Retrieval Conference. 2006.

[76] Kiran, R. S. S.; Atmosukarto, I. “Spam or Not Spam – That is the Question”, em: Technical Report, University of Washington. 2005.

[77] Guzella, T.; Caminhas, W. “A Review of Machine Learning Approaches to Spam Filtering”, em: Expert Systems with Applications, Elsevier, vol. 36, issue 7, p.10206-10222. 2009.

[78] Nelson, B.; Rubinstein, B.; Huang, L.; Joseph, A.; Tygar, J. “Classifier Evasion: Models and Open Problems”, em: Privacy and Security Issues in Data Mining and Machine Learning, vol. 6549, Lecture Notes in Computer Science, p. 92-98, Springer. 2011.

[79] Barreno, M.; Nelson, B.; Sears R.; Joseph, A.; Tygar, J. “Can Machine Learning be Secure?”, em: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, p. 16-25. 2006.

[80] Li, Z.; Shen, H. “SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box”, em: IEEE INFOCOM, p. 1835-1843. 2011.

[81] Chirita, P.; Diederich, J.; Nejdl, W. “MailRank: Using Ranking for Spam Detection”, em: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, p. 373-380. 2005.

[82] Brown, G.; Howe, T.; Ihbe, M.; Prakash, A.; Borders, K. “Social Network and Context Aware Spam”, em: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, p. 403-412. 2008.

[83] Liu, P.; Chen, G.; Ye, L.; Zhong, W. “Anti-spam grid: a dynamically organized spam filtering infrastructure”, em: Proceedings of the 5th WSEAS International Conference On Simulation, Modeling And Optimization, p. 61-66. 2005.

[84] Chen, J. e Guo, C. “Online Detection and Prevention of Phishing Attacks”, em: Communications and Networking in China, p.19-21. 2006.

[85] Cook, D., Gurbani, V. e Daniluk, M. “Phishwish: A Stateless Phishing Filter Using Minimal Rules”, em: Lecture Notes in Computer Science, p.182-186. 2008.

[86] Fette, I., Sadeh, N. e Tomasic, A. “Learning to Detect Phishing emails”, em: International World Wide Web Conference, p.649-656. 2007.

[87] M. Chandrasekaran, K. Narayanan, S. Upadhyaya “Phishing email Detection Based on Structural Properties”, em: Cyber Security Symposium. 2006.